Imprint

The iConference 2017 Proceedings are published by the iSchools.

Copyright © iConference 2017 is held by the authors of individual contributions in the Proceedings. Copyright permissions, when appropriate, must be obtained directly from the authors. The contributing authors are responsible for accuracy of all content. The publishers assume no responsibility for errors or omissions and no liability for any damages related to information in individual documents. Opinions expressed by contributors do not necessarily reflect official positions or policies of the iSchools.

The iConference 2017 Proceedings are available from the Illinois Digital Environment for Access to Learning and Scholarship (IDEALS) Open Access repository at the University of Illinois at Urbana-Champaign.
iConference 2017 Chairs

- **Conference Chairs**: Qing Fang, Wuhan University: School of Information Management; Sam Oh, Sungkyunkwan University: Library & Information Science and Data Science Department
- **Program Chairs**: Wonsik Jeff Shim, Sungkyunkwan University: Library & Information Science and Data Science Department; Chuanfu Chen, Wuhan University: School of Information Management
- **Local Arrangements Chairs**: Lihong Zhou, Wuhan University: School of Information Management; Fei Wang, Wuhan University: School of Information Management; Jie Xu, Wuhan University: School of Information Management
- **Papers Chairs**: Jian Qin, Syracuse University: School of Information Studies; Mika Grundstrom University of Tampere: School of Information Sciences; Jevin West, University of Washington: Information School
- **Poster Chairs**: Ann-Sofie Axelsen, University of Boras: The Swedish School of Library and Information Science; Min Song, Yonsei University: Library and Information Science; Esfandiar Haghverdi, Indiana University: School of Informatics and Computing
- **Doctoral Colloquium Chairs**: Kevin Crowston, Syracuse University: School of Information Studies; Bella Jing Zhang, Sun Yat-sen University, Guangzhou: School of Information Management; Pia Borlund, University of Copenhagen: Royal School of Library and Information Science; Jiangping Chen, University of North Texas: College of Information
- **Early Career Colloquium Chairs**: Marcia Zeng, Kent State University; Kathy Burnett, Florida State University: College of Communication and Information; Antonio Soares, University of Porto: Faculty of Engineering in cooperation with the Faculty of Arts
- **Workshops Chairs**: Noa Aharony, Bar-Ilan University: Department of Information Science; Daqing He, University of Pittsburgh: School of Information Sciences; Fred Ye, Nanjing University: School of Information Management
- **Sessions for Interaction and Engagement Chairs**: Ryan Shaw, University of North Carolina: School of Information and Library Science; Andrew Cox, University of Sheffield: Information School; Jin Ha Lee, University of Washington: Information School
- **Doctoral Dissertation Award Chairs**: Michael Seadle, Berlin School of Library and Information Science, Humboldt-Universität zu Berlin; Soo Young Rieh, University of Michigan: School of Information; Masooda Bashir, University of Illinois: Graduate School of Library and Information Science
- **Chinese Papers Track Chairs**: Ruhua Huang, Wuhan University: School of Information Management; Wei Lu, Wuhan University: School of Information Management; Dan Wu, Wuhan University: School of Information Management
- **iSchool Best Practices Chairs**: Gary Marchionini, University of North Carolina: School of Information and Library Science; Shigeo Sugimoto, University of Tsukuba: Graduate School of Library, Information and Media Studies; Ann-Sofie Axelsen, University of Boras: The Swedish School of Library and Information Science
- **iSchools and Industry Partnership Chairs**: Sean McGann, University of Washington: Information School; Gobinda Chowdhury, Northumbria University: Department of Mathematics and Information Sciences; Joon Lee, Seoul National University, Korea: School of Convergence Science and Technology
- **Conference Coordinator**: Clark Heideger, iSchools
- **Conference Administrative Assistant**: Molly Sanders, iSchools
- **Proceedings Editor**: Wjatscheslaw Sterzer, Humboldt-Universität zu Berlin
- **Proceedings Cover Design**: Thorsten Beck, HEADT Centre, Humboldt-Universität zu Berlin
iConference 2017 Program Committee

- Alessandro Acquisti, Carnegie Mellon University
- Denice Adkins, University of Missouri
- Naresh Agarwal, Simmons
- Jim Andrews, University of South Florida
- Ann-Sofie Axelsson, University of Boras
- Tim Baldwin, University of Melbourne
- Dania Bilal, University of Tennessee, Knoxville
- Leanne Bowler, University of Pittsburgh
- Josep Cobarsi Morales, Open University of Catalonia
- Constantinos Coursaris, Michigan State University
- Gianluca Demartini, University of Sheffield
- Stephen Downie, University of Illinois
- Karen Fisher, University of Washington
- Paul Gandel, Syracuse University
- Lisa Given, Charles Sturt University
- Sara Grimes, University of Toronto
- Lorna Hughes, University of Glasgow
- Jaap Kamps, University of Amsterdam
- Robert Kauffman, Singapore Management University
- Diane Kelly, University of Tennessee
- Kyung-Sun “Sunny” Kim, University of Wisconsin-Madison
- Young Seek Kim, University of Kentucky
- Lars Konzack, University of Copenhagen
- Andy Koronios, University of South Australia
- Serap Kurbanoğlu, Hacettepe University
- Philippe Lenca, Télécom Bretagne
- Claire R. McInerney, Rutgers University
- Julie McLeod, University of Northumbria at Newcastle
- Ruth Nalumaga, Makerere University
- Sanghee Oh, Florida State University
- Virginia Ortíz-Repiso Jiménez, Universidad Carlos III de Madrid
- Jung-ran Park, Drexel University
- Vivien Petras, Humboldt-Universität zu Berlin
- Nils Pharo, University College Oslo and Akershus
- Peter Reid, Robert Gordon University Aberdeen
- David Reitter, Penn State University
- Howard Rosenbaum, Indiana University
- Reijo Savolainen, University of Tampere
- Kalpana Shankar, University College Dublin
- Elizabeth Shepherd, University College London
- Antonio Lucas Soares, University of Porto
- Martin Souček, Charles University in Prague
- Eduardo Vendrell Vidal, Polytechnic University of Valencia
- Iris Xie, University of Wisconsin – Milwaukee
- Yong Jeong Yi, Sungkyunkwan University
- Yan Zhang, University of Texas Austin
- Lihong Zhou, Wuhan University

iConference 2017 Reviewer

- Waseem Afzal, Charles Sturt University
- Denise Agosto, Drexel University
- Noa Aharony, Bar Ilan University
- Isola Ajiferuke, University of Western Ontario
- Dharma Akmon, University of Michigan
- Michael Albers, ECU
- Nicole D. Alemanne, Florida State University
- Daniel Gelaw Alemneh, University of North Texas
- Hamed Alhoori, Northern Illinois University
- Sarah K Alkire, University of Washington
- Suzanne Lorraine Allard, University of Tennessee Knoxville
- Robert B. Allen, Yonsei University
• Morgan G Ames, CSTMS, UC Berkeley
• Liping An, Nankai University
• Lu An, Wuhan University
• Swamy Ananthanarayan, University of Colorado Boulder
• Nazanin Andalibi, Drexel University
• Jack Andersen, University of Copenhagen
• Jim Andrews, University of South Florida
• Clément Arsenault, University of Montreal
• Diana Ascher, UCLA
• Ann-Sofie Axelsson, University of Borås
• Diane Bailey, University of Texas at Austin
• Alex Ball, University of Bath
• Judit Bar-Ilan, Bar-Ilan University
• Veronica Barassi, Goldsmiths University London
• Eric P. S. Baumer, Cornell University, Ithaca
• Brian Beaton, Cal Poly
• Christoph Becker, University of Toronto
• Edith Beckett, New Jersey State Library
• Joeran Beel, Trinity College Dublin
• Nanyi Bi, Cornell University
• Wade Bishop, University of Tennessee
• Jean-François Blanchette, UCLA
• Phuritsabam Bobby, Manipur University
• Toine Bogers, Aalborg University Copenhagen
• Pia Bolund, University of Copenhagen
• Geoffrey C. Bowker, University of California Irivne
• Anne Elizabeth Bowser, Woodrow Wilson International Center for Scholars
• Sarah Bratt, Syracuse University
• Lauren Britton, Syracuse University
• JoAnn M. Brooks, independent
• Harry Bruce, University of Washington
• Sarah A. Buchanan, University of Missouri
• John M. Budd, University of Missouri
• Julia Bullard, University of Texas at Austin
• Gary Burnett, Florida State University
• Christopher Sean Burns, University of Kentucky
• Brian S. Butler, University of Maryland
• Shujin Cao, Sun Yat-sen University
• Vittore Casarosa, ISTI-CNR
• Michelle Caswell, UCLA
• Teresa Cerratto-Pargman, Stockholm University
• Tamy A Chambers, Indiana University Bloomington
• Renate Chancellor, Catholic University of America
• Senthil Chandrasegaram, University of Maryland, College Park
• Tiffany Chao, Graduate School of Library and Information Science, University of Illinois
• Apoorva Chauhan, Utah State University
• Chong Chen, Beijing Normal University
• Chuanfu Chen, Wuhan University
• Dingquan Chen, Sun Yat-sen University
• Hsin-liang Chen, Long Island University
• Hsuanwei Michelle Chen, San Jose State University
• Jiangping Chen, University of North Texas
• Jin Chen, Shanghai Jiao Tong University
• Li Chen, National Library of China
• Minghong Chen, Sun Yat-sen University
• Huanwen Cheng, CN, China, People's Republic of
• Kelvin Ka Sin Cheng, National University of Singapore
• Shih-Yi Chien, University of Pittsburgh
• Sou Hwan Choe, Sungkyunkwan University
• Wonchan Choi, Worcester Polytechnic Institute
• Steven Siu Fung Chong, University of Arizona
• Hans Dam Christensen, University of Copenhagen
• Mike Clarke, Google/University of Waterloo
• Rachel Ivy Clarke, Syracuse University
• Paul D. Clough, University of Sheffield
• Gilbert Cockton, NU UK, United Kingdom
• Monica Colon-Aguirre, Simmons College
• Anthony Joseph Corso, California Baptist University
• Kaitlin Costello, Rutgers, the State University of New Jersey
• Andrew M. Cox, The University of Sheffield
• Michael Douglas Crandall, University of Washington
• Shannon A Crawford Barniskis, University of Wisconsin-Milwaukee
• Kevin Crowston, Syracuse University
• Hong Cui, University of Arizona
• Xu Cui, Northwest University
• Sally Jo Cunningham, Waikato University
• Amber L. Cushing, University College Dublin
• Mats Dahlstrom, University of Bora
• Yang Dai, Anhui University
• Dharma Dailey, University of Washington
• Marija Dabbelo, Rutgers US
• Yuewu Dang, Sichuan University
• Karin Danielsson-Öberg, Umea University
• Ronald Eugene Day, Dept. of Information and Library Science, Indiana Univ.-Bloomington
• Shengli Deng, Wuhan University
• Zhonghua Deng, Wuhan University
• Karin Dessne, University of Bonn
• Sharmistha Dey, Queensland University of Technology,
• Jana Diesner, University of Illinois at Urbana-Champaign
• Ying Ding, Indiana University
• Brian Dobreski, Syracuse University
• Bryan Dosono, Syracuse University
• J Stephen Downie, School of Information Sciences
• Yunfei Du, University of North Texas
• Yufeng Duan, East China Normal University
• Kedma Duarte, Federal University of Santa Catarina
• Patrick M. Dudas, Pennsylvania State University
• Elizabeth Victoria Eikey, The Pennsylvania State University
• Houda El Mimouni, Drexel University
• Volkmann Engerer, Copenhagen University
• Heidi Enwald, University of Oulu
• Donald Everhart, UCSD
• Hao Fan, Wuhan University
• Qing Fang, Wuhan University
• Lesley Suzanne Johnson Farmer, California State University Long Beach
• Melanie Feinberg, University of North Carolina at Chapel Hill
• Yuanyuan Feng, Drexel University
• Antonia Ferrer Sapena, Polytechnic University of Valencia
• Bruce Ferwerda, Johannes Kepler University
• Casey Fiesler, University of Colorado Boulder
• Karen Fisher, University of Washington
• Kenneth R. Fleischmann, University of Texas at Austin
• Rachel Fleming-May, University of Tennessee
• Roger R Flynn, University of Pittsburgh
• Fred Fonseca, Penn State
• Kristin Fontichiaro, University of Michigan
• Andrea Forte, Drexel University
• Helena Francke, University of Boras
• Guo Freeman, University of Cincinnati
• Henry A. Gabb, University of Illinois at Urbana-Champaign
• Maria Gaede, Humboldt-Universität zu Berlin
• Chunmei Gan, Sun Yat-sen University
• Ning Gao, University of Maryland, College Park
• Emmanouel Garoufallou, Alexander Technological Educational Institute of Thessaloniki
• Susan Gasson, Drexel University
• Tali Gazit, Bar-Ilan University
• Qian Geng, Beijing Normal University
• Bela Gipp, University of Konstanz, Germany
• Adam Girard, University College Dublin
• Lisa M. Given, Charles Sturt University
• Yiwei Gong, Wuhan University
• Roberto I González-Ibáñez, Universidad de Santiago de Chile
• Tim Gorichanaz, Drexel University
• William Robert Grace, The Pennsylvania State University
• Elke Greifeneder, Humboldt-Universität zu Berlin
• Melissa Gross, Florida State University
• Mika Grundström, University of Tampere
• Chun Guo, School of Informatics and Computing, Indiana University
• Jing Guo, Shanghai Jiao Tong University
• Kailash Gupta, The International Emergency Management Society India-Chapter
• Stephanie W. Haas, University of North Carolina at Chapel Hill
• Esfandiar Haghverdi, Indiana University
• Oliver Haimson, UC Irvine
• Lala Hajibayova, Kent State University
• Sharon Haleva-Amir, Bar Ilan University
• Martin Halvey, University of Strathclyde
• Rebecca Anne Hardesty, University of California, San Diego
• Jenna Hartel, University of Toronto
• Helen Hasan, University of Wollongong
• Nazmul Hasan, Rajshahi University, Bangladesh
• Helmut Hauptmeier, Siegen University
• Caroline Haythornthwaite, Syracuse University
• Daqing He, University of Pittsburgh

• Yurong He, University of Maryland
• Roberto Henriques, Universidade Nova de Lisboa
• Shuyuan Mary Ho, Florida State University
• Jessica Hochman, Pratt Institute
• Ana Hol, Western Sydney University
• Liang Hong, Wuhan University
• Wolfram Horstmann, University of Göttingen
• Jining Hu, Wuhan University
• Lin Hua, Yunnan University
• Kun Huang, Beijing Normal University
• Ruhua Huang, Wuhan University
• Shuqing Huang, Nanjing Agricultural University
• Yun Huang, Syracuse University
• Isto Huvila, Uppsala University
• Jette Hyldegaard, Copenhagen University
• Salvatore Iaconesi, ISIA Design Florence
• Aylin Ilhan, Heinrich Heine University
• Charles Inskip, University College London
• Josh Introne, Michigan State University
• Vanessa Irvin, University of Hawai‘i-Manoa
• Corey Brian Jackson, Syracuse University
• Isa Jahnke, University of Missouri
• Wei Jeng, University of Pittsburgh
• Jennifer Jenkins, University of Arizona
• Grace YoungJoo Jeon, Michigan State University
• Tingting Jiang, Wuhan University
• Bo Jin, Shanghai University
• Nan Jing, Shanghai University
• Henrik Jochumsen, Copenhagen University
• Jenny Johannisson, University of Boras
• Veronica Johannsson, University of Boras
• Hideo Joho, University of Tsukuba
• Michael Jones, University of Toronto
• Soohyung Joo, University of Kentucky
• Heidi Julien, University at Buffalo
• Nicolas Jullien, Telecom Bretagne
• Jaap Kamps, University of Amsterdam
• Unmil Karadkar, The University of Texas at Austin
• Amir Karami, University of South Carolina
• Ping Ke, Nankai University
• Weimao Ke, Drexel University
• Diane Kelly, University of Tennessee
• Yvon Kermarrec, FR, France
• Heikki Keskustalo, University of Tampere
• Jinsook Kim, University of Illinois at Urbana-Champaign
• Jinyoung Kim, University of Maryland
• Kyung Sun Kim, University of Wisconsin-Madison
• Young Seek Kim, University of Kentucky
• Vanessa Kitzie, Rutgers University
• Bart Knijnenburg, Clemson University
• Emily Knox, University of Illinois at Urbana-Champaign
• Kyungwon Koh, The University of Oklahoma
• Rebeca Koskela, DataONE, University of New Mexico
• Yubo Kou, Purdue University
• Yong Ming Kow, City University of Hong Kong
• Kushwanth Koya, Northumbria University
• Adam Kriesberg, University of Maryland
• Pei-Yi (Patricia) Kuo, Northwestern University
• Diana S Kusunoki, White Ops
• Lynette Kvasny, Penn State
• Nicolas LaLone, Information Science and Technology
• Ari Llampinen, Stockholm University
• Alison Langmead, University of Pittsburgh
• Birger Larsen, Department of Communication and Psychology
• Kiersten F. Latham, Kent State University
• Amanda Lazar, Northwestern
• Christopher Lee, University of North Carolina
• Hosub Lee, University of California, Irvine
• Jin Ha Lee, University of Washington
• Jisue Lee, Florida State University
• Keeheon Lee, Korea Advanced Institute of Science and Technology
• Myeong Lee, University of Maryland
• Noah Lenstra, University of North Carolina at Greensboro
• Dirk Lewandowksi, Hamburg University of Applied Sciences
• Caifu Li, Anhui University
• Guangjian Li, Peking University
• Haitao Li, School of Information Management, Sun Yat-Sen University
• Kai Li, Nankai University
• Mingjie Li, Wuhan University
• Wenyi Li, Zhengzhou University
• Ying Li, Nankai University
• Yongjian Li, Nankai University
• Yueling Li, Nankai University
• Yuhua (Jake) Liang, Chapman University
• Xia Lin, Drexel University
• Zack Lischer-Katz, Rutgers University
• Bin Liu, Tianjin Normal University
• Chang Liu, Peking University
• Ping Liu, Wuhan University
• Xiaozhong Liu, Indiana University Bloomington
• Vincent M. Livoti, Long Island University
• Annemaree Lloyd, Swedish School Library and Information Science
• Yi Long, Shanghai University of Political Science and Law
• Irene Lopatovska, Pratt Institute
• Claudia Lopez, Universidad Tecnica Federico Santa Maria
• Wen Lou, East China Normal University
• Hejian Lu, Anhui University
• Kun Lu, University of Oklahoma
• Quan Lu, Wuhan University
• Xiaobin Lu, Renmin University of China
• Yingjun Lu, Wuhan University
• Yonghe Lu, Sun Yat-sen University
• Ana Lucic, University of Illinois at Urbana-Champaign
• Christopher Lueg, University of Tasmania
• Anna Hampson Lundh, University of Borås and Curtin University
• Caitlin Elizabeth Lustig, University of California, Irvine
• Elizabeth Josephine Lyon, University of Pittsburgh
• Feicheng Ma, Wuhan University
• Haiqn Ma, Heilongjiang University
• Jinxuan Ma, Emporia State University
• Lai Ma, University College Dublin
• Renjie Ma, Anhui University
• Haley MacLeod, Indiana University
• Rachel Magee, University of Illinois Urbana-Champaign
• Agnes Mainka, Heinrich-Heine University
• Jin Mao, University of Arizona
• Janet Marsden, Syracuse University
• Andrea Marshall, Drexel University
• Marianne Martens, Kent State University
• Crystle Martin, University of California Irvine
• Christopher Mascaro, Drexel University
• Kathryn Masten, Texas Woman’s University
• Eleanor Mattern, University of Pittsburgh
• Matthew S. Mayernik, National Center for Atmospheric Research, University Corporation for Atmospheric Research
• Philipp Mayr, GESIS Leibniz Institute for the Social Sciences
• Athanasios Mazarakis, Kiel University
• Claire McGuinness, University College Dublin
• Dana Mckay, Swinburne University of Technology Library
• Elspeth McKay, RMIT University
• Julie McLeod, Northumbria University
• Amanda Menking, University of Washington
• A.J. Million, University of Missouri, United States of America
• Stasa Milojevic, Indiana University
• Alex Mitchell, National University of Singapore
• Matthew Mitsui, Rutgers University
• Eric Monteiro, NTNU Norway
• Robert D. Montoya, UCLA US
• Miraída Morales, Rutgers University
• Camilla Moring, University of Copenhagen, RSLIS
• Atsuyuki Morishima, University of Tsukuba
• Heather Lea Moulaison Sandy, University of Missouri
• Angela Patricia Murillo, 1) University of North Carolina at Chapel Hill; 2) Novozymes
• Karine Nahon, Interdisciplinary Center Herzliya and University of Washington
• Bhuv Narayan, University of Technology Sidney
• Bryce Clayton Newell, Tilburg University
• Chaooquin Ni, Simmons College
• David Martin Nichols, University of Waikato
• Hui Nie, Sun Yat-sen University
• Yonghao Nie, Sun Yat-sen University
• Ragnar Nordlie, Oslo and Akershus University College of Applied Sciences
• David Novick, The University of Texas at El Paso
• Williams Ezinwa Nwagwu, University of South Africa
• Brian Clark O’Connor, College of Information University of North Texas
• Casey O’Donnell, Michigan State University
• Adrian Ogletree, Board Effect US
• ChiYoung Oh, University of Maryland
• Sanghee Oh, Florida State University
• Peter Organisciak, University of Illinois at Urbana-Champaign
• Felipe Ortega, University Rey Juan Carlos
• Virginia Ortiz-Repiso, Universidad Carlos III de Madrid
• Xinru Page, Bentley University
• Drew Paine, University of Washington
• Marco Painho, NOVA Information Management School Universidade Nova de Lisboa
• Galen Panger, UC Berkeley
• Min Sook Park, Florida State University
• SoHyun Park, Seoul National University
• Sun Young Park, University of Michigan
• Vivien Petras, Humboldt-Universität zu Berlin
• Mary Pietrowicz, University of Illinois
• Qing Ping, Drexel University
• Mattthew Pittman, University of Oregon
• University Pittsburgh, University of Pittsburgh
• Angela Pollak, University of Alberta
• Michael Preminger, Oslo & Akershus University College of Applied Science (HiOA)
• Nathan Richard Prestopnik, Ithaca College
• Ricky Punzalan, University of Maryland College of Information Studies
• Michelle Purcell, Drexel University
• Jian Qin, Syracuse University
• Angela U. Rammarine-Ricks, Syracuse University
• Congjing Ran, Wuhan University
• Edie Rasmussen, University of British Columbia
• Susan Rathbun-Grubb, University of South Carolina
• Ankita Raturi, University of California, Irvine
• Peter Reid, Robert Gordon University
• Gabby Resch, University of Toronto
• Scott Robertson, University of Hawaii
• Abebe Rorissa, University at Albany, State University of New York
• Howard Rosenbaum, US, United States of America
• Henriette RouedCunliffe, University of Copenhagen
• Chelecie Rowell, Wake Forest University
• Ehsan Sabaghian, Syracuse University
• Simone Sacchi, Columbia University
• Shaghayegh Sahebi, SUNY Albany
• Ashley E. Sands, UCLA US
• Vitor Santos, NOVA IMS
• Laura Saunders, Simmons College
• Steve Sawyer, Syracuse University
• Linda Schamber, University of North Texas
• Kristen Schuster, King's College London
• Melinda Sebastian, Fordham University
• John S. Seberger, University of California, Irvine
• Yongzhong Sha, Lanzhou University
• Elizabeth Shaffer, University of British Columbia
• Gwen Lisa Shaffer, California State University Long Beach
• Sarika Sharma, Syracuse University
• Ryan Shaw, University of North Carolina at Chapel Hill
• Alain Shema, Syracuse University
• Jian Shi, Nankai University
• Patrick C. Shih, Indiana University Bloomington
• Wonsik Shim, Sungkyunkwan University
• Yotam Shmargad, University of Arizona
• Jessica L. Shumake, University of Arizona
• Li Si, Wuhan University
• Gianmaria Silvello, University of Padova, Italy
• Laura Skouvig, University of Copenhagen
• Richard Aubrey Slaughter, The University of California, Irvine
• Stephen Slota, University of California, Irvine
• Madeline E. Smith, Colgate University
• Jaime Snyder, University of Washington
• Mads Solberg, University of Bergen
• Enmei Song, Wuhan University
• Linlin Song, School of Information management, Sun Yat-sen University
• Min Song, Yonsei University
• Jennifer Sonne, Syracuse University
• iSchool
• Eero Sormunen, University of Tampere, Finland
• Emma S. Spiro, University of Washington
• Ramesh Srinivasan, UCLA US
• Beth St. Jean, University of Maryland, College Park
• Juliane Stiller, Berlin School of Library and Information Science
• Wolfgang G. Stock, Heinrich Heine University Düsseldorf
• Richard Stoffle, University of Arizona
• Caroline Stratton, University of Texas at Austin
• Amber Stubbs, Simmons College
• Besiki Stvilia, Florida State University
• Xinming Su, Nanjing University
• Jianjun Sun, Nanjing University
• Si Sun, Rutgers University
• Yalin Sun, The University of Texas at Austin
• Yongqiang Sun, Wuhan University
• Sebastian Sünkler, Hochschule für Angewandte Wissenschaften Hamburg
• Tanja Svare, Aalborg University
• Sue Yeon Syn Catholic University of America
• Anna Maria Tammaro, University of Parma
• Biyong Tan, Shandong University
• Jian Tang, Central University of Finance and Economics, China
• Qiong Tang, Sun Yat-sen University
• Rong Tang, Simmons College
• Sikana Tanupabrunsun, Syracuse University
• Natalie Greene Taylor, University of South Florida
• Carol Tenopir, School of Information Sciences, University of Tennessee, Knoxville
• Kevin Tew, BYU, United States of America
• Andrea Karoline Thomer, University of Illinois at Urbana-Champaign
• Cheryl Annette Thompson, University of Illinois at Urbana-Champaign
• Violeta Trkulja, Humboldt-Universität zu Berlin
• Tienl Tsai, National Taiwan University
• Kristene Unsworth, Drexel University
• Julian Vamanu, University of Iowa
• Shawn Thomas Walker, University of Washington
• Thomas Walker, Long Island University
• Geoffrey Lee Walton, Manchester Metropolitan University
• Dakuo Wang, UC Irvine
• Fang Wang, Nankai University
• Jieyu Wang, UMBC
• Jing Wang, Pennsylvania State University
• Lin Wang, School of Management, Tianjin Normal University
• Meiling Wang, National Chengchi University
• Na Wang, Samsung Research America
• Peiling Wang, University of Tennessee
• Ping Wang, Wuhan University
• Qian Wang, Nankai University
• Shiwei Wang, Shanghai Academy of Social Sciences
• Tao Wang, UC Irvine
• Xiaoguang Wang, Wuhan University
• Xincui Wang, Wuhan University
• Zhijing Wang, Nankai University
• Amanda Waugh, University of Maryland
• Jenny Waycott, The University of Melbourne
• Nic Weber, University of Washington
• Dawei Wei, National Library of China
• Jingzhu Wei, Sun Yat-sen University
• Martin Weiss, University of Pittsburgh
• Lynnsey K. Weissenberger, Florida State University
• Yanlong Wen, Nankai University
• Brian Wentz, Shippensburg University
• Jevin West, University of Washington
• Jo Lynn Westbrook, University of Texas Austin
• Michael Majewski Widdersheim, School of Information Sciences, University of Pittsburgh
• Gunilla Widen, Abo Akademi University
• Lorna Wildgaard, Copenhagen University
• Kyle Williams, The Pennsylvania State University
• Rachel Williams, University of Wisconsin-Madison
• Bruce Wilson, University of Tennessee
• R. Jason Winning, University of California, San Diego
• Dietmar Wolfram, University of Wisconsin-Milwaukee
• Maria Wolters, University of Edinburgh
• Stacy Wood, University of California Los Angeles
• Adam Worrall, University of Alberta
• Dan Wu, Wuhan University
• Jiang Wu, Wuhan University
• Lin Wu, Sun Yat-sen University
• Peng Wu, Nanjing University of Science and Technology
• Xiaoli Wu, Hebei University of Economics and Business
• Zhenhua Wu, Nankai University
• Zhiqiang Wu, Wuhan University
• Judith Wusteman, University College Dublin
• Huichuan Xia, Syracuse University
• Lu Xiao, Syracuse University
• Qihui Xiao, Wuhan University
• Ximing Xiao, Wuhan University
• Yong Xiao, South China Normal University
• Iris Xie, Univeristy of Wisconsin-Milwaukee
• Wanli Xing, University of Missouri-Columbia
• Jian Xu, Sun Yat-sen University
• Jianhua Xu, Nankai University
• Jie Xu, Wuhan University
• Lifang Xu, Wuhan University
• Tan Xu, AT&T Research
• Xiao Xue, Business School, Nankai University
• Erjia Yan, Drexel University
• Jianyuan Yan, Nankai University
• Lijun Yang, Sun Yat-sen University
• Seungwon Yang, Louisiana State University
• Siluo Yang, Wuhan University
• Yaxing Yao, Syracuse University
• Fred Ye, Nanjing University
• Pippa Yeoman, The University of Sydney
• Ayoung Yoon, Indiana University Purdue University Indianapolis
• Bei Yu, Syracuse University
• Chuanming Yu, ZhongNan University of Economics and Law
• Liangzhi Yu, Nankai University
• Xiaojun Yuan, UAlbany
• Xiaoqun Yuan, Wuhan University
• Qingling Yue, Hubei University
• Ziming Zeng, Wuhan University
• Xianjin Zha, Wuhan University
• Bin Zhang, Wuhan University
• Chengzhi Zhang, Nanjing University of Science and Technology
• Chenwei Zhang, Indiana University Bloomington
• Guangxuan Zhang, Penn State University
• Jin Zhang, University of Wisconsin – Milwaukee
• Jing Zhang, Sun Yat-sen University
• Mei Zhang, University of Wisconsin-Madison
• Xiaojuan Zhang, Wuhan University
• Yan Zhang, School of Information
• Yang Zhang, Sun Yat-sen University
• Zhan Zhang, Drexel University
• Jianping Zhao, Heilongjiang University
• Junling Zhao, Hebei University
• Xing Zhao, East China Normal University
• Yang Zhao, Wuhan University
• Yiming Zhao, Wuhan University
• Yuxiang Zhao, Nanjing University of Science and Technology
• You Zheng, Syracuse University
• Lihong Zhou, Wuhan University
• Lihong Zhou, Wuhan University
• Xiaoyan Zhou, Beijing Normal University
• Xiaoying Zhou, Renmin University of China
• Yaolin Zhou, Wuhan University
• Yi Zhou, Sun Yat-sen University
• Qinghua Zhu, Nanjing University
• Michael Zimmer, University of Wisconsin-Milwaukee
Preface

The iConference 2017 was held on 22-25 March in Wuhan, China. This year’s theme “Effect • Expand • Evolve: Global Collaboration across the Information Community” reflects the attitude and intriguing contributions from experts and aspiring scholars from all over world. Over 500 participants and almost 150 selected written submissions show the growing relevance and evolution of the profession in the field of Library and Information Science (LIS).

30 completed research papers, 36 preliminary results papers, 68 posters, 5 workshop as well as 7 Social Engagement and Interaction (SIE) proposals comprise the two-volume iConference 2017 Proceedings, which are being made available for research purposes as well as for the public interest on the Illinois Digital Environment for Access to Learning and Scholarship (IDEALS) platform. Both proceedings volumes will give readers the opportunity to engage with a multitude of topics from various perspectives, and offer creative solutions to challenges in the field. The tables of contents provide the readers with an overview of this year’s contributions and point them directly to the desired work.

In the spirit of discoverability and accessibility, the proceedings volumes and the individual contributions have been attached to a Persistent Identifier, have a model citation and are deposited in PDF/A format, a long-term preservation standard that captures the inherent look and feel of the documents.

Wjatscheslaw Sterzer (Proceedings Editor)
## Contents

<table>
<thead>
<tr>
<th>Title</th>
<th>Authors</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>VennTags: A File Management System based on Overlapping Sets of Tags</td>
<td>Albadri, N.; Dekeyser, Stijn; Watson, Richard</td>
<td>1</td>
</tr>
<tr>
<td>An In-Depth Analysis of Tags and Controlled Metadata for Book Search</td>
<td>Bogers, Toine; Petras, Vivien</td>
<td>15</td>
</tr>
<tr>
<td>Enrichment of Cross-Lingual Information on Chinese Genealogical Linked Data</td>
<td>Dong, Hang</td>
<td>31</td>
</tr>
<tr>
<td>Re-intermediation in the Fashion Industry: A Qualitative Study on Brokers in the Dongdae-mun Fashion District</td>
<td>Kim, Min-Joon; Na, Nuri; Lee, Joong-Seek</td>
<td>43</td>
</tr>
<tr>
<td>Towards a Graph-based Data Model for Semantics Evolution</td>
<td>Li, Xuhui; Sun, Dewang; Liu, Mengchi; Qian, Tieyun; Ma, Feicheng</td>
<td>54</td>
</tr>
<tr>
<td>Access to Billions of Pages for Large-Scale Text Analysis</td>
<td>Organisciak, Peter; Capitanu, Boris; Underwood, Ted; Downie, J. Stephen</td>
<td>66</td>
</tr>
<tr>
<td>Identifying Users’ Gender via Social Representations</td>
<td>Qian, Tieyun; Zhu, Peisong; Li, Xuhui; Sun, Dewang</td>
<td>77</td>
</tr>
<tr>
<td>Let’s Workout! Exploring Social Exercise in an Online Fitness Community</td>
<td>Zeng, Li; Almquist, Zack W.; Spiro, Emma S.</td>
<td>87</td>
</tr>
<tr>
<td>Welcome New Americans! Investigating the role of hyper-local online communities in integration of immigrants</td>
<td>Farzan, Rosta; Ahmed, Grace; Espinoza, Yvette; Euley, Natalie; Villarreal, Marissa</td>
<td>99</td>
</tr>
<tr>
<td>Personalized Community Detection in Scholarly Network</td>
<td>Gao, Zheng; Liu, Xiaozhong</td>
<td>107</td>
</tr>
</tbody>
</table>
Human-assisted OCR of Japanese Books with Different kinds of Microtasks
Ikeda, Kosetsu; Hayashi, Ryota; Nagasaki, Kiyonori; Morishima, Atsuyuki

Enter a Job, Get Course Recommendations
Li, Nan; Suri, Naren; Gao, Zheng; Xia, Tian; Börner, Katy; Liu, Xiaozhong

Anachronism in Global Information Systems: the cases of Catalogue of Life and Unicode
Montoya, Robert D.; Erickson, Seth R.

Internal/External Information Access and Information Diffusion in Social Media
Xia, Tian; Yu, Xing; Gao, Zheng; Gu, Yijun; Liu, Xiaozhong

Toward Understanding Causes of Anomaly in Dynamic Restaurant Rating
Li, Lee; Lu, Zhao; Zhang, Danchen; Zhao, Sanqiang; Zhang, Ke

Types of Tags for Annotating Academic Blogs
Li, Lee; He, Daqing; Zhang, Danchen; Chi, Yu; Zhang, Chengshi

Automatic Course Website Discovery from Search Engine Results
Meng, Rui; Zhao, Zexin; Chi, Yu; He, Daqing

Analyzing Figures of Brain Images from Alzheimer’s Disease Papers
Tsutsui, Satoshi; Meng, Guilin; Yao, Xiaohui; Crandall, David; Ding, Ying

A Twitter-based Recommendation System for MOOCs based on Spatiotemporal Event Detection
Wang, Yuanyuan; Maruyama, Naoki; Yasui, Goki; Kawai, Yukiko; Akiyama, Toyokazu

Automatic ICD Code Assignment to Medical Text with Semantic Relational Tuples
Zhao, Sanqiang; He, Daqing; Zhang, Danchen; Li, Lei; Meng, Rui

Learning Semantic Representation from Restaurant Reviews: A Study of Yelp Dataset
Zhao, Sanqiang; Han, Shuguang; Meng, Rui; He, Daqing; Zhang, Danchen

SIE: Characterize iSchool Research Territory via Scholarly Data
Liu, Xiaozhong; Lu, Wei; Ding, Ying; Wunells, Sam

Second Interaction and Engagement on Information Research and Learning with Lifelogging Devices (IRLLD 2017)
Joho, Hideo; Gurrin, Cathal; Hopfgartner, Frank
VennTags: A File Management System based on Overlapping Sets of Tags

N. Albadri¹, Stijn Dekeyser¹, Richard Watson¹
¹University of Southern Queensland, Australia

Abstract

File systems (FS) are an essential part of operating systems in that they are responsible for storing and organising files and then retrieving those when needed. Because of the high capacity of modern storage devices and the growing number of files stored, the traditional FS model is no longer able to meet modern users’ needs in terms of storing and retrieving files. So using metadata emerges as an efficacy solution for the limitations of file systems.

In this paper we propose a new model dubbed VennTags to solve the FS problems. We do this by utilising the idea of overlapping the sets as in Venn diagram, and adopting DAG structure (instead of tree) to achieve that we have used tagging capability and exposed a query language at the level of the API. We evaluate the expressive power of VennTags model that shows its ability to resolve the FS limitations compared to other solutions.

Keywords: Operating Systems, File systems, metadata, tagging, Directed Acyclic Graph
Contact: nehad.albadri@usq.edu.au, richard.watson@usq.edu.au, stijn.dekeyser@usq.edu.au

1 Introduction

Personal computer systems, mobile and cloud-based as well as desktop oriented, are permanent companions in users daily lives, for both private and professional activities. Stored collections of files grow steadily as users obtain more files than ever before; whether those files store scientific and experimental data generated by the increasingly sophisticated instruments and computer models, work-related documents, or more personal collections of media and artefacts of our digital life (Lyman, 2003; Alvarado, Teevan, Ackerman, & Karger, 2003; Perišić, 2007). As a result of the increase of the size of the stored collections, the responsibilities that a user faces in storing, organising, and later retrieving files are becoming more complex and problematic. We emphasise our attention on the particular use case where a user stores a file in a file system to which they have write access, and subsequently (perhaps at a much later time) needs to locate and retrieve that file.

Variations of the problem exist, such as the file being created by another user, but the key characteristics are that the user knows that the file they are seeking does indeed exist and they have direct access to the file system (unlike the web search case).

Hierarchical file systems (HFSs) have been the standard for personal data management since 1970s (Carrier, 2005). However, HFS is not able to effectively support the tasks that users employ to manage their file collections (Seltzer & Murphy, 2009; Albadri, Watson, & Dekeyser, 2016). This is because (as mentioned above) that the number of personally created or curated files is so enormous that users typically cannot remember where their files are stored (Barreau & Nardi, 1995; Jackson & Smith, 2011) and how they are named, so support for effective searching is vital.

HFS is a single classification systems which means if an entity belongs to two distinct classes c₁ and c₂ then either c₁ ⊂ c₂ or c₂ ⊂ c₁. This property is handy when building a library classification system for physical books as a book can only reside at a single shelf location, but is limiting because in general many entities, especially files, have properties that violate this constraint; they may belong in two classes that are not in the ancestor relation. So, in HFS, files normally reside only in one particular directory in the hierarchy (Bergman, Gradovitch, Bar-Ilan, & Beyth-Marom, 2013; Lin, Hao, Changsheng, & Wei, 2014) which leads to problems when a user attempts to build a hierarchy of directories that reflects file properties and supports intuitive search strategies. On most Unix-like operating systems (Shacklette, 2004), symbolic and hard links are available to circumvent the shortcomings of hierarchical (tree) file system structures. The links are special files that contain a reference to another file or directory. They are cumbersome to use for personal file management as the link is not updated whenever the target pointed to by a link is moved,
renamed, or deleted; instead, it points to something that no longer exists. In addition, some applications do not handle links as if they were the objects they point to; but rather resolve the referenced path, using that instead of the link’s path. Another related issue is that if the name of the link is changed usually the link is renamed, not its target. The problems that arise from single classification file systems are explored in detail in this paper together with a solution that offers multi-classification.

Numerous attempts have been proposed in order for solving the limitations of HFSs. Some of these proposals rely on a rich collection of file metadata rather than the hierarchical directory structure (A. Ames et al., 2005; S. Ames, Gokhale, & Maltzahn, 2013; Dekeyser, Watson, & Motron, 2008; Gifford, Jouvelot, Sheldon, et al., 1991; Rizzo, 2004; Seltzer & Murphy, 2009). These systems are designed to fully replace the HFS. Another approach to dealing with HFS limitations is to introduce extra functionaility layered on top of the underlying existing file system. Many such approaches have surfaced, and we can group models and applications in terms of the main technique they employ (see Section 6). However, these attempts have some limitations that prohibit considering them as a solution to the HFS problems.

It is clear that for efficiency and practical reasons (e.g. binary files) a search must be conducted in terms of metadata associated with files rather than their contents (though metadata may be obtained automatically from file content). The structure of file management system metadata, and the services provided to manipulate that metadata, becomes a key contributing factor to the efficacy of any search process. We use the term “file management system” as a grouping of both specialised file systems as well as special-purpose applications, designed to improve on the metadata-related deficiencies of traditional hierarchical file systems.

In this work, we propose a new model (named VennTags) that allows overlapping sets (containers) of files which is isomorphic the idea of Venn diagram. In order to achieve that we use the rooted Directed Acyclic Graph (DAG) instead of the hierarchy (tree) structure. In VennTags model, the containers of files (named collections in this model) might have a plurality of membership (having more than one parent except the root at the same) while in the HFSs the containers (folders/directories) are allowed to have just one parent. In addition, in this novel model, we add tagging capability to the fundamental file management system structure (both collections and files) and a query ability as well to avoid the limitations of HFS. This would provide a uniform API that could be used to build richer generic user interfaces that could leverage the enhanced metadata structures to better support user file management activities.

Organisation In this paper we will first expose the motivation of this paper by showing the limitations of the HFSs in terms of file management and search which are summarized in Section 2. In Section 3, we identify the main service requirements with some important definitions. The main contribution of this paper is the novel model called VennTags (Section 4). Section 5 evaluates the proposed model in terms of the solving limitations of HFS as well as comparison to the other solutions. In Section 6, we explore the work related to our proposal and show the differences and the benefits of our proposed.

Contribution The contributions made in this paper include: named and described problems of the hierarchical file system, proposal of VennTags as a file management model to solve the traditional hierarchical file system, allowed overlapping the containers (collections) of files, Directed Cyclic Graph, and tags as a solution for the highlighted limitations, and introduction of a query language as part of the File management System API level to support easy retrieval of files.

2 Motivation Scenarios

Traditional file systems employ a model where files reside in a tree of directories. As such, HFSs support the creation of a user-defined classification system. Classification is a natural human activity that seeks to manage and understand complexity by recursively grouping classes of entities (e.g. files or plants that share common properties) into subclasses. As the classification tree is descended, associated entities have more inherited properties, and the number of members of the subclass decreases.

In the file system instance, the searcher iteratively descends the directory (classification) tree, at each step choosing one directory from the children of the current directory node based on its name (which should reflect its categorical relationship with its parent and siblings). Each step reduces the search space until a relatively small selection of files is presented for selection.

The basic support provided by HFSs to organise files in directories and facilitating iterative, navigational search is one reason for their longevity. However, while simple hierarchical directories may have been
sufficient in the past, ever-growing collections of files mean that HFSs are not able to meet modern users’ needs in terms of organising and retrieving information. Problems of traditional hierarchical file systems have been noted repeatedly in the literature (Seltzer & Murphy, 2009; A. Ames et al., 2005). In our previous research project (Albadri et al., 2016), we detailed HFS problems in this context. The following is a summary of these issues.

1. **Problem 1: Artificial hierarchies**
   Generally, the properties of files do not shape a natural subclass relationship, resulting in artificially constructed hierarchies. Consider files associated with university courses: these items have properties ‘course code’ and ‘year of offer’, but either of the hierarchies as shown in Figure 1 course-code and then year of offer or the second on then course code could justifiably be used to group course files.

2. **Problem 2: Classification**
   In a hierarchy, items can often belong to more than one sub-tree. Assume that the hypothetical course files introduced above are organised by year then course. Now imagine that a file should be included in both courses: in which directory should this file be placed? We can either select one directory to place the file Figure 2 a and b, keep duplicated copies (Figure 2 c), or keep one copy and place a hard or soft link to it in the sibling course directory. None of these solutions are practical and efficient (more details about that in (Albadri et al., 2016)).

3. **Problem 3: Problematic Pruning**
   This problem is a consequence of classification problem (above) where orienteering through an imperfect classification hierarchy leads to users not finding files they are looking for. So if a searcher branches the ‘wrong’ way while orienteering through the directory tree they will never find the file.

4. **Problem 4: Metadata management**
   In HFSs, bulk updates of metadata are inefficient. For example, if a user wishes to add or remove such meta-data, usually because the classification needs to be modified to better reflect reality, it often requires a sequence of non-trivial directory create, delete, and rename operations which must be carried out in the correct order.

5. **Problem 5: Native query support**
   The traditional file system API (e.g. POSIX ("IEEE Standard for Information Technology - Portable Operating System Interface (POSIX) Base Definitions", 2004)) has very limited query ability: either to find a single file given a path (e.g. stat, open) or to open and read the contents of a single directory (opendir, readdir, scandir). This limited query capability supports an orienteering style of search, but does not support file system wide queries of the kind that are provided by the special-purpose applications that are layered on top of the file system.

   We argue that a generic, powerful query mechanism will assist users in understanding the existing organisation of their file system instance and help identify (and hence help rectify) occurrences of incorrect classification.
Figure 2: Multiple classification choices

3 Framework

In the following section, our proposed file management system (VennTags) design is described by the set of functions exposed at the programming level—the application programming interface (API). We exclude from the description the kind of user level file system related operations provided by applications that layer above the file system API.

We further limit our presentation to just the parts of the API that deal with the file management system organisation. For example, file content manipulation operations and file management system privilege and protection operations are not handled. In Section 6, we will compare our approach to other research proposals, including those that also use tags.

Our proposed file management system offers three classes of services that are:

1. Create a file together with associated metadata.
2. Identifying files
   (a) The ability to look up or locate a single file given a metadata based specification. This is needed for the file ‘open’ operation. The file system metadata must be managed to ensure that every file has a unique metadata specification.
   (b) A query service that returns a group of files that meet class membership conditions. At the very least, this would include exposing the contents of a collection of files such as a directory. However we extend this notion to a generic file system query. These operations are critical in the development of user interfaces to file systems.
3. Modifying files’ Metadata
   (a) Update an item of metadata for a single file, for instance to change a name or a tag.
   (b) Reorganise a selected group of files by systematically applying possibly complex changes to those files’ metadata, thereby potentially reclassifying the files.

The fundamental physical entities stored in the proposed file system are tags, files, and collections of files. These are defined as follows, together with the related metadata based path and query concepts.

A tag is an item of metadata associated with a collection or a file. It could be represented as an unlimited length text string. This is a generalisation of the name that is associated with traditional HFS files and directories.

A file is a sequence of bits (or maybe a larger atomic data unit) that is stored in the file system. It has a unique system identifier. The logical organisation of the file system is unrelated to file content—files are simply represented within these structures by the system identifier. So, in the following, the word ‘file’ can usually be interpreted as synonymous for ‘file identifier’.
A file may have associated tags.

A **collection** is a file container. From a logical view it is an object that has some unique system identifier; each collection is associated with zero or more files (file identifiers). To offer the idea of the Venn diagram, the collections are organised as a **Directed Acyclic Graph (DAG)**.

A **DAG** is a directed graph that has no cycles so it is a special kind of directed graph while a rooted tree is a special kind of the DAG. The reason of preferring and selecting this kind of structure for this model is that DAG represents the solution of some of the HFS limitation as it provides a multi classification. This is because a collection might have a plurality of membership so the collections can be placed in several different categories at the same time if they are need to. In addition, as this structure has a single root and there is no cycles, the proposed model will provide the benefits of the semi-hierarchy in organising the collections’ of files

All files in a collection have been placed there because they share some semantic properties (e.g. all these files are associated with a particular project).

A collection may have associated tags and it may have links to any other collections. Every file in the collection is assumed to inherit this tag, as well as any tags associated with ancestor collections (see path discussion below).

A file in a collection can have a set of associated tags which are directly linked to the file. The file inherits from the containing collection and its parent to be the file path. The key semantic difference is that an atomic operation that affects a collection tag broadcasts to involve many files within that collection while a file tag affects just one file.

**A path** is a sequence of collections such that each member is a child of the preceding collection. It is the route from the tree root to a collection, and so unambiguously identifies a collection. Every collection can be uniquely identified by a path. A file path is the combination of a (collection) path together with the identified file within that collection. File system users navigate (Jones, Wenning, & Bruce, 2014) paths to find files.

**A query** specifies a search criterion in terms of collection and file metadata (tags). Performing a search based on such a query returns a set of zero or more files which may reside in many different collections.

The file system query is a key divergence from traditional HFS APIs. The functionality offered by a file system query can be duplicated by a client program of a traditional HFS but will likely suffer from poor efficiency due to the need for repeated file system calls.

## 4 VennTags Model

As mentioned in Section 2, HFS is a single classification system which causes a problem in terms of organising the files and re-finding them when needed. So we propose a file management system based on the Venn diagram idea in order to provide a multi-classification model. This means allowing a collection to belong to plurality of other collection with some constraints as will be shown in the following.

VennTags considers a rooted directed acyclic graph where there are a single root as tree but a collection is allowed to have a plurality of membership instead of just having a single membership and using tags instead of names. In addition, we introduce a basic query language at API level as it has benefits in terms of retrieving files and metadata management as well. So the proposed model represents the solution to the problems listed in Section 1 as we will present every component in full detail.

What follows is a detailed formal description (using Z notation) of the data model (low level); its associated operations to update the data model; and then the model queries that is the high level of the model.

### 4.1 Data Model

Collections are organised in a manner that represents a rooted graph. A collection may contain other collections which are called sub-collections. The terms **parent** and **child** naturally describe the relationship
between sub-collections and collections; more generally a unique path exist between any two collections.
Paths can be expressed as an ordered list of 0 or more items. There is a special collection which does not
have a parent, called root. some collections -as mentioned early- might have a plurality of membership, so in
this case they do have number of paths.

Collections are distinct from files. Files are simply associated with a collection in a belongs-to
relationship. The term ‘collection’ was chosen to explicitly distinguish the concept from the traditional
‘directory’ (or ‘folder’).

1. Graph
\[ G \subset \text{cid} \times \text{cid} \cup \{ \tau \} \]
- \( \text{cid} \) is the type of collection identifier.
- \( G \) describes a graph of collection identifiers with root \( \tau \).
- \( G(s) \) is the parent of \( s \).
- Initial value: \( G = \emptyset \)
- Constraint: \( \forall s \in \text{dom} G \bullet (s, \tau) \in G^+ \)
  All collection identifiers are part of a single rooted graph. This constraint also precludes cycles.

2. Collection tags
\[ S : \text{cid} \cup \{ \tau \} \rightarrow \text{ctag} \]
- \( \text{ctag} \) is the type of collection tags.
- The collection tag for \( \tau \) is the distinguished value root.
- Initial value: \( S = \{ \tau \rightarrow \text{root} \} \)
- Constraint: \( \forall (i, p), (j, q) \in H \bullet p = q \wedge i \neq j \Rightarrow S(i) \neq S(j) \)
  The collection tags of collection identifiers with the same parent must be distinct; collection tags
  are unique within collections.

3. Files
\[ F : \text{cid} \rightarrow (id \rightarrow \text{ftag}) \]
- Files are grouped in collections; each collection is identified by a collection identifier.
- Each file is a bidirectional mapping between file tag (type \( \text{ftag} \)) and a physical identifier (type \( \text{id} \)).
- Initial value: \( F = \emptyset \)
- Constraint: \( \forall (s_1, f_1), (s_2, f_2) \in F \bullet s_1 \neq s_2 \Leftrightarrow \text{dom} f_1 \cap \text{dom} f_2 = \emptyset \)
  A physical file may only be referenced within a single collection.

4. Collection path
The following two functions are derived from \( G \) and \( S \).
\[ P : \text{cid} \rightarrow \text{seq} \text{ctag} \]
\[ D : \text{cid} \rightarrow \text{seq} \text{cid} \]
- A collection path describes the node traversal sequence from the root node to a target collection.
  \( D(s) \) is the collection identifier path to \( s \) while \( P(s) \) is the collection name path.
- A path is defined in terms the function \( \text{up} \), which is the sequence from a node to the tree root.
  \[ \text{up} \tau = \langle \tau \rangle \]
  \[ \text{up} s = \langle s \rangle \cap \text{up}(H(s)) \]
  \[ D(s) = \text{rev}(\text{up}(s)) \]
  \[ P(s) = \{ (n, S(i)) \mid (n, i) \in D(s) \} \]

The syntax of paths is shown in Figure 3. CollPath is the collection path defined by the \( P \) function,
while a path to a file includes an appended file tag. Except for the mandated trailing slash, and the
absence of an unnamed root collection, this is identical to the POSIX style path syntax.
4.2 Operations of the VennTags Model

The possible operations for this model include the function calls below. Rather than precisely formal definitions, we describe their essential properties. Naturally, any user interface built on top of the API may have different operations that translate to these functions.

1. \texttt{CrCollection}(\texttt{newcollection, parent}) : To add a new collection, the precondition is that the collection does not exist. It has to provide the path where the new collection will be with the set of collection tags which must be unique.

   For instance, considering the scenario associated with a university course: these items have properties ‘course code’ and ‘year of offer’, we want to add a new collection \{reference,Git\} within /{2014,CS200} . To do so, all we need function call \texttt{CrCollection}({reference,Git},/{2014,CS200}) to get the target path. This operation returns \texttt{cid} if preconditions are met, or false if not.

2. \texttt{CrCollectionLink}(\texttt{collection, parent}) - create a link between a collection and another collection which refers to create a new membership for a collection. So the collection will have another parent- a new path. The precondition to complete this operation is that the collection and parent have to exist, and there is not link between the collection and the parent before. The semantic of this operation means adding a new path to the collection path to that collection- by allowing to have multiple paths. This operation returns true if preconditions are met, or false if not.

   For instance, suppose that the new collection \{reference,Git\} that we added in within /{2014,CS200} collection has to exist in /{2015,CS200} collection as well, so all we need \texttt{CrCollectionLink}({reference,Git},/{2015,CS200}) function call.

3. \texttt{DelCollection}(\texttt{collection, parent}) : refers to delete a collection. This can be done if the collection is empty which means that all it sub-collections and files have already been deleted (no sub-collection and files at all) and it has not to be linked to another collection as well. This operation returns true if preconditions are met, or false if not.

   For example, to delete \{2015,CS100\} collection within Courses collection, we need \texttt{DelCollection} (/{2015,CS100},/courses) call function which will be false as 2015 collection has subcollections.

4. \texttt{DelCollectionLink}({collection}, parent) : delete a collection link: refers to delete one possible path by deleting one membership of collection with other collection. To complete this operation all we need is the collection and the parent which have to exist and the collection has to have another link with another parent- it means the collection has to have more than one link. The semantic of deleting a link is that it refers to change the collection paths by deleting the collection membership from one path. This operation returns true if preconditions are met, or false if not.

   For Example, if we want to delete existing of \{reference,Git\},/{2015,CS200} collection in /{2015,CS200}, we just need to \texttt{DelCollectionLink}({reference,Git},/{2015,CS200}), so \{reference,Git\} collection will remove just from /{2015,CS200} so in this case one of \{reference,Git\} path will be cancelled.

5. \texttt{UpCollection}(operation, new value, old value) -Update a collection: the updating refers to change the tag value or the path of the particular collection. The input parameters are: operation that means to type of function call whether it is “move” the collection (changing its location) or, “add”, “delete”
Figure 4: VennTags query language

changing the set of associated collection tags; old value always refers to path (whether the operation move, add, or delete as the location of the collection needed in all these operations); and new value means to new path (location) if the operation is “move” while it is the new tag value or without value if the operation is “add or delete respectively” The old and new values will be checked where the old value has to exist and the new one has to not exist and it will not affect the locally uniqueness of the collection. This function means that all the sub-collections and files underneath this collection will be immediately changed as well. This function call returns true if preconditions are met, or false if not.

6. CreFile\(\left\{\text{new file}\right\},\text{parent}\) - creates a new file. The input parameters of this operation are a tag or set of tags and collection path where the file will be. The operation preconditions are that the file does not exist and the new tag (set/subset of tags) has to be locally unique.

7. DelFile\(\left\{\text{file}\right\},\text{parent}\) refers to delete a file from a collection with precondition that the file exists with providing its tags or part of tags which is uniquely identify the file and its collection as well. This operation returns true if preconditions are met, or false if not.

8. UpFile\(\left(\text{operation},\text{old value},\text{new value}\right)\) - Update File: changing a tag value or the path of a specific file with a precondition it has to not affect the uniqueness of the files within its collection. This is done by providing the input parameters that are: operation that means the type of this function call which is either “move”, “add”, “delete” tags; rename operation will be expressed by delete the old one and then add the new one; old value always refers to path (whether the operation move or rename as the location of the collection needed in both operations); and new value means to new path (location) if the operation is “move” while it is the new tag value if the operation is “rename”. The old and new values will be checked where the old value has to exist and the new one has to not exist and it will not affect the locally uniqueness of the file within the collection. This operation returns true if preconditions are met, or false if not.

4.3 Queries

VennTags model adds a query language to the file system API as mentioned early. Figure 4 shows the abstract syntax of the query language. It extends the path language, which is designed to identify a single file system object, by replacing the collection tag by a disjunctive list of tags and the file tag by a disjunctive list of file tags or their inverse. So for files, either the presence or absence of a tag can identify a file to include in the query result.

A query returns a set of files. All files immediately associated with, as well as recursively contained in, the collections identified by the query’s path, are returned in the set. The result set is homogeneous: at first glance it is not possible to determine which file has come from which collection. However, subsequent calls to lower-level operations can retrieve that information.

By adding the query language in the API level, problem 3 of the hierarchy - addressed in §2 - is avoided if a user adopts an orienteering-style search using a simple query while descending the hierarchy. The result at each level shows all files in the remaining subtree.

For example, the query \(/{2014,CS200} \land /{2015,CS200}\) identifies collection that exists in these both collections while \(/{2014,CS200} \lor /{2015,CS200}\) identifies all collections that in \(/{2014,CS200}\) or \(/{2015,CS200}\) and in both queries all files located within those collections will return recursively. An other possible quires:

In addition, complex queries are possible, for example:
Figure 5: University courses

```
courses/2014/CS400 ∧ CS100/ courses/2015/CS400 ∧ CS100/ reference ∨ Git
```

With concrete query syntax, to perform those complex queries, it would need addition of parentheses to resolve this case, but it has been omitted here for simplicity.

5 Evaluation

The evaluation will be in two parts as shown in the below:

5.1 How does VennTags solve HFS problems?

The VennTags file management system structure provides a solution to the problems detailed in section 1.

The provision of multiple tags for a collection (file container) offers a solution for problems 1& 2. In the case of the problem 1 (Artificial hierarchies), VennTags solves this problem as shown in Figure 5. About problem 2 (Classification), our model allows multiple classification schemes (problem 2) by two ways: one by adopting DAG where a collection can belong to more than one collection at the same time (having more than one parent), and using multi-tags for collection and files. So to solve this problem, the user can create a collection with set of tags that reflects the classification and then link the collection to the desired collections that meet the classification as shown in Figure 6. In addition, more visible collection tags can better inform the orienteering style of search that descends a rooted DAG to locate a file (problem 3). Finally, the ease of tag manipulation supports associating more relevant metadata with groups of files (problem 4).

Multiple file tags can also assist users in the latter two cases: file search and metadata management. Both these are potential benefits dependant to a significant extent on the development of appropriate user interfaces that can exploit the opportunities offered by multiple tags.

It should be noted that, while multiple tags give users better tools to organise consistent file hierarchies, the success any user has in doing so depends on their own ability to suitable tag and categorise the files that they create.

The insertion of a generic and powerful tag-based query in the API is a novel feature and one which, like some of the aspects of the tagging structure, depend on appropriate user interfaces to deliver real utility to users. In particular, the correspondence between a query result and the concept of a virtual directory or folder (Gifford et al., 1991) can lead to some real advances in GUI-based metadata management and query (Dekeyser et al., 2008). In particular, if a virtual directory is updatable (this corresponds to an updatable base view (Siberschatz, Korth, & Sudarshan, 2011)) a file can automatically acquire the metadata associated with (files in the) virtual directory. So from all these points, VennTags is powerful model that is a solution to the HFS problems.

5.2 Why rooted graph structure?

To prove that choosing DAG structure provided a more powerful model than other structures, we choose hierarchical (tree) model to compare with as it has been used for long time. So to prove that graph is more expressive than hierarchical (tree) model, consider a graph model where files exist in just one collection and both collections and files have a single tag.

\[ G \subseteq \text{sid} \times \text{sid} \cup \{\tau\} \]

Where \text{sid} is collection identifier.
Figure 6: Supporting multiple classifications

Figure 7: Graph collections

Consider a hierarchical model. Files just exist in one collection and both collections and files have a single tag.

\[ H : \text{sid} \rightarrow \text{sid} \cup \{\tau\} \]

**Theorem:** Graph is more expressive than hierarchical (tree) model.

**Proof:** We show by example that the graph model can exhibit some functionality that tree cannot. Observe that a graph model exhibits the following properties:

1. Preserve metadata.
2. A collection can have more than one parent, so to retrieve a file, there could be more than one path (as defined in section).
3. Metadata updated

For example: if we have a collection that contains other collections - subcollections. The subcollections also contain other subcollections where the files exist, as shown in the in the Figure 7. This figure shows that the collections ‘B’ and ‘C’ shared the same collection which ‘D’.

So the \( f_1 \) file has two paths:

\[ (A, B, D, X) \quad (A \ B \ C \ X) \]

The abstract view is

\[ [f_1, \{(A, B, D, X)\}, \{(A, C, D, X)\}] \]

If we try to transfer this example to the isomorphic of the tree model, there are alternative structures as shown in the Figure 8. We would arbitrarily choose one collection (B or C) in which to place the shared subcollection (D)(Figure 8 a and b), keep two duplicate copies (Figure 8 c). Both of these structures lead to less metadata and having just one path which mean that both do not meet all the three properties addressed earlier. Other structure could be met one of the key properties which is preserved metadata but non of the others (Figure 8 d). Some of those structures provide one correct path with less metadata while others preserve metadata but incorrect path. However, non structures could respect the multi-path property which considers the key property.

6 Related Work

As the amount of data stored on personal computers has grown with the limitations of the existing file systems in terms of organizing and retrieving data which are addressed in (§1), number of attempts has focused on finding solution for those limitations. We can categorise the observed attempts to solve the HFS limitations into three groups. These groups are: proposing a new class of file systems by replacing directory
Alternative hierarchy in favor of a more metadata-centric approach, so these systems often do not rely on a physically existing hierarchy; adding on the traditional file system that can be found as part common operating systems have been extended; or enhancing the HFS by doing some changes that can help improving HFS. The details of each group in the subsections below.

6.1 Replacing Hierarchal File Systems

Alternative post-hierarchical file system architectures have been proposed (Gifford et al., 1991; Dekeyser et al., 2008; A. Ames et al., 2005; S. Ames et al., 2013; Seltzer & Murphy, 2009; Rizzo, 2004; Padioleau, Sigonneau, & Ridoux, 2006) to avoid the problems posed by traditional file systems. The organisation and retrieval of files in the cited systems relies on a rich collection of file metadata rather than the hierarchical directory structure. These systems are designed to fully replace the HFS, though as yet none have succeeded in doing so. This is because it might represent a problem for end-users. This is because using a tree is an easy and simple way to classify objects. Users are familiar with tree structures and they can easily understand them. This is an important issue in acceptance and usage of the HFSs by users.

6.2 Extension- adding on HFS

Another approach to dealing with HFS limitations is to introduce extra functionality layered on top of the underlying existing file system. Many approaches involve the use of tags. Tags have been used effectively in social websites such as Flickr and YouTube (Furnas et al., 2006) which have introduced ways of organising multimedia and allowed users to associate tags with media items and then retrieve those items based on metadata (Livia & Ross, 2010).

TagTree (Voit, Andrews, & Slany, 2011) is an example; it takes user-supplied tags and automatically generates and maintains a navigation tree (folder) structure of tags. The system builds an extensive hierarchy such that multiple paths, each one an ordered permutation of the files’ set of tags, are generated for each file. This novel system is problematic when files have many tags, leading to exponential growth of the tag tree.

Others (Civan, Jones, Klasnja, & Bruce, 2008; Ma & Wiedenbeck, 2009; Schenk, Görlitz, & Staab, 2006; Bloehdorn & Völke, 2006; Lin et al., 2014; Sajedi, Afzali, & Zabardast, 2012) propose models to help...
users to organise their files based on supplied tags. However, the cited systems have not offered a query language so the users cannot easily re-find their files.

Our novel model VennTags does also utilise tags as the cited models - as mentioned earlier; however, VennTags differs from other proposed solutions, which are cited above, in many points. One of these points is that the users are allowed to attach the tags to their collections and files as well not just to files as other models. This promotes the utility of multiple tags in classifying, retrieving files, and manipulating files. The another point is that our model allows overlapping collections which solve some problems of HFS as shown in Section 5. The third point is that VennTags provides a query language built in API, so the users can easily re-find their files.

On the other hand there are proposals that also use DAG and tags such as CoFS (B.-H. Ngo, Silber-Chaussumier, & Bac, 2008; H. B. Ngo, Silber-Chaussumier, & Bac, 2008). However, in these models supplied tags are automatically generated and maintained in a DAG folder structure of tags. The major drawback in this system is that the difficulty of searching to retrieve files which required time for that as the system do not provide a query language for that. In addition, these system lack the metadata management operations to facilitate easy update group/subgroup files because the tags are automatically provided.

6.3 Enhancing HFS

Other proposals to solve HFS problems attempt to find a balance between the HFS replacement (Section 6.1) and the HFS add-on approaches (Section 6.2). TreeTags (Albadri et al., 2016) is an example; it is a modest variation to HFS semantics that adds tagging capability to the fundamental file system structure but retains the familiar and clearly useful hierarchical container structure of traditional file systems. However, VennTags is more powerful than Treetags as proved in Section 5.

FindFS (Chou, 2015) and TrueNames (Parker-Wood, Long, Miller, Rigaux, & Isaacson, 2014) are other examples that offer an enhancing for HFS problems. However, none of them have succeeded to solve all the HFS limitation, they just focus on just one problem and ignore the others.

Hence from the above subsections, it can be seen that different proposed solutions have been utilized tags and/or DAG but none have succeeded in doing so. These approaches can be grouped into four groups based on their drawbacks. The first group includes proposals that support multi-classification, but they are lack inbuilt query system that facilitates advanced use of such classifications. In the second group of attempts, metadata is managed to some extent by the file system, compared to systems where metadata manipulation is under complete user control. The third group contains systems that are only able to update information for one file at a time; that is, all metadata update operations are file-based (rather than updates at directory or collection level).

The members of the fourth group use novel non-traditional approaches; some of these may be more correctly seen as data or information management systems.

The above weakness of the cited attempts are avoided in our proposal VennTags as shown in section 4.

7 Conclusion

The main work in this paper is to introduce with formal description of a file management system structure that utilize the idea of overlapping sets as in Venn diagram and reuse tags but integrates it into the tried-and-trusted rooted graph paradigm. VennTags has been shown to resolve the identified HFS problems.

There are two broad directions that extend the current work. The first one is to continue investigating alternative models that also solve these problems, such as TreeTags (our previous work (Albadri et al., 2016)) but have symbolic/hard links without any limitations that exist in HFS links. We are currently evaluating a model that allows files to exist in multiple collections – links, in essence, but without the problems associated with managing them, and then comparing that with VennTags and show which one more powerful model.

The second direction for future work involves evaluating the VennTags model in a practical sense. A proof-of-concept implementation must make key decisions on data structures and algorithms; comparing the software with traditional file systems then requires the creation of a metadata-oriented benchmark that
could also be used to measure the efficacy of other novel file systems. Perhaps more importantly the user interface design space afforded by the richer metadata and query API of the VennTags file system needs to be explored, prototyped and experimentally evaluated.

References


doi: 10.1145/2611354.2611367


attempt? think to a new file manager for multiple categorization of your personal information. In 6th
ternational workshop on personal information management.

at the 3rd european semantic web conference (eswc).

on hot topics in operating systems.


(6th edition)

quality in e-health. Springer.
An In-Depth Analysis of Tags and Controlled Metadata for Book Search

Toine Bogers¹, Vivien Petras²

¹Department of Communication & Psychology, Aalborg University Copenhagen, Copenhagen, Denmark
²Berlin School of Library and Information Science, Humboldt-Universität zu Berlin, Berlin, Germany

Abstract
Book search for information needs that go beyond standard bibliographic data is far from a solved problem. Such complex information needs often cover a combination of different aspects, such as specific genres or plot elements, engagement or novelty. By design, subject information in controlled vocabularies is not always adequate in covering such complex needs, and social tags have been proposed as an alternative. In this paper we present a large-scale empirical comparison and in-depth analysis of the value of controlled vocabularies and tags for book retrieval using a test collection of over 2 million book records and over 330 real-world book information needs. We find that while tags and controlled vocabulary terms provide complementary performance, tags perform better overall. However, this is not due to a popularity effect; instead, tags are better at matching the language of regular users. Finally, we perform a detailed failure analysis and show, using tags and controlled vocabulary terms, that some request types are inherently more difficult to solve than others.

Keywords: book search, controlled vocabularies, social tagging, query analysis, failure analysis


Copyright: Copyright is held by the authors.
Contact: toine@hum.aau.dk, vivien.petras@ibi.hu-berlin.de

1 Introduction

To locate a scroll in the ancient library of Alexandria, searchers were required to know the genre and author name (Phillips, 2010). Asking a librarian or other scholars at the library was probably the preferred strategy for searchers then. The process of locating relevant or interesting books to read has become substantially easier over the years with the advent of sophisticated book recommender systems, such as those offered by Amazon, which uses purchase history to suggest interesting books, as well as book search engines, such as Google Books which allows for full-text search through millions of books¹.

However, full-text search is not always a practical option. Book metadata is often the only way of satisfying book search requests. Previous work has compared the benefits of different types of metadata elements, such as bibliographic metadata, controlled vocabulary² (CV) terms, and user-generated content (e.g., reviews and tags) and found that user-generated content is the most effective in retrieving relevant books (Koolen, 2014; (Bogers & Petras, 2015)). A second focus of our earlier work in (Bogers & Petras, 2015) was on the specific comparison between tags and CV terms. Tags were found to outperform CV terms significantly overall, but it was not a clear-cut advantage: some book search requests were better served by CV. Both metadata elements appear to provide complementary performance. We ascribed the overall advantage of tags to the popularity effect of tags—identical tags are repeatedly used by different users to describe the same book.

However, many unanswered questions remain: Is there really a true popularity effect for tags? Which types of book search requests are better addressed using tags and which using CV? Which book search requests fail completely for both metadata sources, and what characterizes such requests? In this paper, we provide a deeper understanding of the value of tags and CV for book retrieval by revisiting and updating our study, resulting in the following contributions:

1. A comparative analysis of tags and CV, focusing on complementarity and potential popularity effects.
2. A detailed analysis of book search requests that shows which types of information needs work better with tags or CV.

²In this paper, we use the term controlled vocabulary to denote any form of taxonomy, categorization or language-controlled terminology (e.g., subject headings) that prescribes the form or term for a certain concept that is described (Dextre Clarke, 2008).
3. A failure analysis to determine why certain book search requests succeed while others fail.

The structure of this paper is as follows. We start in Section 2 with an overview of the relevant related work. Section 3 describes the experimental methodology used in this study. Section 4 explains the results of our comparative analysis of tags and CV for fulfilling book search requests. Section 5 contains the results of our request analysis showing which requests are better fulfilled by tags or CV. Section 6 describes our analysis of successful and failed requests. Finally, Section 7 discusses the outcomes of this study and concludes with suggestions for future work.

2 Background

In this section, we briefly discuss the relevant related research on (1) book search and information needs, (2) retrieval using CV versus user-generated content, and (3) aspects of search request analysis.

2.1 Book Search and Information Needs

While full-text search of books has been relatively underrepresented in research (Willis & Efron, 2013), book search in library catalogs has received plenty of attention recently (Slone, 2000; Kim, Feild, & Cartright, 2012; Saarinen & Vakkari, 2013). Magdy and Darwish (2008) showed that for shorter queries using only metadata can be just as effective as full-text retrieval, highlighting the importance of our work.

The Social Book Search (SBS) workshops\(^3\) have been a fertile ground for research on book search since their inception in 2011, using metadata from Amazon, LibraryThing and library catalogs. The 2016 edition (Koolen et al., 2016) focused on the entire process of book search, from automatically detecting and categorizing book search requests to improving retrieval algorithms to investigating how people interact with book search interfaces.

The book search requests used in the SBS workshops were collected from the LibraryThing forums and represent complex, real-world information needs that are typically much longer and richer than the short queries submitted to conventional book search engines like Amazon or Google Books. Koolen, Bogers, Van den Bosch, and Kamps (2015) performed a detailed analysis of these book search requests and found different relevance aspects that went beyond the information present in traditional book metadata, such as novelty, engagement, and familiarity. Mikkonen and Vakkari (2016) found most book search requests for fiction were centered around familiarity and bibliographic information and an earlier study by Buchanan and McKay (2011) found that book requests are rooted in cultural context. Our study contributes to this body of work by providing a detailed analysis of which type of book search requests might be more effectively served by tags or by CV.

2.2 User-Generated Content vs. Controlled Vocabularies

The debate on the relative merits and drawbacks of controlled vocabularies versus free-text (including user-generated content) has been and continues to be a fruitful subject for small-scale case studies, with disagreement on which source is more effective (Cleverdon & Mills, 1963; Rowley, 1994; Dextre Clarke & Vernau, 2016). Recent, larger-scale work using the SBS collection has shown that user-generated content allows for more effective retrieval, with reviews being especially beneficial (Koolen (2014); (Bogers & Petras, 2015)).

Our more specific comparison of tags and CV in (Bogers & Petras, 2015) showed that while retrieval using tags yielded better results, CV and tags were successful for different search requests, a result also found in previous studies on retrieval using natural language vs. CVs (Gross & Taylor, 2005). We provide an in-depth analysis of the reasons behind these findings.

One possible reason for the complementary effects on retrieval performance might be the different characteristics of tags and CV. CV in the form of classes or categories from library classifications or even the Amazon taxonomy are very broad. Studies categorizing LibraryThing tags have found them to contain more subjective, contextual, and personal descriptions (Lawson, 2009; Voorbij, 2012), whereas subject headings tend to be more abstract and are required to be objective, impersonal, and only cover the most important topics of a book (LoC Cataloging Policy and Support Office, 2016).

---

\(^3\)See http://social-book-search.humanities.uva.nl/#/overview (last accessed September 11, 2016) or Koolen et al. (2013).
While some studies concluded that tags and subject headings are complementary (Smith, 2007; Bartley, 2009), other studies found that tags either cover the same topics as subject headings or simply provide a more expansive terminology for book search (Heymann & Garcia-Molina, 2009; Lu, Park, & Hu, 2010). These studies also demonstrated that tags contain self-referential terms, which might introduce noise for a search engine, and might not cover less popular books equally well.

2.3 Search Request Analysis

Research on search requests or query analysis has taken on three forms: (1) query classification, (2) failure analysis, and (3) difficulty prediction. Query classification focuses on determining which types of information needs users of a particular information system have. For instance, Koolen et al. (2015) classified a set of LibraryThing forum requests used in the SBS labs as well as in this study.

Failure analysis looks at why requests fail to retrieve relevant results on particular collections. For instance, a thorough failure analysis of TREC queries found that the reasons for failure are usually due to semantic relationships represented in the query that are not understood by the search system (Buckley, 2009).

Finally, difficulty prediction involves studying how difficult it will be to retrieve relevant documents for a specific search request. Analyses distinguish between pre-retrieval methods, based mostly on linguistic features of the requests, and post-retrieval methods, which analyze requests with respect to documents in the collection or retrieved set. A good summary for these approaches can be found in Carmel and Yom-Tov (2010).

In this study, we perform a failure analysis using some pre-retrieval difficulty prediction indicators such as request or document length to research which book search requests can be more effectively fulfilled by tags or CVs and why some requests are bound to fail.

3 Methodology

This section describes the book metadata that was searched, the book search requests, relevance assessments and evaluation measures we used for the analyses.

3.1 The Amazon/LibraryThing Book Collection

The Amazon/LibraryThing collection has been used for several years in the Social Book Search workshops 2. It was collected by Beckers, Fuhr, Pharo, Nordlie, and Fachry (2010) and contains over 2.8 million book records aggregated from Amazon, the British Library (BL), the Library of Congress (LoC), and LibraryThing (LT). Book records (henceforth referred to as ‘documents’) consist of over 40 different metadata elements, including core bibliographic metadata such as author or title, which were found to benefit retrieval (Bogers & Petras, 2015). In this paper, we only focus on the relative benefits of tags and CV; for an experimental comparison of the other metadata elements we refer the reader to Bogers and Petras (2015).

To put tags and CV on a more equal experimental footing and to be able to examine how they compare for individual documents, we filtered the original Amazon/LT collection so that all book records that did not contain at least one CV term and at least one tag were removed. This resulted in a test collection with 2,060,758 documents.

The tags in the Amazon/LT collection were originally collected from LibraryThing. They make up our Tags test collection 4. The CV terms come from three different providers: Amazon, BL, and LoC. We combined these sources into a single test collection called CV. Table 1 shows the metadata elements making up the CV and Tags test collections.

---

4A test collection is the entire collection of over 2 million filtered book records used for search, where the metadata just consists of the selected metadata element or elements, in this case: Tags.
<table>
<thead>
<tr>
<th>Metadata elements</th>
<th>Provider</th>
</tr>
</thead>
<tbody>
<tr>
<td>CV</td>
<td>Amazon</td>
</tr>
<tr>
<td>DDC class labels</td>
<td>Amazon</td>
</tr>
<tr>
<td>Subject headings</td>
<td>Amazon</td>
</tr>
<tr>
<td>Geographic names</td>
<td>Amazon</td>
</tr>
<tr>
<td>Category labels</td>
<td>Amazon</td>
</tr>
<tr>
<td>LCSH terms</td>
<td>BL, LoC</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Metadata elements</th>
<th>Provider</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tags</td>
<td>LT</td>
</tr>
</tbody>
</table>

**Table 1**: Metadata elements used in test collections and their respective providers.

**Table 2**: Test collections with different metadata elements.

In order to test the popularity effect of tags, we created an additional test collection **Unique tags**. This collection contains the same tags as in the **Tags** collection, but each tag is included only a single time in a book record.

To compare the complementarity effect of tags and CV terms, we also created two collections that are combinations of both sources: **Tags + CV** and **Unique tags + CV**. Table 2 contains an overview of the test collections that were used in this study.

3.2 Book Search Requests & Relevance Judgments

The book search requests that were used to search the collections were collected from LT discussion forums (Koolen, Kamps, & Kazai, 2012). Example requests include asking for (1) suggestions on books about a certain topic or from a particular genre; (2) ideas on books where the user can only remember plot details, but not the necessary metadata; and (3) recommendations based on individual preferences. Frequently, the requesters add books they have already read to their information need description. Figure 1 shows an example book request.

Book search requests taken from the LT forums consist of a title and a narrative, which were used in combination for searching in our experiments. Both can be considered realistic expressions of the information need. Section 5 describes these search requests in more detail.

The book suggestions posted as replies to the LT forum requests are regarded as relevant books for a request (Koolen et al., 2012). A graded relevance scale was used, based on additional criteria, such as whether the book had been added to the catalog of the requester or suggester(s) (Koolen, Kazai, Preminger, & Doucet, 2013). From the 2014 edition of the SBS workshop, we used 340 randomly selected topics for training and 334 topics for testing purposes, ensuring they were filtered to not include any of the ~800,000 book records that were filtered out.

3.3 Retrieval Setup & Evaluation

For our retrieval experiments, we used the Indri 5.4 toolkit and its language modeling implementation with Jelinek-Mercer smoothing, which was shown to perform better on longer queries, such as the LT forum requests (Zhai & Lafferty, 2004).

To optimize the search system’s performance for any of the five test collections, we used 340 randomly selected search requests and their relevant books for training to determine the system’s parameter settings for (i) the degree of smoothing, as represented by the $\lambda$ parameter, which controls the influence of the collection language model (varied in increments of 0.1, from 0.0 to 1.0); (ii) stopword filtering (none or using the SMART stopword list); and (iii) stemming (none or using the Krovetz stemmer). These optimal settings were then used on the 334 test book search requests to produce the results presented in the remainder of this paper.

---


6Available at http://sourceforge.net/projects/lemur/files/lemur/indri-5.4/

7This resulted in 44 different possible combinations of these three parameters, and $5 \times 44 = 220$ training runs in total. Readers interested in these optimal parameter settings are referred to http://toinebogers.com/?page_id=738 for a complete overview.
Figure 1: Book search request from the LibraryThing forums (re-used from (Bogers & Petras, 2015)).

We use NDCG@10 (Normalized Discounted Cumulated Gain cut off at rank 10, see Järvelin and Kekäläinen (2002)) to measure the search performance. This measure has also been used in previous SBS editions, making our work comparable. The single-figure NDCG@10 metric rewards result lists where highly relevant books are ranked higher.

We employ statistical significance testing when comparing the performance of different retrieval runs and use an $\alpha$ of 0.05. When comparing the performance of two different retrieval runs, we use two-tailed paired $t$-tests and also report the effect size (ES) and the 95% confidence interval (CI) as recommended by Sakai (2014).

4 A Comparative Analysis of Tags and Controlled Vocabularies

4.1 Main Results

**Question 1:** Is there a difference in performance between CV and Tags in retrieval?

**Answer:** Tags perform significantly better than CV. The combination of both sources in Tags + CV results in even better performance, but not significantly so.

Table 3 shows the main results of our five runs with Figure 2 representing the same information graphically. There is a statistically significant difference between the five runs according to a repeated-measures ANOVA with a Greenhouse–Geisser ($F(2.529, 842.144) = 4.650, p < .01$). We can see that Tags provide significantly better retrieval performance over our 334 requests compared to CV according to a two-tailed paired $t$-test ($t(333) = 2.171, p < .05, ES = 0.118, 95\% CI [0.0160, 0.0325]$). However, combining the two in Tags + CV results in even better performance, which suggests they are complementary to a degree. While this combination also significantly outperforms the original CV collection ($t(333) = 2.874, p < .05, ES = 0.157, 95\% CI [0.0069, 0.0368]$), Tags + CV does not perform significantly better than the Tags collection ($t(333) = 1.194, p = .253, ES = 0.066, 95\% CI [-0.0031, 0.1263]$).
Table 3: Results for the different test collections using NDCG@10 as evaluation metric. The best-performing run is marked in bold font.

<table>
<thead>
<tr>
<th>Metadata elements</th>
<th>NDCG@10</th>
</tr>
</thead>
<tbody>
<tr>
<td>CV</td>
<td>0.0348</td>
</tr>
<tr>
<td>Tags</td>
<td>0.0519</td>
</tr>
<tr>
<td>Unique tags</td>
<td>0.0524</td>
</tr>
<tr>
<td>Tags + CV</td>
<td>0.0566</td>
</tr>
<tr>
<td>Unique tags + CV</td>
<td>0.0583</td>
</tr>
</tbody>
</table>

Figure 2: Results for the different test collections using NDCG@10 as evaluation metric. Bars indicate average NDCG@10 scores over all 334 topics, with error bars in black.

### 4.2 Popularity Effect

**Question 2:** Do Tags outperform CV because of the so-called popularity effect?

**Answer:** No, the popularity effect does not seem to be the reason for this difference. The Unique tags test collection (without tag frequency information) performs even better than Tags (albeit not significantly so).

One possible explanation for the difference between Tags and CV is the so-called popularity effect in tagging systems as first described by Noll and Meinel (2007): popular books received more (and more of the same) tags than unpopular books, whereas CV terms are more evenly distributed across books. In our previous study (Bogers & Petras, 2015), we also put forward this phenomenon as a possible explanation for the performance difference, similar to Koolen (2014).

One component of this popularity effect is books receiving more unique tags than CV terms, which requires an examination of the type and token statistics of our collections. Table 4 shows type and token counts for the different collections, both as total counts and averages per document. It shows that books do not receive more unique tags (= types) per document than CV terms. In fact, the average number of types assigned per document for CV is nearly three times higher at 36.52 compared to 13.08 for Tags. The high number of types for CV can be explained by the aggregation of several controlled vocabulary metadata fields in one test collection, as explained by table 1.

<table>
<thead>
<tr>
<th>Metadata elements</th>
<th>#types</th>
<th>#tokens</th>
<th>avg. types/doc</th>
<th>avg. tokens/doc</th>
</tr>
</thead>
<tbody>
<tr>
<td>CV</td>
<td>2,208,694</td>
<td>109,793,695</td>
<td>36.52</td>
<td>53.3</td>
</tr>
<tr>
<td>Tags</td>
<td>2,272,393</td>
<td>246,313,480</td>
<td>13.08</td>
<td>119.5</td>
</tr>
<tr>
<td>Unique tags</td>
<td>2,272,393</td>
<td>47,253,002</td>
<td>13.08</td>
<td>22.9</td>
</tr>
<tr>
<td>Tags + CV</td>
<td>2,353,659</td>
<td>354,046,417</td>
<td>29.43</td>
<td>171.8</td>
</tr>
<tr>
<td>Unique tags + CV</td>
<td>2,353,659</td>
<td>154,985,939</td>
<td>29.43</td>
<td>75.2</td>
</tr>
</tbody>
</table>

Table 4: Type and token statistics for the five different metadata element sets.

The other aspect of the popularity effect is that the books receive more of the same tags, i.e., that tag frequency
plays an important role in the performance difference. Table 4 does show that the average number of tokens assigned to a document is more than twice as high for Tags at 119.5 than it is for CV at 53.5.

To further examine this, we created our Unique tags collection, removing tag frequency information from the Tags collection and only retaining single occurrences of each tag. If tag frequency is indeed the deciding factor, then Tags should perform better than Unique tags. However, the opposite is true: Unique tags achieves better performance than Tags, even though the difference is not statistically significant according to a two-sided paired-samples t-test ($t(333) = 0.139, p = .890, 95\% \text{ CI } [-0.0070, 0.0080]$). Moreover, Unique tags shows an even bigger, statistically significant performance increase over CV than Tags did ($t(333) = 2.135, p < .05 \ (0.033), = 0.117, 95\% \text{ CI } [0.0014, 0.0338]$). This strongly suggests that it is the quality of the tags themselves that makes the difference with CV instead of a popularity effect. On average, CV simply do not appear to match the user’s vocabulary as well.

4.3 Complementarity

**Question 3:** Do Tags, Unique tags, and CV complement or cancel each other out in terms of retrieval performance?

**Answer:** Tags, Unique tags, and CV complement each other: they are successful on different sets of requests.

The best-performing of our five runs is the Unique tags + CV collection. The success of this combination suggests that the two individual representations Unique tags and CV provide the best complementary performance. This is also confirmed by the per-request difference plots in Figure 3.

![Per-topic differences](image.png)

**Figure 3:** Differences in retrieval performance ordered by per-request difference between (a) the Unique tags and CV, and (b) the Unique tags and Tags collections. Bars above the horizontal axis represent requests where Unique tags perform better, bars below the horizontal axis represent requests where the other collections perform better.

It shows the per-request differences in NDCG@10 between (a) a search using Unique tags and CV; and (b) a search using Unique tags and Tags.

Figure 3a shows how many of the requests were better served by Unique tags (bars above the horizontal line) and how many requests were better served by retrieving using the CV (bars below the horizontal line). As the area above the horizontal axis is larger than the area below it, this again demonstrates that Unique tags show a
small advantage over CV. It also shows that there are different types of search requests: for most requests, searching either Unique tags or CV makes no difference, but for certain requests one of the two test collections outperforms the other. The pattern in Figure 3a confirms that Unique tags and CV indeed offer complementary performance.

We see an identical pattern for the difference between Unique tags and Tags in Figure 3b. This suggests that while, on average, tag frequency information hurts retrieval performance slightly more than it helps, there are also several requests where tag frequency information actually improves the ranking enough so that Tags outperforms Unique tags.

Despite their complementarity, the majority of requests shown in Figures 3a show no performance difference between Unique tags \(^8\) and CV. Most of these requests actually failed to retrieve any relevant documents at all: 247 out of 334 test topics (or 74.0\%) fail completely.

This leads us to two very interesting follow-up questions: (1) what is it in these Unique tags and CV collections that helps successfully retrieve relevant documents, and (2) what makes the overwhelming majority book search requests so difficult that both metadata elements fail completely at retrieving relevant documents? We attempt to answer these two questions in Sections 5 and 6 respectively.

5 Analysis of Book Search Requests

In the previous section we learned that despite overall performance differences, Unique tags and CV terms offer complementary performance. In this section, we take a closer look at 87 requests (or 26.0\%) that succeeded in at least one relevant document being retrieved in the top 10 results by Unique tags or CV. What makes certain representations better at satisfying some types of book search requests than others?

5.1 Relevance Aspects in Book Search Requests

**Question 4:** What types of book requests (in terms of what makes them relevant to users) are best served by the Unique tags and CV test collections?

**Answer:** CV show a tendency to work best for requests that touch upon aspects of engagement, whereas requests that focus on content-based, familiarity, known-item, or socio-cultural aspects are best served by Unique tags.

One way of categorizing book search requests is by the relevance aspects that are expressed in it: what aspects make a book relevant to the original poster? While some LT users are trying to re-find a book from their childhood with only vague plot points and memories of characters to go on, others express a desire for books that match a specific mood or provide a certain reading experience.

To analyze the difference between Unique tags and CV in terms of such relevance aspects expressed in the requests, we use the relevance aspects identified and annotated by Koolen et al. (2015) and inspired by Reuter (2007). They annotated a large set of SBS book requests (which include our 334 test requests as a subset) with one or more of a set of eight relevance aspects\(^9\). Table 5 contains brief descriptions of these eight relevance aspects.

It shows that among the successful requests, Content is the most common relevance aspect in 79.3\% of all 87 topics, followed by Familiarity and Metadata.

If we compare the search requests where one of the test collections outperforms the other by a margin of at least 120\%, then we see a clear difference in the distribution of aspects. Apart from Engagement, which is best served by CV representations, all other aspects are best satisfied with Unique tags.

Figure 4 shows how well the two test collections perform on requests of different types as measured by NDCG@10, and it shows largely the same pattern as Table 5. While all aspects except Engagement perform better when using Unique tags, the difference between Unique tags and CV is only statistically significant for Familiarity according to a two-tailed paired-samples t-test (t(35) = 2.268, p < .05, ES = 0.377, 95\% CI [0.0119, 0.2147]) and for Content (t(62) = 3.489, p < .005, ES = 0.440, 95\% CI [0.0489, 0.1800]).

When inspecting the relevant documents, it is easy to see why these patterns occur. A good example for the Content aspect is topic #63529 (“I just finished and enjoyed Climb the Wind by Pamela Sargent. Can anyone recommend other science fiction and or alternate history about Native Americans?”). Several of the relevant retrieved documents are

---

\(^8\)In the remainder of this paper we will use Unique tags as our collection representing tags, because they provide the best individual performance.

Table 5: Distribution of the relevance aspects over all 87 successful book requests (column 1), the requests where Unique tags outperform CV terms by 120% or more (column 2), and the requests where CV terms outperform Unique tags by 120% or more (column 3). More than one aspect can apply to a single book request, so numbers do not add up to 100%.

<table>
<thead>
<tr>
<th>Relevance aspect</th>
<th>Description</th>
<th>Requests overall (N = 87)</th>
<th>UniqueTags &gt; CV (N = 53)</th>
<th>CV &gt; UniqueTags (N = 27)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accessibility</td>
<td>Language, length, or level of difficulty of a book</td>
<td>9.2%</td>
<td>7.5%</td>
<td>11.1%</td>
</tr>
<tr>
<td>Content</td>
<td>Topic, plot, genre, style, or comprehensiveness</td>
<td>79.3%</td>
<td>83.0%</td>
<td>70.4%</td>
</tr>
<tr>
<td>Engagement</td>
<td>Fit a certain mood or interest, are considered high quality, or provide a certain reading experience</td>
<td>25.3%</td>
<td>22.6%</td>
<td>33.3%</td>
</tr>
<tr>
<td>Familiarity</td>
<td>Similar to known books or related to a previous experience</td>
<td>47.1%</td>
<td>49.1%</td>
<td>37.0%</td>
</tr>
<tr>
<td>Known-item</td>
<td>The user is trying to identify a known book, but cannot remember the metadata that would locate it</td>
<td>12.6%</td>
<td>17.0%</td>
<td>7.4%</td>
</tr>
<tr>
<td>Metadata</td>
<td>With a certain title or by a certain author or publisher, in a particular format, or certain year</td>
<td>23.0%</td>
<td>24.5%</td>
<td>14.8%</td>
</tr>
<tr>
<td>Novelty</td>
<td>Unusual or quirky, or containing novel content</td>
<td>3.4%</td>
<td>3.8%</td>
<td>0%</td>
</tr>
<tr>
<td>Socio-cultural</td>
<td>Related to the user’s socio-cultural background or values; popular or obscure</td>
<td>13.8%</td>
<td>15.1%</td>
<td>7.4%</td>
</tr>
</tbody>
</table>

Performance grouped by relevance aspect

Figure 4: Results for the Unique tags and CV test collections, grouped by the eight relevance aspects expressed in the 87 successful book search requests. Average NDCG@10 scores over all requests expressing a particular relevance aspect are shown in grey and as horizontal bars, with error bars in black.
only indexed with Science fiction in CV, but not with terms like alternate history and native americans, which are present in Unique tags and greatly improve their chances of being ranked at the top of the results list.

Known-item requests are by their very nature difficult to fulfill using CV as the user typically only remembers some vague plot elements and characters, which are not the most important topics that subject headings tend to cover. For example, in request #73796 (“I read this book 5 to 10 years ago. It was like Francine Rivers, but doesn’t seem to match any of her titles that I can find. It started with 3 older men of a small church searching for a new pastor and hiring a young man who seemed promising. The new pastor had great success but as the church grew into a mega church with building projects, etc, he strayed away from the Word.”), the tags church growth, pastor, mega churches are what rank the relevant document near the top. The more generic CV terms Church buildings and Clergy are not enough to provide effective retrieval.

Other aspects like Socio-cultural and Novelty are also more likely to be present as tags than as CV, resulting in their improved performance with Unique tags. Given typical indexing guidelines, one could perhaps expect that Accessibility and Content aspects would be covered well by CV. Figure 4 shows that this is not the case. Perhaps surprisingly, it is actually the Engagement topics that are better served on average by CV than by Unique tags. This difference, however, is not significant ($t(20) = 0.767$, $p = .452$, ES = 0.167, 95% CI [-0.1132, 0.0524]). An inspection of the Engagement requests revealed no CV or Unique tags terms related to reading engagement, suggesting that the difference is coincidental.

5.2 Book Type: Fiction vs. Non-fiction

**Question 5:** What types of book requests (in terms of fiction or non-fiction books that are requested) are best served by Unique tags or CV?

**Answer:** Unique tags work much better than CV for fiction book requests. CV show a tendency to work better for non-fiction book requests, but the difference is not significant.

Search requests typically ask for one of two types of books: fiction or non-fiction. For our previous study (Bogers & Petras, 2015), we annotated all 334 requests as for works of fiction or non-fiction. Fiction and non-fiction requests are unevenly distributed in our test set: the majority of the requests (75.3%) were for works of fiction.

![Performance grouped by book type](image)

**Figure 5:** Results for the Unique tags and CV test collection, grouped by type of book(s) requested (fiction or non-fiction). Average NDCG@10 scores over all requests for a particular book type are shown in grey and as horizontal bars, with error bars in black.

An analysis of the performance of Unique tags and CV with respect to the nature of the book(s) being requested (see Figure 5) shows that Unique tags are significantly better in serving requests for fiction books than CV ($t(58) = 3.571$, $p < .005$, ES = 0.465, 95% CI [0.0568, 0.2016]). While CV is better than Unique tags in serving non-fiction requests, this difference is not statistically significant ($t(27) = 1.194$, $p = .243$, ES = 0.226, 95% CI [-0.1699, 0.0449]).

What could be the reason for the large difference between Unique tags and CV for fiction requests? To explain this, we can consult the distribution of relevance aspects by the type of book(s) requested in Figure 6. Aspects that are more commonly expressed in requests for fiction are exactly those aspects that Unique tags tend to be better at solving. For example, the Known-item aspect occur more often in fiction requests: 41.0% of all fiction requests
cover this aspect versus 30.8% of non-fiction requests. As we saw in Figure 4 in the previous section, these are more likely to be solved by Unique tags, which explains (part of) the better performance on fiction requests.

![Figure 6: Distribution of relevance aspects by the type of book(s) requested (fiction (N = 256) vs. non-fiction (N = 78). Horizontal bars represent the percentage of all request of a particular book types that express a specific aspect. For example, 41.0% of all 256 fiction requests express a Known-item aspect.](image)

**Metadata** is another aspect that is more common in fiction requests at 21.5% versus 6.4% of non-fiction requests. Normally, requests for books from a certain author or from a specific publication year would be solved by core bibliographic metadata. However, in this pure comparison between Unique tags and CV the former performs better, because in free tagging schemes these metadata elements will inevitably be added as tags by some users, whereas CV will not describe them. Claiming this as an advantage for tags is unfair, however, as any normal digital library search engine would index such core bibliographic metadata regardless.

Perhaps counter-intuitively, aspects such as **Content**, **Engagement**, and **Accessibility** are more common in non-fiction requests. **Engagement** was the one aspect that was better addressed by CV (albeit not significantly), which could help explain the better performance of CV on non-fiction requests. Requests that express a **Content** aspect are also understandably tied to non-fiction requests, which commonly include elements like the topic and the degree of comprehensiveness. All of this suggests that the relative benefits of Unique tags and CV are strongly dependent on the types of book requests made and on the aspects of books relevant to the requester.

### 6 Failure Analysis

Despite the complementarity of Unique tags and CV, the majority of book requests fail: 74.0% of all 334 book search requests fail to retrieve any relevant books. Combining the two test collections in Unique tags + CV produces non-zero results for 7 more requests, due to improved ranking, but for 240 requests even Unique tags + CV fails to retrieve relevant documents in the top 10 results. In this section, we perform a failure analysis of these 240 requests: What kind of requests fail most frequently? What is the reason for this? And is this a problem that tagging and controlled vocabularies could ever be expected to solve?

#### 6.1 Sparsity, Recall Base, and Example Books

One common cause of poor performance in retrieval and recommendation systems is data sparsity: many algorithms breakdown when not provided with enough data. Sparsity could affect book retrieval in two different ways: (1) book search requests could be too short and thereby provide inadequate information to locate relevant documents, and (2) book metadata could be too short, making it difficult to match them against rich book search requests.
**Question 6:** Do failed book search requests fail because of data sparsity, a lower recall base, or a lack of examples?

**Answer:** Sparsity does not appear to be a reason for retrieval failure and neither is the size of the recall base. The number of examples provided by the requester does have a significant positive influence on performance.

Table 6 shows the average length of search requests and relevant documents for successful and failed requests. This suggests that the length of search requests is unlikely to be the underlying cause. Not only is there no significant difference between the two groups according to an independent-samples t-test \(t(332) = 0.907, p = .365, 95\% CI [9.915, 10.933]\), but the difference actually goes the other way as failed requests are longer on average than successful ones. Document length does not appear to be the reason either: the relevant documents for failed search requests are longer on average and statistically significantly so \(t(3889) = 6.257, p < .001, 95\% CI [-5.580, 0.892]\). Instead of sparsity, a possible cause could be that the retrieval algorithm is unable to distinguish well enough between important and unimportant terms and that request and document length exacerbate this problem.

<table>
<thead>
<tr>
<th></th>
<th>Avg. book search request length (in words)</th>
<th>Avg. relevant document length (in words)</th>
<th>Avg. no. of relevant documents</th>
<th>Avg. no. of example books provided</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Successful</strong></td>
<td>86.7</td>
<td>73.9</td>
<td>13.3</td>
<td>1.63</td>
</tr>
<tr>
<td><strong>Failed</strong></td>
<td>96.6</td>
<td>79.5</td>
<td>11.0</td>
<td>0.54</td>
</tr>
<tr>
<td><strong>Overall</strong></td>
<td>93.8</td>
<td>77.7</td>
<td>11.7</td>
<td>0.84</td>
</tr>
</tbody>
</table>

Table 6: Breakdown of book search requests by request length, length of the relevant documents, size of the recall base, and the number of examples provided by the original requester.

Another possible reason for retrieval success is the recall base: more relevant documents means it is relatively easier to return relevant documents in the results list. Again the numbers in Table 6 do not bear this out: the two groups have an average difference of only 2.3 documents, which is not statistically significant \(t(332) = 1.269, p = .205, 95\% CI [-2.301, 1.812]\).

Finally, the number of examples provided by the original requester may influence performance: with more relevant examples, the retrieval engine could potentially provide better results. This explanation appears to have some merit. There is a significant difference in the number of examples provided for successful and failed requests \(t(332) = 4.638, p < .001, 95\% CI [-1.098, 0.237]\), as shown in Table 6. There is also a weak positive correlation \(r = 0.175 (p < .005)\) between NDCG@10 score for Unique tags + CV and the number of provided examples.

### 6.2 Relevance Aspects and Book Types

**Question 7:** Do book search requests fail because of their relevance aspects?

**Answer:** No. The relevance aspects are distributed equally for successful and failed requests. Only **Accessibility** and **Metadata** related search requests seem to fail more often.

Figure 4 in Section 5 showed that some of the performance differences between search requests could be explained by the different relevance aspects that are expressed in them. For example, **Known-item** and **Metadata** requests achieve higher NDCG@10 scores than **Accessibility** and **Socio-cultural** requests. A possible explanation for the failed requests could be that they contain proportionately more of the difficult aspects. Table 7 and Figure 6 show the distribution of relevance aspects over the successful and failed requests.
Table 7: Tabular distribution of the relevance aspects over all 94 successful and 240 failed requests.

<table>
<thead>
<tr>
<th>Relevance aspect</th>
<th>Successful (N = 94)</th>
<th>Failed (N = 240)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accessibility</td>
<td>9.6%</td>
<td>15.4%</td>
</tr>
<tr>
<td>Content</td>
<td>79.8%</td>
<td>75.8%</td>
</tr>
<tr>
<td>Engagement</td>
<td>14.9%</td>
<td>12.5%</td>
</tr>
<tr>
<td>Familiarity</td>
<td>24.5%</td>
<td>21.3%</td>
</tr>
<tr>
<td>Known-item</td>
<td>45.7%</td>
<td>35.8%</td>
</tr>
<tr>
<td>Metadata</td>
<td>13.8%</td>
<td>19.6%</td>
</tr>
<tr>
<td>Novelty</td>
<td>24.5%</td>
<td>22.5%</td>
</tr>
<tr>
<td>Socio-cultural</td>
<td>4.3%</td>
<td>3.3%</td>
</tr>
</tbody>
</table>

Figure 7: Visual distribution of the relevance aspects over all 94 successful and 240 failed requests.

The two distributions for successful and failed requests do not differ greatly from one another. Accessibility and Metadata occur in greater proportion among failed requests, suggesting that these are among the harder types to solve. While the 13 Metadata requests that successfully retrieve relevant documents tend to achieve good results, the majority of Metadata requests (N = 47) actually fail to retrieve any relevant documents.

The same holds for the Known-item requests: while 42 requests are successful, more than double that number of requests—86 in total—fail to rank any relevant documents in the top 10 results. This suggests that even for Unique tags, these request types are far from a solved problem.

Question 7: Does the type of book that is being requested (fiction vs. non-fiction) have an influence on whether requests succeed or fail?

Answer: Requests for works of fiction result significantly more often in failed book search requests.

Finally, another way of categorizing our book search requests is by the type of book requested: fiction or non-fiction. Of the 240 failed requests, 193 (or 80.4%) were for fiction, whereas only 63 out of 94 successful requests (or 67.0%) were for fiction. This difference is statistically significant according to a Chi-square test ($\chi^2(1) = 6.771, p < .01$) and shows that fiction requests tend to be harder to solve.

7 Discussion & Conclusions

In this paper we have presented a large-scale empirical comparison and in-depth analysis of the performance of controlled vocabulary vs. social tagging metadata for book search. Using a large collection of book records and book search requests that go beyond simple bibliographic queries, we showed that tags offer a richer vocabulary for answering complex book search requests than CV terms in general, but that the performance of the two representations is complementary. We also provided compelling evidence against the often suggested popularity effect of tags as the reason for their superior performance. Instead, it is the quality of matches between user-provided tags and search requests that results in better performance. A detailed request analysis showed that the relative performance of tags and CV appears to be dependent on the type of request, both in terms of the relevance aspects expressed in them and the types of works being requested. These factors appear to be more predictive of which representation performs best.

Finally, a comparative analysis of successful and failed search requests showed that addressing complex search requests using book search engines is far from a solved problem as most requests fail to retrieve relevant documents. Especially fiction book search requests are hard to fulfill, probably because requesters are asking for more aspects (engagement, familiarity, plot details) that are less covered by tags and possibly not at all by CV. We posit from
these results that existing indexing practices for books will have to change if complex book search requests are to have a chance of being met. Plot details could be added by harvesting the full-text of books for keywords. Indeed, Amazon already adds character and place names from its books to some of its controlled metadata. A detailed genre classification, a reading speed or engagement level estimate are highly subjective data points, which is why controlled vocabularies have avoided to use them for indexing. However, this type of information is often exactly what searchers are looking for. A critical look at existing subject indexing guidelines is required if we want book search engines to solve all book-related information needs and not just the simple ones.

For future work, we are considering developing a predictive model of query difficulty that takes all of the relevant factors into account. Another interesting avenue of research would be to examine the completeness of the current set of relevance judgments: currently, these are restricted by the suggestions made by other users, but an inspection of the data suggests these may be incomplete.

References


Table of Figures

<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Book search request from the LibraryThing forums</td>
</tr>
<tr>
<td>2</td>
<td>Results for the different test collections using NDCG@10 as evaluation metric</td>
</tr>
<tr>
<td>3</td>
<td>Differences in retrieval performance between two pairs of collections</td>
</tr>
<tr>
<td>4</td>
<td>Results for Unique tags and CV, grouped by relevance aspects expressed the book requests</td>
</tr>
<tr>
<td>5</td>
<td>Results for Unique tags and CV, grouped by book type</td>
</tr>
<tr>
<td>6</td>
<td>Distribution of relevance aspects by the type of book(s) requested</td>
</tr>
<tr>
<td>7</td>
<td>Distribution of the relevance aspects over all successful and failed requests</td>
</tr>
</tbody>
</table>

Table of Tables

<table>
<thead>
<tr>
<th>Table</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Metadata elements used in test collections and their respective providers</td>
</tr>
<tr>
<td>2</td>
<td>Test collections with different metadata elements</td>
</tr>
<tr>
<td>3</td>
<td>Results for the different test collections and their combinations using as evaluation metric</td>
</tr>
</tbody>
</table>
Table 4 Type and token statistics for the five different metadata element sets. . . . . . . . . . . . . . 20
Table 5 Distribution of the relevance aspects over all 87 successful book requests . . . . . . . . . . . 23
Table 6 Breakdown of book search requests by length, recall base and example count . . . . . . . . . 26
Table 7 Distribution of the relevance aspects over all successful and failed requests . . . . . . . . . 27
Enrichment of Cross-Lingual Information on Chinese Genealogical Linked Data

Hang Dong

1University of Liverpool
2Xi’an Jiaotong-Liverpool University

Abstract
With the emergence of non-English Linked Datasets, discrepancy in language has become a major obstacle for cross-lingual access of resources in the Semantic Web. To prevent non-English monolingual Linked Datasets to form ‘islands’ in the Web of Data, it is suggested to enrich a further layer of multilingual information on the Linked Open Data cloud. In the domain of culture heritage, enriching cross-lingual information can enhance the multilingual retrieval of cultural heritage resources, and promote international communication in the field. In this article, methods to enrich cross-lingual information for Linked Data are summarized, with a review on the cultural heritage domain. The mobile App Demo, Learn Chinese Surnames, winning the Shanghai Library Open Data Application Development Contest on 2016, is then introduced as a case study, to present the practice of enriching English-described information on a Chinese genealogical Linked Dataset, through consuming multilingual sources in the Linked Open Data cloud. Further in the data validation and conclusion, the issues of data quality and experience of consuming Linked Data are summarized.

Keywords: Linked Data; Cross-lingual information; Chinese genealogies; Open data applications; Consuming Linked Data


Contact: hangdong@liverpool.ac.uk

1 Introduction
Since the last decade, there has been an enormous growth in the amount of information on the Semantic Web, marked by the initiative of Linked Data. The Linked Open Data cloud (LOD cloud) has increased from a dozens of dataset to a large data space containing a thousand of datasets today (Schmachtenberg, Bizer, & Paulheim, 2014), forming a Web of Data. Linked Data refers to a set of best practices to publish and connect data, organized by RDF triples, via de-referenceable URIs on the Semantic Web (Bizer, Heath, & Berners-Lee, 2009). The vast Linked Datasets cover domains in the media, government, academics, user-generated content (Schmachtenberg et al., 2014). Linked Data initiatives and projects have also begun in the libraries (Singer, 2010) and in the cultural heritage domain (Oomen, Baltussen, & Van Erp, 2012).

Although the majority of datasets in LOD cloud are described in English, leading to a language-biased Semantic web, it is witnessed that many non-English Linked datasets have been published. For example, Dogmazic1, the French open music vocabularies, GeoLinkedData.es2, the open initiative of creating Spanish geospatial dataset, and the Shanghai Library Genealogical Linked dataset3, etc. These non-English datasets are rather valuable since they represent the diverse culture dependent on the languages and geographical areas. It is however that, with the issue of language discrepancies and cross-lingual access, these monolingual non-English Linked datasets could form “islands” on the Semantic Web (Gracia et al., 2012, p. 64). Similar

1http://www.dogmazic.net/
2https://datahub.io/dataset/geolinkeddata
3RDF structure in http://gen.library.sh.cn:8080/ontology/view, data content in http://jp.library.sh.cn/jp/home/index
issues about multi-lingual and multi-cultural are also the main challenges for cultural heritage data (Hyvonen, 2012, p. 5).

Currently, in the cultural heritage domain, multilingual Linked Data projects are still rare. It is believed in this article that, through the enrichment of cross-lingual information on current linked dataset in the cultural heritage domain, it is possible to enhance the multilingual access of cultural heritage information, boost international communication on researching these cultural heritage resources, and develop more international public data services.

Based on the principles of Linked Data and the ontologies adapted from BIBFRAME, Shanghai Library transformed and described the family books from MARC and those recorded in the General Catalogue of Chinese Genealogy to RDF triples (Xia, Liu, Zhang, & Zhu, 2014). The new genealogy data service platform of Shanghai Library supports multiple interfaces to consume data, including a web-based portal, a SPARQL endpoint and JSON-LD (Xia, Liu, Chen, & Zhang, 2016). In 2016, Shanghai Library held an Open Data Application Development Contest lasting about two months' time till the end of May, which is the first of this kind organized by Chinese libraries. A case study, an Android App Demo named Learn Chinese Surnames, is introduced to provide experience on the enrichment of information described in English from Linked Data sources, on the Chinese Genealogical dataset.

In this case study, the designer presents the methods to match entities in three famous Linked Dataset, DBpedia, Wiktionary and GeoNames, to the Genealogical dataset from Shanghai Library, to add a new layer of English-described information for cross-lingual access of Chinese genealogical information. This not only can enhance the international cooperation on research of Chinese genealogies, but also can benefit a wider user group of data services from native users to international users. From the Semantic Web point of view, the work constitutes an effort to realize the Multilingual Web of Data (Gracia et al., 2012).

In Section 2, the idea of enriching cross-lingual information for Linked Data is explained and related projects in the cultural heritage domain are reviewed. In Section 3, the Android App Demo, Learn Chinese Surnames, is presented as a case study to share the practice of consuming LOD to enrich English-described information, followed by a discussion on data quality and validation of Linked Data sources. Conclusion and future studies are in the Section 4.

2 Research Background

2.0.1 Enrichment of Cross-Lingual Information for Linked Data

The idea of enrichment of cross-lingual information for Linked Data is originally from the “multilingual Web of Data”, coined by Gracia et al. (2012). The “multilingual Web of Data” has been proposed as an enhancement of the current Linked Open Data cloud with an extra layer of resources and services that boost and improve the internal linking and multilingual access of the Web of Data. Potential resources and services are listed below (Gracia et al., 2012).

1. multilingual linguistic information to describe resources in different natural languages;
2. multilingual mapping between ontologies/vocabularies that establish cross-lingual connections and between entities/instances in different natural languages;
3. (semi-)automatic services for accessing and traversing Linked Data across languages dynamically, such as multilingual Linked Data generation, ontology localization and translation, automatic cross-lingual matching.

In this proposed article, adding the first two resources (1 and 2), multilingual linguistic information and multilingual entity mapping or linking, is summarized as the task of enrichment of cross-lingual information on Linked Data. The reason for this definition is a division of micro-level and macro-level: resources 1 and 2 are on a micro-level, thus easy to be implemented in individual applications, while services such as general tools for ontology localization are on a macro-level.

5http://pcrc.library.sh.cn/zt/opendata/
It is also necessary to distinguish the level of Linked Data to enrich: on the entity level (“instance level”), and on the ontology level (vocabulary level or “conceptual level”) (Gracia et al., 2012, p. 68). A group of studies in cross-lingual ontology matching deals with the mapping of ontological vocabularies in different languages (Trojahn, Fu, Zamazal, & Ritze, 2014; Mejía, Montiel-Ponsoda, de Cea, & Gómez-Pérez, 2012). Besides, it is possible to add multilingual information on the entity-instance level, for example, state that the page “Liu”\(^6\) in the DBpedia is the same as “[chinese character]” and “[chinese character]” in the Genealogical Linked Dataset from Shanghai Library, regardless of the ontological structure used to organize these entities in each Linked Dataset.

While the two levels of data enrichment are different, they share some common types of methods (methods on the ontology level in Trojahn et al. (2014), methods on the entity level in examples from Section 2.2): First, through translation by human or by machine to derive a new vocabulary or an entity description; second, through matching or linking of vocabularies or entities in two data sources; third, through a crowdsourcing interface enabling users to add multilingual information to the Linked Dataset.

All methods have their advantages and disadvantages. Current machine translation techniques are still preparing themselves to be intelligent enough to carry the intent behind languages and cultures, while manual translation effort needs much human effort. Through data matching based on semantic similarity, no translation effort is required, but the precision and recall of matching largely depend on the comprehensiveness of available semantic sources; also, without adding linguistic description such as synonym and hyponym, simple string matching can produce poor results (Pazienza & Stellato, 2006), so complex metrics to measure similarity or advanced methods based on machine learning have been used in large cross-lingual matching projects such as the alignment of English WordNet to Chinese HowNet (Ngai, Carpuat, & Fung, 2002), and the semantic linking of online collaborative encyclopedia, Chinese Baidu Baike\(^7\) and English Wikipedia (Wang et al., 2013). Through crowdsourcing, represented by Wikipedia and DBpedia (Lehmann et al., 2015), it is possible to obtain more accurate cross-lingual information based on the actual meaning that reflects culture background of resources, but the participation of a group of expert users is required. Further examples of the three methods are given in the culture heritage domain.

2.1 Enrichment of Cross-Lingual Information in the Cultural Heritage Domain

In the cultural heritage domain, due to the variety of representations and the commonality of cultures in the world, it is necessary to integrate resources of multi-culture to satisfy people’s information needs; while at the same time, resources in different cultures are sometimes originally represented in different languages, without an enrichment of cross-lingual information it would be hard to link them to each other, making these resources inaccessible to users in other languages. Currently, there are several practices for adding cross-lingual vocabularies and descriptions to cultural heritage resources, but few of them have published a Linked Data version.

An example on the enrichment of cross-lingual information on the vocabulary level is in the Digital Archives Sub-Project of Antiquities in the National Palace Museum\(^8\) under Taiwan e-Learning and Digital Archives Program (TELDAP)\(^9\). To satisfy the users’ needs to retrieve Chinese art resources using English, it is necessary to align the controlled vocabularies from the National Palace Museum (NPM) in Taiwan with the Art & Architecture Thesaurus (AAT) developed by the Getty Research Institute in US (Chen & Chen, 2012). The challenge of this task is the heterogeneity in the “conceptual structures” (the manner that a concept is involved in the hierarchical or associative relationships) (Chen & Chen, 2012, p. 285) of two controlled vocabularies, due to the discrepancy in language and culture. Pure manual effort was made to map the complex conceptual structures of different degrees of similarity in the two vocabularies. This mapping of two controlled vocabularies of different language and culture can enhance their interoperability, enable multilingual search, integration and sharing of cultural heritage resources among users in Chinese and English language background.

Instead of the enriching cross-lingual information on the vocabulary level, a US academic library has led a project on enriching cross-lingual information with a focus on the entity level. English-described

\(^6\)http://dbpedia.org/page/Liu
\(^7\)http://baike.baidu.com/
\(^8\)http://www.npm.gov.tw/digital/index2_2_8_en.html
metadata to the Collection of Chinese Scrolls and Fan Paintings has been added to ensure their retrieval in English (Matusiak, Meng, Barczyk, & Shih, 2015). First, the project group adapted the Dublin Core template to define a metadata scheme that incorporates the unique features for Chinese art resources and the bilingual fields. Then, human translation was carried out to generate the English described metadata of fields such as Main Text, Other Text, Seal Content, Subject, Coverage, etc. Compared to machine translation, human translation is suggested being more appropriate to capture the special cultural and linguistic features for the Chinese art resources (Matusiak et al., 2015).

A multilingual digital heritage project based on Linked Data structure is MOLTO10 (Damova, Dannélls, Enache, Mateva, & Ranta, 2014). The project enables using natural language in 15 European languages to search museum artifacts. This is realized by conversions from natural language to SPARQL and from RDF triples to natural language outputs in multiple languages. For the generation of multilingual descriptions of artifacts, manual translation was carried out for the vocabularies in the ontology and the entities or instances in the classes Material and Colour. Instances of the classes Painter and painting titles was, however, untranslated due to the unavailability of lexicons and long time for human translation. The translation of museum names was derived through mapping of the article names to Wikipedia, which achieved 90% correctness for the 5 major languages French, Italian, German, Russian and Spanish.

To give a background for the case study in Section 3, Chinese traditional family books are also an important cultural heritage resource for its special value on research in history and sinology. Enriching cross-lingual information to Chinese genealogical resources, can benefit their retrieval and access in an international environment. It is however that, the only available multilingual retrieval system for family books is FamilySearch.org, operated by The Church of Jesus Christ of Latter-day Saints, previously known as the Genealogical Society of Utah. FamilySearch is the biggest genealogy project in the world, and its digital Chinese Collection of Genealogies11 can date back from the year of 1239, cooperated with Shanghai Library in China from 201212. The general FamilySearch project uses a crowdsourcing interface with a friendly tutorial webpage13 to encourage users from the world to participate in the indexing, such as tasks to recognize texts from an old family book image, of multilingual genealogy resources. Also, FamilySearch has a Research Wiki portal14, enabling users to collaboratively edit articles about genealogies in different countries and cultures.

To sum up, in the domain of cultural heritage, due to the relatively narrow vocabularies in the ontology, and the lack of mature, easy-to-use techniques to translate cultural-dependent texts, researchers tend to enrich cross-lingual information manually, while for large projects, crowdsourcing methods are adopted. Multilingual Linked Data projects in the cultural heritage domain are still rare, although it is believed that they can enhance the interoperability, integration and searching of cultural heritage resources for a wider community (Hyvonen, 2012, p. 8).

3 Case study: Learn Chinese Surnames

In this study, a method is presented to enrich cross-lingual information through consuming multilingual resources in the Linked Open Data Cloud. A use case of mobile data services based on cultural heritage Linked Data published by a public library is provided. The work, an Android App Demo named Learn Chinese Surnames, winning the Shanghai Library Open Data Application Development Contest in 2016, is used as a case study.

3.1 Project Information and Data Use

Learn Chinese Surnames is an Android App designed for non-native Chinese learners to get familiar with Chinese surname culture and surname characters. The App made use of the 400 most popular modern Chinese surnames released in April 201315, and presented their origin, related people, metadata of early

10http://museum.ontotext.com/
11https://familysearch.org/search/collection/1787988
13https://familysearch.org/indexing/
15https://zh.wikipedia.org/wiki/%E4%B8%AD%E5%9B%BD%E5%A7%93%E6%B0%8F%E6%8E%92%E5%90%8D#2013.E5.B9.B44.E6.9C.88 [Webpage in Chinese showing 400 surnames, originally from the book Yida, Yuan & Jiuru, Qiu. (2013). Zhong guo si bai da xing ([chinese character]). Nan chang: Jiang xi ren min chu ban she. WorldCat Catalog:
family books and information about Chinese characters, etc. Users can browse the 400 surnames by their alphabetic ranking or by popularity, learn the stroke and meaning of Chinese surname characters, the origin of surnames as well as the knowledge about family books.

The task of enriching cross-lingual information is required, since English descriptions related to Chinese characters and family books are necessary for the targeted non-native Chinese users of this App. The figure (Figure 1) listed all information demonstrated in the App. The open data from Shanghai Library already created some data described in English or as numbers, displayed in colour black, including Pinyin of surnames, metadata of family books, location where the book is stored, etc. The other information, which are acquired through consuming the external Linked Data sources or other online resources, are displayed in red, including the address where the family book is compiled, introduction, origin and notable people of a surname, meaning and stroke of Chinese surname characters, etc.

![Figure 1: Information demonstrated in the App, Learn Chinese Surnames](image)

Apart from the genealogical dataset from Shanghai Library, three other Linked Open Datasets have been used in this App, namely, DBpedia (extracting information from Wikipedia), Wiktionary and GeoNames. The table (Table 1) presents the data source, the part of information used, method of consumption, usage in the App and copyright notice for each dataset.

The interface of this App is presented in the figure (Figure 2) to show the usage of newly enriched English descriptions for Chinese genealogical data. The figure (Figure 2(a)) displays the Wiktionary and Wikipedia URL links corresponding to the surname Su ("Chinese character"), while the figure (Figure 2(b)) displays the translation of geographical place data to English in the Chinese genealogical dataset (from "[chinese character]" to “Xiuning Xian, Anhui China”). GIF Images of Simplified and Traditional Chinese Characters were called at runtime using URLs within the domain of WrittenChinese.Com\(^{16}\).

### 3.2 Consuming Linked Data to Enrich Cross-Lingual Information

Three pieces of information in English were enriched in through consuming three Linked Open Data Sources.

- GeoNames is used to obtain the English-described information about the place of compilation of Chinese family books:

\[\text{http://www.worldcat.org/title/zhong-guo-si-bai-da-xing/oclc/910234509}\]

\[^{16}\text{https://www.writtenchinese.com/}\]
Table 1: Linked Open Datasets Used in the App, Learn Chinese Surnames

<table>
<thead>
<tr>
<th>Linked Data Source</th>
<th>Information Used</th>
<th>Method of Data Consumption</th>
<th>Usage in the App</th>
<th>Copyright Notice</th>
</tr>
</thead>
<tbody>
<tr>
<td>Shanghai Library Ge-</td>
<td>Metadata of family books (Title, year of compilation, place of compilation, location (@en)), pinyin and number of ancestors related to surnames</td>
<td>Real-time calling data through Restful service supported by SPARQL Endpoint</td>
<td>Display texts on App</td>
<td>(1) CC BY-NC-SA 2.0; (2) Right to use the data during contest</td>
</tr>
<tr>
<td>nealogical Linked Data</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DBpedia and Wikipedia</td>
<td>Articles in English about Chinese Surnames</td>
<td>Offline querying data through SPARQL Endpoint</td>
<td>Use the whole mobile webpage within App</td>
<td>(1) CC BY-SA 3.0; (2) GNU Free Documentation License</td>
</tr>
<tr>
<td>Wiktionary</td>
<td>Articles in English about a Chinese Character</td>
<td>Real-time calling data from URL</td>
<td>Use the whole mobile webpage within App</td>
<td>CC BY-SA 3.0</td>
</tr>
<tr>
<td>GeoNames</td>
<td>Chinese Geographical location information described in English</td>
<td>Real-time calling data through official API</td>
<td>Display texts on App</td>
<td>CC BY-SA 3.0</td>
</tr>
</tbody>
</table>

- DBpedia, which extracts information from Wikipedia, is used to get the information about Chinese surnames;
- Wiktionary is used to obtain the information corresponding to a surname character.

The methods used in this project to consume these data are various, including direct access with URLs, querying the SPARQL endpoint with Restful service or calling the official API. Data can be cached locally, as it is done for DBpedia, or called online at runtime, as applied on the Chinese Genealogical data and on GeoNames. The methods are chosen for the ease of implementation, and due to access limitations in terms of speed and connection, quality of data, and data needs for the App. A brief note on the methods to consume the Linked Data sources is in the table (Table 1). The figure (Figure 3) illustrates the path of consuming Linked Open Data sources to enrich Shanghai Library Genealogical Data.

3.2.1 Enrichment of geographical data of genealogy compilation places: through API interface of GeoNames

The GeoNames dataset\(^{17}\) contains over 10 million names and 9 million unique features, such as population and alternate names, of geographical places in the world. Geographical names are multilingual, thus can enable users to get the English names of places where their official languages are not English. Users may edit, correct and add new names using a wiki interface. The GeoNames ontology is available through datahub\(^{18}\). GeoNames dataset can be fully downloaded, queried using a third party SPARQL endpoint powered by FactForge\(^{19}\), or requested using the official API from GeoNames.

Considering the speed of connection and storage issues, the official API is used to retrieve GeoNames data at runtime. As shown in the figure (Figure 3), it is possible to use the hierarchical data (Country-Province-City-County) obtained using SPARQL endpoint from Shanghai Library to get a JSON format output using GeoNames API. Key parameters to construct the URL for API are featureCode, name and country.

To obtain GeoNames data of Hu Xian of City Xi’an in Province Shanxi in China (“[chinese character]”), which is the place where the family book “Duan Shi Shi Xi” (“[chinese character]”) is compiled on 1731, the URL to query API is constructed as follow, \(\text{http://api.geonames.org/searchJSON}\)

---

\(^{17}\)http://www.geonames.org/about.html

\(^{18}\)https://datahub.io/dataset/geonames-semantic-web

\(^{19}\)http://factforge.net/
(a) Wiktionary and Wikipedia Links to a Character and its Corresponding Surname

There are 195 family books for 苏 (su), where 493 names are recorded.

See traditional 苏 (su) and early family books in the next page.

(b) Metadata in English of Earliest Chinese Family Books Corresponding to a Surname

Figure 2: Interface that Displays Data in the App, Learn Chinese Surname

?name_equals=%E6%88%B7%E5%8E%BF&featureCode=ADM3&country=CN&maxRows=10&username=XXX, where featureCode value ADM3 corresponds to the third-order administrative division; country corresponds to the abbreviation of the Country to query; username is a registered ID on GeoNames website; name_equal is a complete string match value and %E6%88%B7%E5%8E%BF is the percent-encoding in URI of the Chinese character “????”. The JSON output displays the name of this place with its longitude, latitude and population.

```
{"totalResultsCount":1,"geonames":[{"adminCode1":"26","lng":"108.58764","geonameId":1806562,"toponymName":"Hu Xian","countryId":"1814991","fcl":"A","population":556377,"countryCode":"CN","name":"Hu Xian","fcodeName":"third-order administrative division","adminName1":"Shaanxi","lat":"33.99969","fcode":"ADM3"}]}
```

If multiple results are returned due to the fact that some Chinese location names on the city or county level along can map to several geographical places, the province level name (adminName1 in the JSON output) can be further used to select the exact one. For places in Taiwan, the country input should be TW according to the GeoName database. Through the method above so far, no wrong or missed matches are discovered, which means that both the precision and recall nearly reach 100%.

3.2.2 Enrichment of surname culture information: through SPARQL endpoint of DBpedia

The DBpedia project\(^\text{20}\) extracts structured information from Wikipedia based on crowdsourcing, making it a cross-domain, community-based, constant evolving and multilingual knowledge base. DBpedia has been localized to versions of 125 languages describing 38.3 million things. Serving as the core Linked Data source on the Web, the knowledge base covers 3 billion pieces of RDF triples and connects to other Linked Datasets by around 50 million RDF links.

\(^\text{20}\)http://wiki.dbpedia.org/about
For this case study, DBpedia and Wikipedia are ideal sources since they contain information about Chinese surname and its culture in English. In the RDF of DBpedia, foaf:isPrimaryTopicOf links a thing in DBPedia with a page URL of Wikipedia. Although DBpedia organizes resources with standard URIs, due to the inconsistency of vocabularies or fields among pages in DBpedia and the inconsistency among page names, it is not suggested to conduct batch processing of information in articles of Chinese surnames. In the case study, a semi-automatic approach is used, combining SPARQL queries with human decision, to achieve an acceptable data matching.

A good knowledge on the ontological vocabularies of DBpedia is vital for querying data with SPARQL. Through the dct:subject relation, the resources can be refined to a very specific domain; through dbo:abstract, the corresponding abstracts of articles in multiple languages can be extracted; by foaf:isPrimaryTopicOf, the corresponding Wikipedia URL can be obtained. The two examples below show the extraction of Wikipedia URL of Chinese surname "[chinese \ character]' (Simplified\Traditional Chinese) using abstract to match Chinese characters or using URL to match pinyin. The number of output is set as the most relevant URL.

Below is an example SPARQL query of obtaining the most relevant Wikipedia URL from a Chinese character "[chinese \ character]'". The measure of relevance is through the number of times the Character appears in the abstract of an article; the bigger the value of number, the more relevant this abstract is. The Wikipedia URL linked to the most relevant abstract is output as the result.

```sparql
# DBpedia SPARQL Endpoint
# Method 1: Matching strings of Chinese characters to the abstract field in DBpedia.
select distinct ?url ?ex count(?res) as ?count
where{
  filter(contains(str(?a), "??") || contains(str(?a), "?w").
  optional {?res dbo:wikiPageExternalLink ?ex}
}
order by desc(?count)
limit 1
```
Below is an example SPARQL query of obtaining the most relevant Wikipedia URL from a Chinese pinyin “Zeng” of Character “[chinese character]”. The measure of relevance is through the length of matched string to the Chinese pinyin; exact matching produces the lowest length. The matched Wikipedia URL which has the lowest length is output as result.

```
# DBpedia SPARQL Endpoint
# Method 2: Matching pinyin of Chinese characters to page URL in DBpedia
select distinct ?u
where{
  filter (contains(str(?m), ’Zeng’)).
  ?m foaf:isPrimaryTopicOf ?u.
}
order by asc(fn:string-length(?u))
limit 1
```

Through a final manual matching mediated by the two methods above, it is possible to achieve 100% precision and recall of extracting a Wikipedia URL from a Chinese character.

### 3.2.3 Enrichment of character related information: through URLs of Wiktionary

The Wiktionary project\(^{21}\) creates a multilingual dictionary of “all words in all languages” through the collaborative contribution of volunteers online; the freely open dictionary is available in 172 languages\(^ {22}\). DBpedia has extracted RDF triples from Wiktionary and opened the SPARQL endpoint as a side project\(^ {23}\), however, currently only RDF graphs of 4 languages have been fully created, not including Chinese. Wiktionary has also its official MediaWiki API\(^ {24}\), but considering the inconsistency of structures among Wiktionary articles, the API is not used in the project of this case study.

In the case study, URLs of Wiktionary are directly constructed and used to get and display the HTML information on the Android App. Unlike Wikipedia, the URL of Wiktionary are more standard, thus enabling direct construction of the Wiktionary URL with any Chinese character as an input. For example, the Wiktionary URI for Chinese character “[chinese character]” is \(\text{https://en.wiktionary.org/wiki/%E5%88%98}\), where “\%E5%88%98” is the percent-encoding in URI of “[chinese character]”. Apart from Wiktionary, another possible multilingual dictionary Linked Data source that can be used for enriching cross-lingual information is BabelNet\(^ {25}\).

### 3.3 Data Validation

Errors and conflicts are common in datasets. When it comes to matching data in different sources, it is necessary to validate the data and solve the conflicts originated from different structures and semantics. In this case study, comprehensiveness and conflicts of data are measured.

For data comprehensiveness, over the 400 most popular modern Chinese surnames, 377 of them (94.25%) are included in the Shanghai Library Genealogical dataset, and 295 of the 400 surnames (73.75%) have a corresponding article in English Wikipedia. Based on the 377 surnames in the Shanghai Library dataset, 93 of them (24.67%) do not have a corresponding article in English Wikipedia. This result can be later used to add new entities in these data sources.

For data conflicts, through the matching of Wikipedia articles to Shanghai Library dataset, 5 conflicts on the entity level were discovered in pinyin or English translation of Chinese surnames, as listed in the table (Table 2). There are also some slight internal conflicts on the forms and suffixes of page

\(^{21}\)https://www.wiktionary.org/
\(^{22}\)https://en.wikipedia.org/wiki/Wiktionary
\(^{23}\)http://wiki.dbpedia.org/wiktionary-rdf-extraction
\(^{24}\)https://en.wiktionary.org/w/api.php
\(^{25}\)https://datahub.io/dataset/babelnet
names in URLs of DBpedia and Wikipedia. For example, surname “[Chinese character]” corresponds to Liu; “[Chinese character]” corresponds to [Chinese character] with a tone in the Pinyin; “[Chinese character] corresponds to Bo_(Chinese_surname) with the suffix (Chinese_surname); “[Chinese character]” corresponds to Pang_(surname) with a different suffix (surname). This inconsistency is common and caused some effort on consuming the data.

<table>
<thead>
<tr>
<th>Surname Characters</th>
<th>Pinyin in Shanghai Lib Data</th>
<th>English Translation of Surnames in Wikipedia</th>
<th>URLs of the Wikipedia articles</th>
<th>Notes</th>
</tr>
</thead>
<tbody>
<tr>
<td>??</td>
<td>fang</td>
<td>pang</td>
<td><a href="http://en.wikipedia.org/wiki/Pang_(surname)">http://en.wikipedia.org/wiki/Pang_(surname)</a></td>
<td>Both applicable in different situations</td>
</tr>
<tr>
<td>??</td>
<td>qu</td>
<td>ou</td>
<td><a href="http://en.wikipedia.org/wiki/Ou_(surname)">http://en.wikipedia.org/wiki/Ou_(surname)</a></td>
<td>Should be “ou”</td>
</tr>
<tr>
<td>??</td>
<td>wei</td>
<td>ngai</td>
<td><a href="http://en.wikipedia.org/wiki/Ngai_(surname)">http://en.wikipedia.org/wiki/Ngai_(surname)</a></td>
<td>Both applicable; “ngai” is special for Cantonese</td>
</tr>
</tbody>
</table>

Table 2: Conflicts of English Translations of Chinese Surnames between Two Data Sources

In addition, on the ontological vocabulary level, internal inconsistency exists among the pages in DBpedia: the internal structure or the list of fields among surnames does not follow a strict standard. For example, the ontological vocabularies in DBpedia pages for “[Chinese character]”27 and “[Chinese character]”28 are very different, due to the nature of crowdsourcing and the level of popularity among different surnames. Similar issues were also found in Wiktionary. This issue has made it hard to process the semantics of surnames on a higher granularity using the vocabularies, for example, currently it is not possible to extract all notable people related to all Chinese surnames using one SPARQL query.

4 Conclusion and Future Research

A comprehensive multilingual Web of Data is a realizable dream that could boost the international interoperability of semantic resources on the Web, based on general services on the macro-level and small or middle applications on the micro-level, as the one in the case study. The overall practice of Linked Data for many organizations is to construct a worldwide Web of Data, and offer advantages such as promoting data re-use and enhancing the discoverability and interoperability of resources on the semantic level. It is however that, without adequate resources or services to overcome language barriers in the Semantic Web, data networking across language and culture would not be possible.

In this research, the author recapitulated the idea of enriching cross-lingual information of Linked Data to create a Multilingual Web of Data. The task of enriching cross-lingual information for Linked Data has been defined as adding a layer of information on the entity level or vocabulary level to the existing infrastructure, the Linked Open Data cloud. On the entity level, the layer of cross-lingual information can contain (a) descriptions that express the entity in a different language, or (b) URLs or links to an entity described in a different language.

Digital heritage resources in places where English is not a native language, are facing barriers of languages on the Web. Language barriers could impede the accessibility of cultural heritage resources and thus international cooperation in the field. Among recent digital heritage projects that aim at enriching cross-lingual information, manual effort including crowdsourcing, has been mostly used to derive the translations.

---

26 DBpedia URLs are http://dbpedia.org/page/PAGE_NAME, using the same page names as Wikipedia URLs https://en.wikipedia.org/wiki/THE_SAME_PAGE_NAME
27 http://dbpedia.org/page/D%C7%92ng
28 http://dbpedia.org/page/Liu
which highly depend on culture. It is also discovered that there are only few multilingual Linked Data projects in the cultural heritage domain.

For some cultural heritage resources, as the one in this case study, matching to existing multilingual data sources is a better choice, which can reduce the human effort to re-define the same entity in another language. The major part of this article, therefore, introduced the case study as a practice of enriching English-described information for Chinese Genealogical Linked Data through consuming multilingual resources in the Linked Open Data cloud.

Throughout the project, the designer consumed three multilingual semantic resources, DBpedia, Wiktionary and GeoNames, with the approaches that call the API, SPARQL, URL of these resources. The issue of data quality, including the comprehensiveness and conflicts among the three datasets, is considered to make appropriate choices for methods to consume data. The variable choices of interface is also a key for the development of open data, for example, all the four Linked Datasets used in this study (Table 1) have multiple data consumption interfaces available online, no matter official or third-party.

For future studies in terms of the project in the case study, first, it is expected to extract other cross-lingual information with higher semantic granularity, for example the relationship among all notable people of a Chinese surname and their relationships in English Wikipedia, through intelligent data matching techniques. Second, the comprehensiveness of dataset could be enhanced: it is possible to use the Genealogical data from Shanghai Library to supplement data in English Wikipedia; and use the GIS information in GeoNames to enhance the visualization of Chinese family books. Third, for copyright concerns on the use of data, such as embedding webpages and images in a mobile App through URLs, future surveys should be conducted before formally releasing the App to the public.

For a wider picture, it is worth discussing the impact of enriching cross-lingual information on digital heritage projects: how it addresses the information needs of users in the area related to a type of cultural heritage resource, and how it would benefit the cultural heritage organizations. In addition, technically, for any other multilingual digital heritage or digital library projects, the enrichment of cross-lingual information may only be a first step: Further issues are awaiting to be addressed, including the representation, query and automatic services for multilingual information on Linked Data.

References


Re-intermediation in the Fashion Industry: A Qualitative Study on Brokers in the Dongdae-mun Fashion District

Min-Joon Kim¹, Nuri Na¹, Joong-Seek Lee¹
¹Seoul National University, Korea

Abstract
Traditionally, middlemen had a large role in physical markets, serving as intermediaries by evaluating and distributing products. However, the rise of online markets and intermediaries have diminished their roles. This study investigates the change of intermediaries' roles in the presence of infomediaries, and the conditions that necessitate the re-introduction of the middlemen as “re-intermediaries”. We observed a group of brokers who work in the Dongdae-mun (DDM) fashion district in Seoul, Korea through multiple qualitative research methods including observations, contextual inquiries, and in-depth interviews. Our findings show that the reliability and depth of information relayed by human sources, along with the subjective nature of fashion have contributed to the brokers playing a major role again, despite infomediaries. The DDM fashion district relies heavily on the brokers' evaluations on trends, fashion, and product popularity, in addition to their traditional role of distributing goods in a quick manner.

Keywords: Intermediary; Infomediary; Human Agent; Re-intermediary; Human Factors in Information


Copyright: Copyright is held by the authors.

Acknowledgements: If there is an acknowledgement add it here. This field is optional. For review purpose, do not add information to this field until after the paper is accepted.

Contact: jseo@snu.ac.kr

1 Introduction
The concept of ‘intermediary’, which refers to a third party that offers intermediation services between two parties, has been a necessity in many fields for a long time. It has been noted by business historians that the entrance of the professional trader at the end of the Middle Ages was an important driving force for the development of the society(Heilbroner, 1962). These traders have now become real-estate agents, car dealers, and fund managers; the intermediaries in our society(Chircu & Kauffman, 1999). In the marketing domain intermediaries are commonly known as brokers, who act as functional middlemen. The marketing literature identifies “functional middlemen”, whose principle business is “specializing in performance of one or more specific marketing tasks, especially those concerned with negotiation. Their compensation is in the form of a commission or fee for a service rendered, NOT a profit on the sale of goods.”(Marketing Channels, 1982). At least some form of intermediary contact is used in the process of providing nearly every good and service in the economy(Hackett, 1992). For this reason, intermediaries are very important players in the market. Both consumers and producers benefit heavily from the roles of middlemen, who ensure that there is a seamless flow of goods in the market by matching supply and demand(Bailey & Bakos, 1997). Buyers also gain from the services of intermediaries in forms of promotion and delivery(Olsson, Gadde, & Hulthen, 2013). The role of intermediaries, however, have been reduced and altered significantly with the advance of information technology(Howells, 2006).

The Internet has brought the middleman to the digital space(Ordanini & Pol, 2001). A new word combining ‘information’ and ‘intermediaries’ has been invented to properly capture its role and place. Fisher defines “information intermediaries or infomediaries” as who are “concerned with enabling access to information from multiple sources and engaged in informing, aggregating, compiling, and signaling information.”(Fisher, 2010). Traditional intermediaries were simply human agents who took on the task of collecting and dispersing information and products relevant to the market(Snehota & Gadde, 2001). Infomediaries are able to do this on a much broader scale with high efficiency: infomediaries are machine, after all(Malone, Yates, & Benjamin, 1987). Compared to traditional intermediaries, infomediaries have a huge advantage in reachability and accessibility. For example, local used-car dealerships cannot simply compete with infomediaries like cars.com.
The increasing information accessibility of the general public is pushing intermediaries obsolete (Gellman, 1996).

Despite these trends, at the Dongdae-mun (DDM) Fashion District in South Korea, offline traditional intermediaries absorb the infomediary role and become highly specialized to the level where they become a necessity for the market to function and thrive. Technology and the Web are replacing intermediaries, but these human agents have found a way to adapt by taking on additional tasks like gatekeeping and coaching on top of traditional intermediary roles. In order to further investigate this new human-infomediary we use observations, contextual inquiries, and in-depth interviews to address the following questions:

- RQ1. How did the roles of human agents change with the arrival of infomediaries?
- RQ2. What are the information characteristics and areas of human intermediaries that infomediaries are not able to replicate?

The rest of the paper is organized in six sections. In Section 2, we provide further explanation of the DDM Fashion District, the human-infomediary that we will call as brokers, and the relation that this human-infomediary has with the market and potential customers. Based on these observations, we detail our research methodology and approach in Section 3. Findings and results from the research are presented in Section 4, followed by the discussion of our contributions and expandable issues in Section 5. Lastly, our work is summarized and concluded in Section 6.

2 Background

2.1 The Dongdae-mun Fashion District and Korean Fast Fashion

The Dongdae-mun (DDM) Fashion District can be traced back to 1905, where local shops took in and modified uniforms and blankets from US G.I.s in the outskirts of Gwangjang Market. The assembly of the Pyunghwa Market in 1961 laid the first stone of the fashion district, enabling the market to grow in size and quality. The building of the Art Plaza in 1990, and the following construction of the Milliore Building in 1998 has pushed the market to even further heights. The market now is a booming center for all things clothing, from wholesale to retail businesses (Hong, 2007). This was possible because the DDM Fashion Market naturally integrated a steady production infrastructure as well. Fabric, subsidiary materials, and sewing machines are all present inside the fashion district, and it is this link that has allowed DDM to be what it is today (Lim, 2010). With over 30 million wholesale stores producing multiple products in batches, DDM is considered to be the place for fashion design, production, merchandising, and distribution in Korea (Choi, 2013).

The closely-knit relation between wholesale stores and fabric/sewing factories allows the fashion district to speed up its production cycle. While it can be said that DDM’s multi-item batch production is similar to fast fashion, the nature of the two are vastly different (Lim, 2010). Represented by brands such as ZARA and H&M, fast fashion is spear-headed by big companies that design, produce, and distribute small amounts of trendy clothing over a short period of time. In contrast, wholesale merchants inside the DDM Fashion District are mostly consisted of four to five people and utilizes a Quick Response (QR) system: When a designed product falls out of favor, production is stopped immediately and production becomes geared towards more popular clothing (Lee, 2010). For example, if a dress worn by world-famous actress Song Hye-gyo in the Korean drama “Descendants of the Sun” is gaining popularity, multiple versions of the dress will be on display the very next day in the DDM Fashion District. Merchants and designers are able to get real-time feedback from customers, and production is always flexible and fast-paced (Park, 2012).

The incorporation of internet technology has further accelerated the already-fast manufacturing process. Internet shopping is huge in Korea. According to a market research in 2013, Korea’s online market boasts 35.7 billion dollars in market share and is continually growing (Kim, 2013). There are orders to be made and delivered, and with Korea’s busy cultural repertoire (e.g. the “hurry hurry” mindset) and exceptionally accessible geological characteristics (a small nation, where traveling from one end to another takes only eleven hours), there are expectations in parcel delivery time (two business days) that must be met (Kang, 2010). At the DDM fashion district, brokers make sure that deliveries from big wholesale stores to smaller retail stores are carried out almost instantly.
2.2 Human Agents and Infomediaries

The definition of intermediaries are especially covered in the field of marketing and business. While the roles of intermediaries have changed over the course of history, the more traditional roles can be boiled down to three main dynamics of traders, distributors, and providers (Snehota & Gadde, 2001). Initially, intermediaries were big as traders. The lack of specialties of other occupations required intermediaries to be a jack-of-all-trades professional (Heilbroner, 1962). Traders were able to connect customers to products, and products to customers by the sharing of information (Ramírez, Parthasarathy, & Gordon, 2013). This intermediary bridges the gap between markets, facilitating interaction among its members (Bayer, Geissler, & Roberts, 2011). This used to be the primary role of the intermediary, until the subsequent development of production technologies that came with the industrial revolution pushed intermediaries from traders to distributors and providers.

Intermediaries acting as distributors during and after the industrial revolution helped producers push the sudden uptick of production to customers. As more and more products were readily available, it was up to the intermediary to push the products so that the market would generate a flow (Malone et al., 1987). Intermediaries were able to serve multiple manufacturers, and some would start focusing on certain products that was better suited to their customers. This manufacturer’s perspective of the middleman is what previous studies mainly mention about intermediaries (Olsson et al., 2013). The reverse would also be applicable, with the intermediary delivering the wants and needs of the customers to the manufacturers. The manufacturers would then again provide the goods that the customers have demanded. A cycle was created and maintained by intermediaries (Malone et al., 1987). This too, however, would change with the a new wave of technology.

Technology and the changing of times have definitely forced intermediaries to change in order to survive. It can be argued that the most drastic changes have come with the rise of information technology (IT) in the 1900s (Hagel & Singer, 1999). A new digital intermediary called infomediaries came into being armed with a vast information pool that would surely outperform traditional middlemen (Bakos, 1998). IT allowed the customer to be able to contact the producer directly, establishing a direct contact which left out the middleman and the services they brought (Vandermerwe, 1999). IT also introduced a better control of material flows, further diminishing the roles of intermediaries (Hagel & Singer, 1999). Bleak consequences were observed, with even some scholars arguing that “many intermediaries will die out” (Pitt, Berthon, & Berthon, 1999). This process was even coined the term dis-intermediation by Gellman (Gellman, 1996).

However intermediaries not only did not disappear, but also have took on more specified roles. Infomediaries and intermediaries would exist in harmony, unlike the unpromising predictions of many scholars. This phenomenon is equally observed in the DDM Fashion District. Here, wholesale and retail merchants have access to an infomediary known as the “New Products Market”. This is where all newly designed clothing are uploaded for ease of viewing and ordering. Once an order is placed through this website, it is up to intermediaries to deliver the products between retail and wholesale stores. It may seem as a simple job and that the intermediary’s role has been reduced, but as the investigation shows below that is not the case. It turns out that intermediaries have a strategic and important role within the market such that the fashion district may not function without them.

3 Research Method

The investigation was carried out over a three-month time period in three distinctive stages, starting from April to June of 2016. The participants of this study are working or had experience working as a broker at the DDM Fashion District. After an initial observation period of two weeks, contextual inquiry interviews were conducted with the research team following a broker throughout his workday. Final in-depth interviews were conducted as the last step. A detailed description of each stage is as below.

3.1 Observations

A two-session observation period was held over two weeks at the DDM Fashion District. Observations were carried out in order to have a better idea of how the fashion district operated. Each session lasted over the entire work “night” (DDM brokers work schedule starts late evening, around 8pm and finishes around 9am), minus the last two hours where the main workload is finished and delivery packages are prepared. The
first observation session was focused on the fashion district itself, with the people surveyed having included brokers as well as retail and wholesale merchants, as shown in the figures below. The second observation session was geared towards the merchants. Observation sessions were conducted in a non-intrusive manner.

Figure 1: Obersvation Photos - Conducted in a non-intrusive manner, the researchers did on-the-field research on the DDM fashion district.

3.2 Contextual Inquiry

We conducted two contextual inquiry (CI) interviews with two current brokers. CI interviews were conducted in order to better understand the daily routines of brokers, and to find out first-hand the process of information exchange. Due to the nature of the brokering job, it was only possible to conduct contextual inquiries with one person at a time. Each CI lasted over the entire work schedule, which ranges from eight to ten hours. Each CI session was conducted by following a broker over the course of the interview. Both researchers participated fully as an apprentice to the broker, and took part fully in ordering, delivering, and conversing tasks as shown in Figure 2. Permission was granted to follow and film the broker during his work routine, and the footage was then analyzed and transcribed.

Figure 2: Contextual Inquiry Photos - Researchers dressed in appropriate fashion and followed a broker throughout the market.

3.3 In-depth Interviews

A total of six individuals were available for in-depth interviews. The interviews were conducted to obtain detailed information that we could not gather from observations and CIs, and to hear about their perspective
directly. Due to the rather spontaneous nature of the brokers’ work schedule, we were able to conduct three face-to-face interviews after hours in groups of two. Four individuals currently work as a broker at the DDM fashion district, and the remaining two have had prior experience working as a DDM broker. The interviews conducted were based on semi-structured guidelines, and were recorded under the participants’ agreement. The recorded interviews were then transcribed.

4 Findings

Our findings show that brokers who work in the DDM fashion district have a unique work schedule, with the use of digital devices versus pen and paper strictly divided between the time and kind of task. This divide proves to be a deciding factor in the brokers’ dual-role as an intermediary and infomediary. For better understanding, we first start off with providing a detailed view on what a typical workday for a DDM broker is like.

4.1 Work Routine

The typical work schedule of a broker at the DDM fashion district starts in the evening at around 8pm. Brokers check into their office and fire up their computers. As shown in Figure 3, the first task for them is to check their email and KakaoTalk\textsuperscript{1} accounts for orders. Orders are mostly based off of items listed in the “New Arrivals Market”. Experienced brokers organize orders by grouping them by wholesale dealers while at the same time considering their route around the fashion district. Unfortunately for the newcomers, most of them figure out their routes for collecting by trial and error. The phones are busy as the brokers start calling wholesale merchants to relay the order sheet. Jargons and expressions unique to the fashion district is a must when placing orders. After a row with the phones and emails, the brokers are ready to head out to the street. They take their printed-out order sheet along with pen and paper and head out to retrieve the goods.

Figure 3: Work routine of a broker - Brokers take in and organize orders in their offices before going out to retrieve the goods. After taking care of the returns and orders, the products are delivered to the retailers, all before the end of the night.

“I do use computers, but it’s mainly before I go out on my run. I capture parts of the orders from the retailers and send them to the respective wholesale merchants using KakaoTalk. I’m basically telling them to prepare the items so I can grab them quickly. This is because it’s so difficult to do so when I’m out running and retrieving the goods.” <P01>

Before they head on out and tackle the orders, they must deal with the enormous amount of returns. As mentioned earlier, the fashion district runs on the fast-paced quick response (QR) system, and many retailers end up returning a good amount of the orders they placed a few days ago. In order to avoid confusion,

\textsuperscript{1}KakaoTalk is a free mobile Instant messaging application for smartphones with free text and free call features. It is the most popular instant messaging application in Korea.
the brokers tend to take care of the returns first, and then proceed to their daily order list. If the number of returns is too high, returning and retrieving orders are done simultaneously; they must finish their order list before sunrise, when the orders are loaded on to trucks and delivered to retailers before the start of the work day.

Brokers at the DDM fashion district have to take care on average a thousand orders per shift. These orders are spread out over the entire fashion district and it is all too common for brokers to climb up and down tall buildings and even retrace their paths. The sheer amount of orders and complicated navigation of the market requires brokers to be smart and diligent, while also having the strength to take on the daunting size of orders — A single order may even weight up to 10 kilograms. For this reason, it is extremely rare to see women working as brokers in the DDM.

“At first, I only ran through the lists in whatever order they appeared on my sheet. Now that I have experience, I can see the list of wholesalers and somewhat calculate the shortest path I can take. I also take into account the number of items from each store, so I don’t overload and slow myself down.” <P02>

Because the need of both hands is critical, digital devices such as phones and tablets are usually put away. In contrast to popular belief, it is much more cumbersome for the brokers to work while recording on their phones. This is why the use of pen and paper is preferred by brokers. Things are done and recorded in their way instead of a standardized digital input. They incorporate wireless headsets to free up their hands during their runs, and their use of digital devices remain strictly to phone calls — activities that do not require the use of hands. After each building is cleared of orders, brokers take the goods to a centralized pick-up location and head off to tackle the next batch of orders. It is interesting to note that no monetary transactions happen when the brokers pick up the products from the store. Costs are taken care of after the day is over, and little to none is recorded, with verbal exchanges being the majority of confirmations.

“It’s quite difficult for me if I have to take or give I.O.U.s during my shifts. So I try to push and pull the merchants a bit, depending on what their personalities are. Sometimes I’ve got to be stern, and sometimes I’ll let some things slide.” <P03>

“First, I need all my hands, and second, it’s really hectic as it is. I’ve tried calling a few times using a Bluetooth headset, but other than that it is really difficult to use a digital device when doing work. I’ve got to trust my instinct and experience.” <P02>

It is through these runs where the exchange of information takes place. When returning items, brokers give feedback to the merchants on the clothes’ quality, design, price, and so on. When taking in the goods listed on the orders, the brokers learn about the fashion trends and can also give tips to retailers who put in orders about what is trendy and how many of an item should one order. This kind of information is crucial for the broker to expand his customer list, and also helps in matching retailers to wholesale merchants in the possible future.

When the night is almost over and all the goods are collected, they are loaded onto a truck and shipped to the retailers before the start of the work day. Retailers are now able to deliver the goods to their customers on time. However, work is not finished for the brokers. They need to organize and balance their order list, and after that their work “day” can be finished. The brokers are a bridge to many wholesale merchants and retailers — relations with each and every customer has a direct impact on the business of each broker. Having a good reputation is essential to acquiring new retail customers and also trust. If a broker has a good relationship with wholesalers, they may be able to get free “samples” from time to time. Having a possible demo item before an official product launch proves to be vital help in securing new lines of business.

4.2 New Roles for Brokers

In regards with the first research question of this investigation, we have found that in addition to the three traditional roles of intermediaries — traders, distributors, and providers — an additional two roles can be identified: one as a matchmaker, and the others as a gatekeeper. The three traditional tasks are carried out much of the same way as mentioned above in Section 2. While there may not be an exact “trader” role
for the DDM brokers, the distributor and provider are very much alive. Both parts cannot be explained separately, as they are closely-knit together and function as a two-way street. The main difference is that distributing tends to be a physical task, while providing tends to be informational. Brokers as a distributor can be observed when brokers are physically collecting the items ordered by the retailers and gathering them up at a loading zone. On the other hand, during the process of collecting and organizing orders, the brokers are able to learn about the most recent trends in the people’s taste in fashion. This process is diagramed in Figure 4. In turn, the provided information is reflected on the next batch of newly designed clothes made by the wholesale merchants, fulfilling the duty of the provider.

“I am constantly in touch with my retail clients, so it’s easy for me to know the current customer trends. If there’s a particular style or product that is ordered from multiple stores, I can see what is popular.” <P04>

“Talking to the popular stores, the big ones especially, is usually a good measure to see what is popular. The wholesalers can’t really get this information, so I forward these along to my customers.” <P05>

The same can be said when orders are being returned to the wholesalers. Returned products represent a change in people’s fashion, whether it is due to seasonal changes or just simple fads fading out. The information going back towards the wholesalers show brokers acting as providers. Wholesale merchants are able to decipher the wants and needs of the market based on both orders and returns. In addition, both wholesale merchants and retail shops This influences the DDM’s quick reaction system on fashion, and due to the quick deliveries made by the brokers, the market is able to adjust and adapt to the market in a short period of time.

“If there are products that are returned at a high rate, I call up my customers to see what the issue is. I also take a look at the products to see if I can find the problem. Then on my return trips I tell the wholesalers what the problem is.” <P04>

It is not only the wholesalers and retailers that benefit from this information. The brokers are the ones taking orders and returns, which makes them aware of the fashion business as well. First, by dealing with numerous clients and taking care of countless orders, the brokers are able to have a better understanding of trends of the market. When retailer puts in seemingly outdated orders, brokers are able to step in and give appropriate feedback to the client. The same goes with the quality of the clothing. Brokers obtain various feedback on the clothes’ fabric, quality and design from the returns by retailers. The retailers provide which piece of clothing is satisfactory and which is not, along with the specific reasons. By combining the details, the brokers themselves are able to function as a gatekeeper, introducing a more direct quality check in the DDM fashion district.

This process pushes the boundaries of intermediaries, and with the additional knowledge, brokers are given a new role as Gatekeepers within the DDM fashion district. Not only do they function as a gatekeeper, controlling the ins and outs within the market, they also act as middlemen, brokering and linking new retailers to the appropriate wholesale merchants. Brokers can coach new businesses how to approach the fashion district and also give plenty of advice. Gatekeeping and coaching roles help expand the DDM fashion district’s market size, and their traditional intermediary activities help in sustaining the market. Their contributions often go beyond their roles of brokers. Armed with the knowledge gained from working as a broker, they move on and take on a different job in the market, perhaps as a wholesaler or retailer. Their network in the fashion industry serves them well.

“At DDM there are so many shops, selling all kinds of different styled clothing, so it is difficult for newcomers to choose a wholesaler to buy from. We do the searching for them, because we know a lot more about the market than they do. We practically live here.” <P05>

“I’ve been working as a broker for seven years, and I’m still working as one, but I’ve started a side-job last year. My wife is working as a designer now, and we’re running a small wholesale designer shop. Because of my experience and I know many people around the market, it was relatively easy to start this.” <P02>
4.3 Brokers as Fashion and Trend Experts — Dealing with Subjective Information

Infomediaries have perhaps forcibly changed the roles of intermediaries, but another important factor is the conditions that give the need for the re-entrance of intermediaries. As mentioned earlier, the creation of infomediaries should have had a negative impact on existing intermediaries. To a more extreme level, infomediaries were supposed to bring an end to the human agent, but that is absolutely not the case. Human intermediaries have a solid footing in the fashion district, and even more so is critical to the maintenance of the DDM market. We have to ask why brokers, in the existence of infomediaries, must go around the market by foot, and act as the delivery man? What is it that infomediaries do not provide versus the brokers? The answer lies in the type of product and information the DDM brokers are dealing with in their work.

In a bigger scale, a look into the logistics of the DDM fashion district provides detail on what the brokers have to deal with. First of all, the sheer size and quickness of the fashion district can be pointed out. There are more than 30,000 wholesale stores inside the market, and each of these wholesalers are operating under the quick response system, pumping out multiple new items every few weeks. With this kind of information flooding to the DDM infomediary called the “New Arrivals Market”, it is almost impossible for the retailers to see which product is gaining interest. Under the QR system, manufacturing new clothes is rapid and instant changes in preference and popularity is crucial in controlling the production line. Merely relying on the infomediary to push out information on hot items is not enough for the merchants. The system may not be updating quickly, or there might be multiple orders and cancellations by customers, causing confusion within the system. And with all purchases being spread out over the diverse pool of clothes in the infomediary, distinguishing what is popular proves to be a difficult task. However, since the DDM brokers are the ones delivering and returning orders to and from the fashion district, their information on the pulse of the market is reliable and trustworthy to both wholesalers and retailers.

“A lot of new retailers who start their own business know people working in DDM from the start. It’s really difficult to start a business from scratch” <P06>
outside of the market. So they usually rely on us to get information.”  <P01>

“I have some stores that I’m really close with. I tend to tell them more about what I’ve learned, and they listen to my advice as well. I sometimes link new retailers to them if it’s a match.”  <P06>

Another factor is the item of trade itself in the market: clothes. People want to wear good looking, pretty, beautiful clothes. The problem is that clothing items in general do not have a set standard on defining what is “beautiful”. After all, beauty is a subjective measure — even two pairs of the same shoes in different color sparks plenty of disagreement on what is better looking. Simply put, infomediaries do not have the capability to determine what piece of item is better than the other. The only information that the “New Arrivals Market” can provide to its users is the items that are for sale, how much it is, and the popularity of each item; all numerical values. But with the subjective nature of the preference in clothing, this type of data is not useful. Style and fashion cannot be quantified nor standardized. Recommendations on style is a uniquely human ability, so it is only natural that the DDM brokers, who have the ability to make value judgements, are an absolute necessary in this market. Brokers with years of experience of watching and interacting with the market have the knowledge, or a sense, on what item would be a hit. They know what kind of designs, colors, and styles that are popular, and are in perfect position to relay that information to both wholesalers and retail shops. Brokers can also determine if a product is deemed too expensive, another information that infomediaries cannot provide. The subjective nature of the items in circulation require human agents to be present.

“Last summer, open-shoulder clothing were a hit. But, open-shoulder clothes that had some strap on them, compared to being totally open were more popular. When I pick up on information like this, I let the wholesalers know about it.”  <P04>

“A lot of the retail shops that opened recently are geared towards people in their mid-to-late thirties. So naturally more elegant clothing, such as long dresses became popular. It was different than a couple years ago when mini skirts and shorts were a hit.”  <P03>

5 Discussion

This research proposes two discussable points: 1) what are the new roles imposed on intermediaries by infomediaries? And 2) is subjectivity what makes human intermediaries a necessity, enabling the co-existence with infomediaries? The changing roles of intermediaries have been very well documented, with seemingly big technological advancements exerting the most influence on the human agents. However, the findings from this investigation contrasts the suspected reduced role of the human intermediary. In addition to maintaining the familiar roles intermediaries had in the past, DDM brokers have taken on a more active role inside the community. We suggest that the nature of information — subjective beauty and style of clothes in this environment — has contributed to the enhanced role the brokers play. Infomediary solutions such as the “New Arrival Markets” are simply a congregation of available items sorted by popularity, and is definitely not capable of providing the stylistic information merchants on both sides need. Wholesale merchants need information on current trends and design styles to manufacture clothes, and feedback from retailers need to come in fast and reliable. The existing roles that the brokers had in the market naturally placed them to be the perfect messengers for both physical items and intellectual knowledge and advice.

Another implication this research has is the applicability to a broad range of problems that pit humans against machines. Research and applications of computers, artificial intelligence and automated systems are on the rise, and present many discussion points within the job sectors. There are studies that show how telemarketers and sports referees are the first jobs that would be replaced by automated machines(Frey & Osborne, 2013). While the observations from this research is limited to the case in Korea and the DDM fashion district, the findings from this study can be applied to a wider range of environments and occupations. The shortcomings of computer-aided systems are clear. As displayed in this body of work, the lack of ability to make judgement calls on subjective matters is an area that must be improved. On the other hand, we believe this investigation has pointed out the direction on how humans can compete and also get along with technological advancements. In areas where humans workers are in direct competition with technology, the kind of insight displayed in this study can prove to be helpful.
6 Conclusion

This research investigated a group of brokers working inside the DDM fashion district to understand the changes infomediaries brought to traditional human intermediaries. Through contextual inquiry sessions and in-depth interviews, this research aimed to understand the brokers themselves and the relationships the brokers had in and out of the fashion district. By focusing on the brokers, the research has found the co-existence of old and new roles of intermediaries carried out by the DDM brokers. On top of traditional distributor and provider roles, DDM brokers have been observed to take on the gatekeeping and matchmaking roles inside the market, simultaneously growing and maintaining the fashion district. The investigation also led to findings related to the conditions that contribute to the re-introduction of intermediaries. In environments where subjectiveness is the main information characteristic, the need for human intervention is high. Unless infomediaries are able to decide and make value judgements on subjective matters, human intermediaries are likely to continue to be needed. Infomediaries can only deliver “shallow” data, and it is the human agents that can delve deeper to find insights from the information. The second coming of intermediaries, or “re-intermediary”, is surely to exist elsewhere, and it would be interesting to discover more environments where this phenomenon is observable.

References


Towards a Graph-based Data Model for Semantics Evolution

Xuhui Li\textsuperscript{1,}\textsuperscript{*}, Dewang Sun\textsuperscript{1}, Mengchi Liu\textsuperscript{2}, Tieyun Qian\textsuperscript{2}, Feicheng Ma\textsuperscript{1,}\textsuperscript{*}
\textsuperscript{1}School of Information Management, Wuhan University, Wuhan, China
\textsuperscript{2}State Key Lab of Software Engineering, Wuhan University, Wuhan, China
\textsuperscript{*}Corresponding Author

Abstract
Semantic information comes from the things being recognized and understood gradually, and thus it is often in evolution during the modeling process. Existing semantic models usually describe the objects and the relationships in an application-oriented way, which is unsuitable to reuse the schemas during the semantics evolution. In this paper, we propose a new graph-based semantic data model to overcome the limitation. SemGraph adopts a meaning-oriented approach to specifying the subjective view of the things and uses the certain meta-meaning relationships to build a graph-based semantic model. The model is simple but expressive, and is especially fit for the semantics evolution. We introduce the basic concepts and the essential mechanisms of the model, demonstrate its features with examples and typical cases of semantics evolution.

Keywords: Semantics Evolution; Semantic Data Model; Meaning-oriented Approach; Graph Data Modeling


Copyright: Copyright is held by the authors.
Acknowledgements: This research is supported by the NSF of China under contract No.61272110, No.91646206, No.71420107026, No.61572376 and No.61272275 , and the China Postdoctoral Science Foundation under contract No. 2014M562070.
Contact: lixuhui@whu.edu.cn, fchma@whu.edu.cn.

1 Introduction

Semantic data model is always a central topic in knowledge and data engineering. Since the late 1970s, researchers have designed various models to specify semantic information in a way that computers can manage and manipulate. The semantic information of the things in the real world is often treated as the instantiation of various semantic concepts. These concepts are often interrelated by the certain meta-semantic relationships such as attribute function, aggregation, generalization/specialization, role, derivation, etc.(Hull & King, 1987; Peckham & Maryanski, 1988). Therefore, semantic information is recognized and understood with the concepts in which it is involved. How to present, organize and find the semantic information according to the concepts and their interrelationships is thus a fundamental issue for the data models to concern.

Early studies on semantic data modeling often explicitly or implicitly focus on describing the “entities” (or “objects” in some literatures) and the “relationships” among the entities, known as the Entity-Relationship models, and treat them as the “semantics” of the real world. The semantics embodied in the models is often presented as “attributes”, i.e., the meaningful functions between the source and the target entities, and the “aggregation” which are essentially n-ary relationships of entities or subordinated “aggregation”. The semantic models have been proposed following two philosophical approaches: one approach places an emphasis on explicitly type constructors especially the aggregates (Abiteboul & Hull, 1987), and the other stresses the use of the attributes(Hammer & McLeod, 1981; Shipman, 1981). The models such as enhanced ER modeling approaches (Thalheim, 2011) also adopt various ways of specifying the generalization/specialization of the concepts and the derived concepts.

The later studies are overwhelmingly affected by the rapid development of object-oriented paradigms, which mainly uses the attributes of the objects to describe the semantic relations to other objects(Atkinson et al., 1989; Cattell & Barry, 2000). Generally, these models often adopt taxonomical approach to classify the semantic information into various categories corresponding to various concepts, and use various semantic features such as the object identity, the aggregation, the generalization/specialization, the class hierarchies, the non-monotonic inheritance, etc., to describe and refine the concepts. Some early object-oriented models
mainly concerned about the static aspects of the real world and normally required an object to be an instance of a most specific class. Then the literatures (Dahchour, Pirotte, & Zimányi, 2004; Liu & Hu, 2009), are more interested in describing the dynamic semantic features of the entities, e.g., transient roles under certain context, which is common in practical modeling.

Another kind of data models, the graph data models, attempt to overcome the limitations of the conventional models with respect to capturing the inherent graph structure of data appearing in application, such as hypertext or geographic information systems, where the interconnectivity of data is an important aspect (Angles & Gutiérrez, 2008). As the entities and their relationships are also the fundamental issues for these models, they are deeply influenced by the semantic data models, especially the object-oriented data models, and can present diverse semantic features using graphs or hyper-graphs (Hidders, 2003; Choudhury, Chaki, & Bhattacharyya, 2006). Many of them can represent the attributes and aggregations, and some of them also adopt the inheritance and derivation. In fact, some semantic data models like IFO (Abiteboul & Hull, 1987) are designed as a graph-based one which can present most semantic features.

The recent decade has seen the popularity of NoSQL databases (datasets), e.g., the document databases and the graph databases. These databases (datasets) are to contain and manipulate the information involving more semantics than conventional relational databases, even though there is an insufficiency of the semantic modeling study for them. For the document databases managing the XML documents or JSON-like documents, the tags or the keys of the data elements are often regarded as semantic labels of the data, and the hierarchically organized data schema can also present the compound structure of the semantic concepts (Mani, Lee, & Muntz, 2001; Ahmed, Lenz, Liu, Robinson, & Ghafoor, 2008). On the other hand, the popular graph databases (Buerli & Obispo, 2012) use big graph datasets describing the entities and their relationships as the graph nodes and edges. However, these graph databases seldom deploy the aforementioned graph data models to present the semantic information (Angles, 2012). An exception is the RDF datasets deploying the triple-based attribute-oriented data models like RDFS (Brickley & Guha, 2014) to build the knowledge bases like the Semantic Web (Heath & Bizer, 2011).

The above mentioned data models are quite expressive and comprehensively cover the various aspects of semantic modeling. However, these models are often designed in an application-oriented way. That is, the semantic concepts and their interrelationships, especially the taxonomical relationships, are well-designed in a top-down process according to the application requests. The models usually adopt a pragmatic way to describe the semantic information, but these descriptions are not easy to be reused for other applications. For example, in describing a person concept, the designers usually use the attributes like “name”, “birth-date” and “occupation”. However, such information is only used as the social labels of a person under certain context, e.g., social security, and thus it is not actually natural for a physical person who is just a human being. In practical data modeling, the semantic information about the entities and the relationships are often in evolution. During the evolution, new semantic information is generated from the old one through various data operations, but the semantic manipulations in existing models cannot sufficiently support these evolution requests. This insufficiency severely hinders the reuse of the semantic information, as would be shown in the next section.

In this paper, we propose a new graph-based semantic data model named SemGraph (standing for Semantics Graph) to overcome the inadequacy. SemGraph is a data model which uses meaning-oriented schemas to present interrelated concepts in a graph structure. The contribution of the paper lies in two folds. On one hand, we propose a meaning-oriented approach to modeling semantics. Existing approaches almost focus on the objects and the relationships, whereas our approach treats the subjective meaning of the things as the major source of semantic information. Therefore, our approach can simplify the data structure by adopting some simple and essential meta-semantic relations among the meanings. On the other hand, we design a set of mechanisms to work with the meta-semantic relationships. They can efficiently present the semantic information in the evolution and can effectively reuse the existing semantic schemas and data.

The remainder of this paper is organized as follows. Section 2 uses a scenario to show the motivation of exploring a new approach to semantic modeling for semantics evolution. Section 3 introduces the SemGraph data model, including its basic concepts and the four meta-semantic relationships which are useful for presenting the structure and the features of the semantic information. Section 4 further discusses the issues of representing semantics using SemGraph with the major mechanisms in common semantics evolution. Section 5 concludes the paper.
2 Motivation

2.1 A scenario of semantic data modeling

Let us consider the evolution of the semantic information in a teaching scenario where the semantic information of the entities and the relationships is gradually enriched as shown in Figure 1 (a)-(j).

Figure 1: A scenario of modeling semantics evolution

In this scenario, the teachers and the students are initially treated as the people related to each other by the “teaches” and the “taughtby” relationships (Figure1 (a)). Then the teaching content is involved, and the figures (b) and (c) depict the attribute-based and the aggregation-based approaches to modeling the case. Since a teacher often teaches several students the same content, the 1:1 teaching relationship between the teachers and the students needs to be extended to 1:n as shown in (d). However, the current teaching relationship contains redundant couples of the teacher and the content which can be separated as a new nested relationship “teaching content”, as shown in (e).

In a broader scope, a school, as shown in (f) is modeled to organize the teachers and the students and to separate the students into classes. Here a teacher plays the role of the president, and the courses are introduced as the official teaching content and a score is attached to a pair of student and course to assess his/her study.

At this stage, we find that the school is an organization consisting of the members, and an organization is often specialized to the one with some geographical locations for their members to work and to take certain activities like meetings or teachings, as shown in (g) and (h). Correspondingly, we need to consider the school again with the locations including the offices for the teachers, the dormitories for the students and the teaching rooms for the teaching events; after that, the school model would be associated with a geo-information model.
to do some geo-sensitive works, as in (i) and (j). A practical use of such an information system might be a request on the geo-information of a school, e.g., to seek a free teaching room for an optional course in the mid of a term. This task can be fulfilled by a certain geo-information sorting algorithm to choose the minimum expense of time for the students and the teacher to arrive at the destination teaching room from their dormitory rooms or office.

2.2 Limitations of conventional semantic data models

The above scenario might not be difficult to model for some semantic models if the complete knowledge and the application requests, e.g., the geo-information of a school and the teaching room selection, are known in advance. However, it is often not the case in practical application modeling. We usually only have partial information about the world in the beginning and cannot see the whole scene until the model is almost done (or even worse). A common request for the modeling approach is that they can provide a proper evolution mechanism which can keep up with the evolution of our recognition to the real world. That is, the model of the partial world can be reused as much as possible in the future. However, for the existing semantic models, there are some issues not properly addressed, as listed below, which hinder the models to be evolved and be reused in practice.

a. Extending attributes. The attributes denoting the binary relationships between the entities, such as “teaches” or “taughtby” in Figure 1-(a), often need to be extended to contain more information like the “teaching content” in Figure 1-(b)(c). This extension can be carried out by extending the attribute as the multi-valued one (as in (b)), which is adopted by some models like FDM(Shipman, 1981) or RDFS(Brickley & Guha, 2014). However, the inverse relationship like “taughtby” and the new relation introduced by the extension, e.g., the one between the student and the teaching content, are not easy to represent, and the possible further extension would make the model more complicated. Another solution is to use the aggregation approach to redesign the extended concepts (like in (c)). Both solutions severely restrict reusing the previous modeling result.

b. Projecting aggregations. The aggregations gather the related entities as n-ary relations. These relations, e.g., the teaching relation in Figure 1-(d), are formed in a pragmatistic way and often need to be adapted for new situations such as in Figure 1-(e). However, existing models usually provide the mechanisms, such as the subtyping or the nested aggregation, to extend or combine the aggregations. But they seldom provide a mechanism to project an aggregation, just like the projection operation in the relational model, into smaller parts which represent subordinate and original semantic units.

c. Representing context-dependent concepts. Many modeling approaches recognize the importance of the context-dependent concepts, i.e., the concepts which only has actual meaning in certain situations. Some early models like IFO use the specialization mechanism to specify the context-dependent concepts, meanwhile the common way in the recent studies is the role-based approach using the context-dependent roles associated with other entities. These two approaches focus on representing different aspects of the context-dependent concepts. Some concepts like a person being a teacher in a school are static and should be treated as context-dependent subtypes, and the others like a person being a president of a school are dynamic and should be treated as a role. However, existing models seem seldom use a consistent mechanism to integrate the two approaches.

d. Specifying logical dependency. The semantic concepts are usually dependent to one another logically, and this dependency is important in representing knowledge and searching latent semantic information. For example, from the 1:n teaching relationship in Figure 1-(c), the simpler 1:1 teaching relationship in (b) should be easily inferred. Such logic dependency should be checked by the model to guarantee the completeness of the semantic information. However, existing models seldom deploy proper mechanism to specify and check it. Some models allow users to manually specify the logical dependency as user-defined constraints, but these constraints are optional and the specification is legal either with or without the constraints. Therefore, a database with a sound model might contain incomplete data logically.

e. Specifying the generalization. As shown in Figure 1-(f) and (g), the specialization is often required to be dynamically specified during the modeling process, so as to enable the generalization of existing semantic
types, e.g., generalizing the “organization” concept from “school”. However, the existing models often only allow the specialization of the concepts be statically declared in their definition, and thus the generalization of the concepts have to be restrained to the union of the concepts.

f. Presenting some important constructed types. Existing semantic models often support certain kinds of type construction, like the “union” and the “product” (through aggregation). The “intersection” and the “natural join” of the semantic types, which are common and important in building the links of semantics objects as shown in Figure 1-(i) and (j), are often presented through the multi-inheritance mechanism or the derived types of the product using the constraints to simulate the general join operation. However, the current multi-inheritance mechanisms focus on combining existing concepts, e.g., the “member_loc” and the “teacher” in Figure 1-(i), but seldom consider the cases of the intersection involved in other semantic modeling issues such as generalization and attributes. On the other hand, the instances of the derived types are apt to be affected by the update propagation of the joined values, but we are often just interested in the statically related objects sharing common attribute targets as in Figure 1-(j).

Due to the above inadequacy of the semantic models in presenting the evolving semantic information, we propose a new graph-based model named SemGraph to meet the requests of frequent semantics evolution.

3 The SemGraph Model

3.1 Basic concepts of SemGraph

SemGraph adopts a meaning-oriented approach to semantic data modeling. In SemGraph we don’t directly describe the objective entities and relationships, instead we use a “meaning” to denote the subjective view in our mind when we think about a thing, e.g., an entity, a relationship, an event, etc. For example, to describe a school entity, we don’t consider what the school is, but dwell on the information about the school from a certain viewpoint, e.g., teaching or geo-information. A “meaning” is a (partial) understanding of a thing, and the related “meanings” from various aspects can form an overall and considerate understanding of the thing.

As the fundamental semantic units, the “meaning” are interrelated to model complex semantic information in a directed graph structure called semantics graph, as the name “SemGraph” indicates. In a semantics graph, the nodes represent the “meanings”, and the edges represent the meta-semantic relationships between the meanings. A meaning node has a label, e.g., “a:teaching”, denoting its meaning type and thus the node a is treated as an instance of the type “teaching”. A meaning type is specified in a schema indicating how its instance nodes should be linked with other nodes. For brevity, we often call a meaning type a “meaning” and a meaning node a “node”.

SemGraph allows four kinds of meta-semantic relationships: the ”composition”, the ”specialization”, the ”reference” and the ”equivalence” relationships respectively denoted by the symbols “→”, “⇒”, “⇔” and ”⇔” between pairs of nodes. Fig.2 depicts a semantics graph describing some simplified information about the school, the teaching, the teachers and the students.

![Figure 2: A semantics graph of simplified school information](image)

In a “composition” relationship, e.g., “b:#Teaching→c:Teacher”, the source node a represents a composite one and the target node b represents a component meaning (or component for short) node of a. The
composite node can have one or more component nodes, e.g., there might be another relation “b:#Teaching→ g:student”. The component m’ of the meaning m is also denoted as m.m’. In the figure and the following description, the meaning “m.m’” is often abbreviated as “m’”, e.g., “e:Teacher” is the abbreviation of “e:#School.Teacher”.

In a “specialization” relationship, the target meaning specializes the source meaning. For example, in the relation “s:#Person⇒t:Student”, the meaning “Teaching.Student” called the sub meaning is a specialized kind of the meaning “#Person” called the super meaning. The source node and the target node are thus respectively called the super node and the sub node.

In a “reference” relationship, the source meaning represents the reference to the target meaning and the target meaning node can be dynamically altered. For example, the reference relationship “w:Age→ 15:#Nat” indicates that the age w should be a reference to the natural number 15; and the relationship “r:President→ e:Teacher” indicates that the president role r be played by the teacher e.

The “equivalence” relationship indicates that the two meanings be semantically “equivalent”. For example, “e:Teachers⇒x:Teacher” indicates that the two nodes e and x represent the same “teacher” meaning of the same person. Such equivalence is common in object oriented programming, e.g., Java, where an equals() method is always provided for a class to specify how the two instances of the same class are equivalent to each other.

The following subsections would dwell on the basic features of SemGraph using the above meta-semantic relationships. In this paper the features are mainly introduced with the examples, please refer to (Li, 2016) for a more formal description of SemGraph.

3.2 Composition of meanings

SemGraph uses a simple but expressive composition mechanism to specify how the component meanings constitute a composite meaning. A meaning not being the component of another meaning is called a “root” meaning and is denoted with a “#” prefix. A component meaning can be composed of lower-level component meanings and thus forming a composition hierarchy. A semantic graph containing a composite node and all its descendant nodes in the composition hierarchy (maybe empty) is called a unit graph. It is required that the composition hierarchy be a directed acyclic graph, and so do the unit graphs. For example, the root meaning “#School” consists of the component meanings “Teaching”, “Teacher”, “Student” and “Course” meanwhile “Teaching” consists of the components “Teacher”, “Course” and “Student”.

The composition of the meanings indicates a semantical dependency, that is, the component meanings make sense only under the context of the composite meaning, and thus the composite node is also called the “context” of its component nodes. This feature makes SemGraph different from aggregation-oriented models where the components are loosely associated and do not depend on the aggregation. The semantic dependency is required to be determined by the users from the beginning stage of modeling and be statical during the semantics evolution, which enhances the reusability of the schemas.

For example, for the concepts “Teacher”, “Student” and “Teaching” involved in specifying the activity “a teacher teaches a student”, the meanings “Teacher” and “Student” depend on “Teaching” because they are two agents of the activity, and thus they should be the components of “#Teaching”. However, for a school teacher, the “teacher” concept makes sense only in the school and thus it can be defined as the component meaning “Teacher” of the “#School” meaning.

The schema of a composite meaning is specified as the combination of the components with certain modifiers and occurrence operators. For example, the composition information of the “#Teaching” and the “#School” meanings are specified as follows:

meaning #Teaching -> {public Teacher, public Student}
meaning #School -> {Teacher[*], Student[*], Course[*],
Coursegrade -> {A | B | C | D},
Teaching -> { Teacher, Student[*]->{Score | Grade}, Course, Location?} }
SemGraph adopts a regular-expression-like mechanism to specify how the component nodes can occur in its context. It uses the conjunctive modifier ",", the disjunctive modifier "|", the optional modifier "?" and the multiple modifier "[*]" for the component meanings to indicate their instance meanings’ occurrence. For example, in the composite meaning "#School.Teaching", the occurrence declaration "(Teacher, Course)" indicates that the teacher and the course should occur conjunctively once in the context, and "(Score | Grade)" indicates that either the "Score" or the "Grade" meaning occurs once, "Location?" indicates that a "Location" meaning would occur or not, and "Student[*]" indicates that one or more student meanings occur.

### 3.3 Specialization and reference

SemGraph deploys the specialization relationship to build the taxonomy of meanings and to enrich the sub meanings with more features. A sub meaning node would inherit all its super node’s component nodes (if exists), meanwhile it can develop new component nodes which usually specialize other meaning nodes (including the inherited component nodes) allowed to access. The sub node is often relatively independent to its super node unless the super meaning being specified with a "!" symbol (to be introduced later), because a sub node just represents a subjective extension to the super node and thus the existence of the sub node should not affect the super node itself. A meaning schema often embeds the specialization relationship declaration into the composition and contains certain constraints to specify its features.

For example, the following schemata show the specialization version of "#Teaching" and "#School".

```plaintext
meaning #Entity -> {Name <= #String}
meaning #Person <- #Entity -> {Age <= #Nat}
meaning #Teaching -> {public Teacher <= #Person, public Student <= #Person, public Content}
meaning #School -> {Teacher[*] <= #Person with {this.Age >= 18}, Student[*] <= #Person, ...
Teaching -> {
  Teacher <= #Teaching.Teacher & .Teacher,
  Student[*] <= #Teaching.Student & .Student ->
  {Score <= #Int | Grade <= .Coursegrade },
  Course <= #Teaching.Content & .Course, Location? }
with { virtual #Teaching (-> $x, ->$y, ->$z) :-
    this (-> Teacher <= $x:#Teaching.Teacher,
    -> Student <= $y:#Teaching.Student,
    -> Course <= #Teaching.Content } } }
```

As the example shows, the “Teacher” and the “Student” components of “#Teaching” are declared as the sub meanings of “#Person”, and so do these components of “#School”. In the “#School.Teaching” meaning, “Teacher”, “Student” and “Course” are declared as the intersection sub meanings, i.e., the meaning indicating the intersection of multiple super meanings, of the corresponding components of “#Teaching” and “#School”. For example, the specialization declaration “Teacher <= (#Teaching.Teacher & .Teacher)” indicates that a “#School.Teaching.Teacher” component node has a super node of “#Teaching.Teacher” and a super node of “#School.Teacher” meanwhile the two super nodes should have a common “#Person” super node. The intersection meaning would be introduced in more detail in the next section.

The constraints of a meaning indicates how to validate it and its related meanings. The constraints are appended with a meaning using a “with” clause in the schema declaration. There are two kinds of constraints: one kind is the validating constraints which impose restrictions on the properties of its (inherited) components. For example, a “#School.Teacher” has a constraint rule “this.Age >= 18”, which means that its “Age” component value be not less than 18. The validating constraints enable specifying the derived meanings in SemGraph. A derived meaning is the sub meaning of an existing one. For example, a derived meaning “big_school” can be defined as the school has over 1,000 students. Due to the independency of the sub node to the super node, the derived meanings can be dynamically appended to the super node and be deleted if the derivation condition is no longer satisfied.

The other kind of constraints are to guarantee the completeness of the related meanings nodes. The so-called complementary constraints are in the form of “p1 :- p2” where p1 and p2 are two patterns indicating the specialization relationships among the nodes of the specific meanings. For example, in a “#School.Teaching” node, the component nodes “Teacher”, “Student” and “Course” are assumed to be sub nodes of the certain “virtual #Teaching” nodes. These virtual nodes are not materialized in the graph, but
they can be queried through the rewriting mechanisms derived from the constraint. Furthermore, the virtual 
“#Teaching” nodes involved in these component nodes are required to be complete, i.e., each virtual node 
have necessary materialized components in the graph to form a complete unit graph. For example, each 
component node belonging to a certain “#Teaching” node are covered in the constraint, and each virtual 
“#Teaching” node has its components fully determined by the constraint. The completeness of the virtual 
nodes and the associated unit graphs should be determined by the schema constraints. SemGraph deploys a 
certain checking mechanism to guarantee that the schema has been equipped with the proper complementary 
constraints to avoid the incomplete information(Li, 2016).

The reference relationship is a variant of specialization which allows the super node being altered. This is analogous to the reference types in the programming languages. Here the sub meaning and the super 
meanings are respectively called the “delegate” and the “host” meanings. A delegate meaning can inherit 
the components of the host, and these components can be the hosts of other components in the delegate. However, since the host is alterable, the delegate meaning should not put restriction on the host and its 
components in its constraints, but these components cannot be specialized as the new ones in the delegate 
meaning.

An important use of the reference relationship is to implement the role models. It has the following 
features which are exactly required by a role model: a delegate node can refer to mutable host nodes; the 
delegate meaning can be a composite one and be extended, and the extended component meanings only 
make sense under the context of the delegate meaning. For example, in a school the president is obviously a 
role and thus can be defined as the component meaning “President→{Office, Telephone}” of the “#School” 
meaning with the reference relationship “President→Teacher”. Here the component meanings “Office” and 
“Telephone” only pertain to the president role rather than the teacher who plays the role.

The reference relationship can also be used to emulate the conventional O-O mechanism, as the 
mutable fields of an object can be treated as delegates and its values correspond to hosts. For example, the 
meaning “#Person.Age” can be declared as a delegate of the “#Nat”, which means the age of a person refers 
to a mutable natural number value.

3.4 Equivalence

The equivalence relationship is actually an isomorphism between two unit graphs. It is useful both at the 
meaning schema level and at the semantic graph level. The equivalence between the schemas is used to 
declare two meaning types originated from different situations are equivalent to each other. Therefore, the 
equivalence declaration should specify the correspondence between the components of the two meanings.

The equivalence between the nodes in a semantic graph is often more interesting and important in 
semantics evolution. A meaning node can have partial information, i.e., lacking some components, during the 
semantics evolution. For example, when we try to resolve the case like “a school teacher teaches the course 
101” in a semantic graph, we can build a “#School.Teacher” and a “#Person” meaning is created accordingly, 
but his/her “Name” and “Age” is not determined, and there might be another “#Person” node which is later 
determined to be equivalent to the former “#Person” node. To determine the equivalence between the nodes 
is an important semantic computation where recognizing and aligning the equivalent things are common in 
building semantic graphs in practice.

The equivalence relationships in semantic graph are not always physically represented as the equivalence edges. Sometimes they can be inferred based on the schemas and the nodes’ components, which prevents the 
graph from being trivially complicated with all the equivalence or specialization edges. The meanings 
whose equivalence can be (recursively) inferred are called intensional meanings, which are universally used 
to denote the value types, e.g., integers and strings. For example, the “#Person.Age” meaning can involve 
numerous specialization edges from all the “#Person.Age” nodes to their “#Nat” nodes. Instead, the 
specialization edges nodes can be replaced with a certain value node of “#Nat”, and the equivalences among 
the value nodes are determined through certain computations. The physical equivalence edges are thus 
transformed to the logical equivalence relations in data management.
4 Representing semantics using SemGraph

Based on the four meta-semantic relationships listed above and certain associated mechanisms, SemGraph can easily specify various kinds of semantic information and its evolution.

4.1 Dependency of meanings

SemGraph concerns the dependencies of meanings from both the context aspect and the logic aspect. Representing context-dependent concepts is straightforward in SemGraph. As mentioned previously, composite meanings are supposed to be the context of the components. Therefore, the components always represent the context-dependent concepts. These concepts can also be treated as the roles under the context. For example, for the case of a component “Teacher” of a composite meaning “#School” specializes a meaning “#Person”, it indicates that a person play the role of teacher in a school. When the component’s super meaning is alterable, i.e., the reference relationship is considered, the role becomes dynamic, as the case of the “President” component described previously.

Logical dependency is mainly specified by the specialization relationship and the related constraints, as shown in Section 3.3. A specialization relationship indicating the taxonomy of meanings represents a basic logical dependency between the meanings. Furthermore, the validating constraints and the complementary constraints provide practical mechanisms to specify the subtle logical relationships and the restrictions between the meanings. For example, the logical dependency between the 1:n teaching relationship in Figure 1-(c) and the 1:1 one in Figure 1-(b) is easily and completely addressed with the specialization and the constraints described in Section 3.3.

4.2 Type Construction

Constructing the compound semantic types to present complex meanings is the most important task for a semantic data model. Existing data models usually provide some common constructors like the product (or aggregation) and the group, meanwhile some models also try to use the preliminary generalization mechanism to implement the union type.

SemGraph can easily implement various type constructors, since the composition relation and the component occurrence mechanisms has already laid the foundation of the product(or aggregation), the group and the union types. SemGraph especially provides a product type constructor “ˆ”, a group type constructor “*” and a union type constructor “||”, to explicitly declare a compound meaning type of other root meanings. (A compound meaning is a special composite meaning generated by the constructors.) For example, the compound meaning “#Schoolˆ#GIS” consists of two components “(#Schoolˆ#GIS).School” and “(#Schoolˆ#GIS).GIS” which are respectively the sub meanings of the two root meanings “#School” and “#GIS”, and the two components should occur conjunctively. For the compound meaning “*#School”, the component meaning “(*#School).School” should be declared to have multiple instances. For the compound meaning “#School||#Institute”, it is the super meaning of “#School” and “#Institute” whose components should occur disjunctively in the union.

SemGraph does not support the compound meaning of the components of different root meanings, because the existence the components always assume the existence of their root meanings and thus the compound of their root meanings should be declared previously. However, the components under the same context can be composed with the type constructors.

Although the compound meanings can easily implement the functions of the conventional type constructors, they are different in their intension. The latter like the aggregation often represents the loose associations of the member entities or relations, and the member types don’t depend on the compound semantic types. However, the former is declared as the context under which the components can make sense. Therefore we restrict the components of the product or group meaning to be the sub meanings of the original root meanings.

Intersection of semantic concepts is usually considered by the logic-based semantic models, e.g., the ontology models, and is seldom directly supported by the data models. Some data models can indirectly implement the limited features of the intersection through multiple inheritance. SemGraph introduces a special intersection compound meaning using an intersection constructor “&”. For an intersection meaning
\( l_a \in l_b \), the meanings \( l_a \) and \( l_b \) are both the sub meanings of a certain meaning \( l \), and \( l_a \in l_b \) is automatically declared as the sub meaning of \( l_a \) and \( l_b \) respectively. The \( l_a \in l_b \) meaning inherits the components of its super meanings (through a certain mechanism avoiding the name conflict). Recursively, under the context \( l_a \in l_b \) the certain intersection component meanings \( l_1 \in l_2 \) can be declared where the meaning \( l_1 \) and \( l_2 \) are respectively the components of \( l_a \) and \( l_b \) and they are both the sub meanings of a component \( l' \) of \( l \). At the instance level, a node of \( l_1 \in l_2 \) certainly shares the same super-node of \( l' \) as \( l_1 \) and \( l_2 \).

For example, for the meanings “\#School” and “\#Institute” which are both the sub meanings of “\#Organization”, the component meaning “\#Organization.Member” is specialized as the meanings “\#School.Teacher” and “\#Institute.Researcher”, and in the intersection “\#School\&\#Institute” a new component “Teacher\&Researcher” can be declared. In the semantic graph, if a node of “\#Organization” has a sub node of “\#School” and a sub node of “\#Institute”, they would have a sub node of “\#School\&\#Institute” which not only inherits the components of the components “Teacher” and “Researcher” but also has the new nodes of “Teacher\&Researcher” which specialize the “Teacher” and the “Researcher” nodes being of the common “Member” super nodes.

To avoid the possible combinatorial explosion of the compound meaning types produced by the type constructors, SemGraph deploys a declare-to-use policy. That is, the compound meanings with the type constructors only make sense if they are explicitly declared in the schema. The nodes of a compound meaning are virtual on default, and the nodes of its user-defined sub meaning are physical in the semantic graph. Furthermore, since the intersection meaning only works as the two super meanings \( l_a \) and \( l_b \) have the common ancestor \( l \), the amount of the intersection node and its component nodes can be restricted in the polynomial space.

With the aid of intersection, SemGraph can easily present the natural join operation. For the two meanings \( m \) and \( m' \) where the components \( m.x \) and \( m'.y \) share a common ancestor super meaning \( z \), the natural join of \( m \) and \( m' \) through \( m.x \) joining \( m'.y \) is presented as the sub meaning of the product meaning \( m \cdot m' \) where there exists a component \( m.x \in m'.y \), and thus is denoted as \( m \cdot z \in m.x \& m'.y \). Since \( m.x \in m'.y \) shares the super meaning \( z \) with \( m.x \) and \( m'.y \), the join meaning \( m \cdot m' \) can exactly combine \( m \) and \( m' \) using \( m.x \in m'.y \) as the joint. For example, the case in Figure 1-(j) can be specified as the natural join of the meanings “\#School_loc” and “\#GIS” using a certain joint meaning “\#School_loc.location.geo_id \& \#GIS.loc_geo.geo_id”, assuming that the “\#GIS” meaning be defined following the above guidelines.

### 4.3 Generalization

The specialization relationships between meaning nodes are often established as the sub nodes are instantiated. However, SemGraph also supports creating sub nodes from super nodes by declaring the specialization with the “\{" symbol. The relationship “\( A \Leftarrow !B \)” means that for every node of the super meaning \( B \), there does exist a node of the sub meaning \( A \). This mechanism is used to declare generalization of existing meanings.

The following schema shows an example of declaring a meaning as the generalization of an existing meaning.

```plaintext
meaning #Organization -> { Member[*], Head --> .Member }
meaning #School_O <- !#School & #Organization -> { 
  !.Teacher & .Member, !.Student & .Member, !.President & .Head }
```

In this example, an “\#Organization” meaning is defined after the “\#School” meaning. Then a new meaning “\#School_O” is declared as the intersection meaning of “\#School” and “\#Organization”, and especially, the components are specified as the intersection meanings of the corresponding component in the super meanings. Since the “\#School” and its components are decorated with the “\{" symbol in “\#School_O”, every node of “\#School” is required to correspond to a node of “\#School_O” and so do their components. As “\#School_O” is also a sub meaning of “\#Organization”, the generalization of “\#School” from the viewpoint of “\#Organization” is actually formed because every component of “\#School” corresponds to a sub component of the one in “\#Organization”.

### 4.4 Representing and extending attributes

Many conceptual modeling approaches use attributes to denote binary relationships between entities. In SemGraph, the attributes can be emulated easily using the generalization based on a binary relationship
meaning “#BiRel”. The following example specifies how to represent the “#Teaching” meaning as a binary relationship “#Teaches”.

\[
\text{meaning #BiRel} \rightarrow \{ \text{First} \leq \text{#Entity}, \text{Second} \leq \text{#Entity} \}
\]

\[
\text{meaning #Teaches} \leftarrow \#\text{Teaching} \& \#\text{BiRel} \rightarrow \{ \text{!.Teacher} \& \text{.First}, \text{!.Student} \& \text{.Second} \}
\]

As the example shows, the “#Teaching” meaning corresponds to a special case, i.e., “#Teaches”, of the generalized meaning “#BiRel”. Therefore, “#Teaches” can be used as an attribute of “#Teaching.Teacher”, and its inverse relationship “#TaughtBy” can be defined similarly.

Using the common specialization and the above generalization mechanisms, attributes can be easily extended. On one hand, to extend an existing attribute, e.g., “#Teaches”, we can just extend the meaning with new component like “Content”. On the other hand, to create a new attribute from the existing meaning, e.g., “student studies content”, the attribute like “#Studies” can be specified with another “#Teaching & #BiRel” where the intersection component meanings “!.Student & .First” and “!.Content & .Second” exist.

4.5 Projecting composite meaning

The components in a composite meaning sometimes have subtle interrelationships. The tightly-coupled components can form a new composite meaning which is a projection of the original composite meaning, as described previously. For example, in Figure 1-(e) the “teacher-content” couple is recognized as a new meaning and needs to be specified separately.

The projection of a composite meaning is also a variant of generalization. Since the components in the projection meaning have the same information as the ones in the original meaning, they need to share the content with the original components through the intersection meanings.

The following schemas show the projection example of Figure 1-(e).

\[
\text{meaning #Teaching} \rightarrow \{
\quad \text{public Teacher} \leq \#\text{Person} \rightarrow \{ \text{ID} \leq \#\text{String} \},
\quad \text{public Student} \leq \#\text{Person} \rightarrow \{ \text{ID} \leq \#\text{String} \},
\quad \text{public Content} \rightarrow \{ \text{Name} \leq \#\text{String} \}
\}
\]

\[
\text{meaning #Teach_content} \rightarrow \{
\quad \text{public Teacher_p} \leq \#\text{Person} \rightarrow \{ \text{ID_p} \leq \#\text{String} \},
\quad \text{public Content_p} \rightarrow \{ \text{Name_p} \leq \#\text{String} \}
\}
\]

\[
\text{meaning #Teaching_proj} \leftarrow \!\#\text{Teaching} \& \#\text{Teaching_content} \rightarrow \{
\quad \text{!.Teacher} \& \text{.Teacher_p} \rightarrow \{ \text{!.ID} \& \text{.ID_p} \}
\quad \text{!.Content} \& \text{.Content} \rightarrow \{ \text{!.Name} \& \text{.Name_p} \}
\}
\]

In this example, the intersection meaning of the components “Teacher” and “Teacher_p” exists and further their components “ID” and “ID_p” also have intersection meaning. Therefore, in a “.Teacher & Teacher_p” node, the “ID” and the “ID_p” components specialize the same #String node, i.e., they have the same value. Similarly for the “Content” and “Content_p” components. As a “#Teaching_proj” node is mandatorily instantiated by its “#Teaching” super node, its “#Teach_content” super node should have the information according to the “#Teaching” node, which actually forms a projection of original node.

5 Conclusion

Semantic information comes from the things being recognized and understood gradually, and thus it is always in evolution during the modeling. In this paper, we have proposed a new graph-based semantic data model named SemGraph to overcome the limitations of existing semantic data models in semantics evolution. SemGraph deploys a meaning-oriented approach to specify the subjective view of the things and uses the certain meta-meaning relations to build a graph-based semantic model, which is simple but expressive, and especially fit for the semantics evolution.

Our work on the SemGraph model is in the preliminary stage. Based on the meta-semantic relationships and the mechanisms listed in the paper, we are studying the proper approaches to physically present and store the semantic graphs and the expressive mechanisms for presenting more complex semantics such as regular graph structures and temporal information. Further, we would introduce the new meta-semantic relations to represent the knowledge rules and the natural language semantics to build a knowledge
base. A full-fledge query language is also under development to explore the graph of semantic information and extract meaningful results.

References


Access to Billions of Pages for Large-Scale Text Analysis

Peter Organisciak, Boris Capitanu, Ted Underwood, J. Stephen Downie
University of Illinois at Urbana-Champaign

Abstract

Consortial collections have led to unprecedented scales of digitized corpora, but the insights that they enable are hampered by the complexities of access, particularly to in-copyright or orphan works. Pursuing a principle of non-consumptive access, we developed the Extracted Features (EF) dataset, a dataset of quantitative counts for every page of nearly 5 million scanned books. The EF includes unigram counts, part of speech tagging, header and footer extraction, counts of characters at both sides of the page, and more. Distributing book data with features already extracted saves resource costs associated with large-scale text use, improves the reproducibility of research done on the dataset, and opens the door to datasets on copyrighted books. We describe the coverage of the dataset and demonstrate its useful application through duplicate book alignment and identification of their cleanest scans, topic modeling, word list expansion, and multifaceted visualization.

Keywords: non-consumptive research; feature extraction; large-scale text analysis; datasets; text mining


Copyright: Copyright is held by the authors.
Contact: organis2@illinois.edu.

1 Introduction

Individual and consortial library digitization efforts around the world have been scanning a massive number of books and other library items. Projects such as Google Books and the HathiTrust have resulted in non-trivial, petascale collections of the world’s published cultural heritage. Such digitization efforts are remarkable for the span of time they represent, often hundreds of years, and their span of cultures. For example, the HathiTrust collection, housed at the University of Michigan, alone comprises 14.7 million volumes taking up 659 TB of disk. Though library digitization efforts are primarily intended for preservation and document access, they also pave the way to new forms of large scale research. As data, they provide for a wealth of linguistic, cultural, historic, or even structural insights, providing researchers evidence for modelling language across more sub-facets and broad time periods. We present the HTRC Extracted Features (EF) dataset, a public research-oriented dataset of page-level features extracted from 4.8 million digitized public-domain books (referred, generally, as volumes) provided by the HathiTrust Digital Library. The EF dataset is notable because of its scale, the ease of access allowed by its non-consumptive design, and the ease of use for reproducible research enabled by its preprocessed and cleaned format.

The EF dataset is a general-purpose dataset, oriented toward research uses that necessitate a long view of the published word or breadth across different topics or languages. Here we demonstrate a number of those uses for information science: duplicate book alignment and identification of cleanest scans, topic modeling, word list expansion, and multifaceted visualization.

With typical research datasets the text analysis process starts with feature extraction, followed by computation over those features (e.g. modeling, counting). In contrast, the EF dataset has already completed the costly first part of that process. In addition to eliminating the effort- and resource- costs associated with feature extraction, this type of dataset allows for more readily reproducible experiments. While the EF dataset is currently derived from works that are in the US public domain, there is also a pragmatic access reason for a feature dataset: it abstracts away from the original full-text in a way that does not redistribute access-restricted scans, and offers a roadmap toward datasets from texts that are copyrighted or of unknown status (orphan works). Such use is referred to as non-consumptive because it cannot be enjoyed in a traditional sense by a person, but little is lost from a computational point of view.

The most important markers of a text’s meaning are the words, and it follows that of the various features offered in EF, the most broadly useful ones are term frequency counts. The EF dataset offers such
counts at the page-level for each page of each volume, tagged by the part-of-speech in the context that they are used. Though positional information is not included, these bags-of-words (BOW) are in a small enough context that they can be quite discriminating in informing a scholar what a page is about.

Particularly useful for clean use of the data, token counts are disambiguated by head, body, and footer on the page. Token counts and other features are noted by section, which makes it very easy to focus solely on content of a page without confounding textual information such as page titles, chapter titles, and page numbers. Other features provided in EF include counts of sentences, lines, and empty lines on the page, page-level language inference, counts of characters that occur at the start and end of lines, and a count of the longest length of capital characters starting a line.

A massive but granularly facetable corpus of scanned texts can support numerous information science uses. In this paper, we demonstrate using EF to:

- **Identifying duplicate books and selecting a best-scanned copy**: By processing a collection of literature in EF into smaller-dimensional ‘fingerprints’, we show that pairwise similarity between books is tractable for identifying multiple copies of a book in the HathiTrust collection, linking a book not only to its identical printings but to different publications of that book. Not all texts are digitized equally, so we introduce a method to surface candidates for the cleanest copy of a book.

- **Topic modeling concepts across books**: We provide a demonstrative example of how EF can support mixed-model soft clustering with Latent Dirichlet Allocation (LDA) in order to find conceptual ‘topics’. Since topic modelling is concerned with coherent conceptual term collocation patterns, various training approaches can be used with EF, including filtering to more interesting parts of speech and training at a page-level document frame.

- **Word list generation**: Using a list of related words unfolded from a seed word is a portable way of following higher-level concepts in a text. It is also used in information retrieval for query expansion, expanding searches beyond the exact keyword of a query. By leveraging a topic model built on literature, we show a simple case of word list generation derived from EF.

- **Multifaceted visualization**: To ease exploratory analysis against the overwhelming scale of 5 million books, we built a multifaceted, interactive visual interface to the EF. Built on top of the open-source tool Bookworm, our publicly-accessible implementation can display trends in different subsets of the data; for example, comparing the use of the word ‘lady’ in British versus American texts.

The possibilities for use of the EF dataset extend well beyond those that we demonstrate. In information retrieval, for example, it can assist retrieval by augmenting collection models for historical or domain-specific collections, or be used to training structural classifiers for book-parsing to improve index. The EF dataset is also appropriate for supporting research in full-text book search, a direction of research that has been impeded by access to large corpora, the variability of documents, and difficult of evaluation (Kazai, Kamps, Koolen, & Milic-Frayling, 2011). EF’s accessibility gives it potential as a test book collection, though its scale and breadth makes it particularly viable for use in developing models of text in different domains, time periods, and languages. In cataloguing, the broad coverage of the dataset can be used for outlier detection to find possibly misclassified texts. Beyond higher-level modelling, we anticipate the value of studying the content itself, for scaling up research questions in computation social sciences and the digital humanities. A deliberate use of EF can follow discourse of a topic over time, look at the rise of a cultural trend or linguistic shift, or observe how the structure of the book has changed.

2 Dataset

The HTRC Extracted Features dataset covers slightly over 4.8 millions volumes, comprised of 1.8 billion pages. Each volume is represented as an individual file, structured in JSON and accessible compressed using the rsync utility that is common on Unix-like systems. Details for access are available at http://dx.doi.org/10.13012/J8X63JT3.

The volumes represented in the EF dataset are from the HathiTrust Digital Library, a consortium of institutions collecting their digitized collections into a single digital library. This prominently includes
Table 1: Most-represented classes (Library of Congress)

<table>
<thead>
<tr>
<th>LC class</th>
<th>count of volumes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Language and Literature</td>
<td>385044</td>
</tr>
<tr>
<td>General and Old World History</td>
<td>244976</td>
</tr>
<tr>
<td>Social Sciences</td>
<td>212951</td>
</tr>
<tr>
<td>Science</td>
<td>184441</td>
</tr>
<tr>
<td>Philosophy, Psychology, and Religion</td>
<td>162939</td>
</tr>
<tr>
<td>Law</td>
<td>126709</td>
</tr>
<tr>
<td>Technology</td>
<td>120953</td>
</tr>
<tr>
<td>General Works</td>
<td>108515</td>
</tr>
</tbody>
</table>

Table 2: Most-represented languages

<table>
<thead>
<tr>
<th>Language</th>
<th>Count of Volumes</th>
</tr>
</thead>
<tbody>
<tr>
<td>English</td>
<td>2705542</td>
</tr>
<tr>
<td>German</td>
<td>564463</td>
</tr>
<tr>
<td>French</td>
<td>528817</td>
</tr>
<tr>
<td>Spanish</td>
<td>142874</td>
</tr>
<tr>
<td>Italian</td>
<td>125561</td>
</tr>
<tr>
<td>Latin</td>
<td>114643</td>
</tr>
<tr>
<td>Japanese</td>
<td>73447</td>
</tr>
<tr>
<td>Russian</td>
<td>59936</td>
</tr>
</tbody>
</table>

volumes from libraries and other institutions across the US, including those scanned by the Google Books project, but also holds contributions from non-US institutions. The underlying materials were scanned and their full text is parsed from those scans. The EF dataset does not share full text, however: only the quantitative feature counts that may be needed in computation analysis or modeling of the volumes.

In the interest of long-term preservation, the EF dataset release is maintained with a DOI (digital object identifier) through the University of Illinois Library, which is intended to persistently point to the current hosting for the EF data. The dataset is licensed with a Creative Commons Attribution License, which hews closely to academic convention by allowing any form of usage or redistribution in return for attribution.

2.1 Coverage

There are 512.1 billion unigrams across the 1.8 billion pages of the EF dataset. The version of the dataset described here is specific to public domain works, but this was a temporary restriction: we have released an expanded dataset including in-copyright and orphan works, three times larger, since this paper was accepted for publication.

The EF dataset covers 344 languages. As the source materials are primarily digitized from US-based academic libraries, English is the best-represented language, with 58.5% of the collection, though German, French, Spanish, Italian, and Latin all have over 100000 texts represented. The computed part-of-speech tags are only accurate for a subset of the languages.

Temporally, 95% of the dataset covers documents published between the years 1722 to 1987. Figure 1 shows the date distribution. Dates are taken from metadata records for the represented volumes. A troublesome but common classification quirk in libraries is confounding accurately classified dates with century-floored approximate dates (e.g. entering 1800 to denote a 19th century text rather than one published in exactly that year), and our use of bibliographic metadata means these errors are retained in the EF dataset.

Cross-referencing the texts with bibliographic metadata from the HathiTrust, we can find Library of Congress classifications for approximately 49% of the volumes, showing that the texts span all 21 top-level classes. The best-represented is class P (Language and Literature).

The corpus underlying the EF dataset benefits from a mostly indiscriminate digitization policy, meaning that much of the collection cuts broadly across the holdings of the participating institutions. There
are benefits to smaller carefully curated digital collections, like the ability to correct individual OCR or 
metadata errors. At the scale of the HathiTrust many such fixes are intractable; instead, individual errors in 
a book are smoothed over by the larger statistical significance afforded.

Despite the broad coverage, there is no reason to assume that the collection is balanced in its 
strengths, and understanding potential biases is useful for proper usage of the dataset. While the digitisation 
was often indiscriminate, the actual holdings of contributing institutions has biases: certain types of text 
are favoured more by academic institutions (the most common type of contributor), and some texts may 
be popular enough that the collection holds duplicate copies from multiple contributors. Figure 2 shows 
the distributions of the top classes within a selection of languages. Note that the classes are distributed 
differently; for example, Spanish is proportionally well-represented by literature, Latin by philosophy, and 
German by science. It is unknown how much of these distributions represents collection bias (what materials 
were held) and which represent the distribution of what is published in that language.

A more significant change in distribution representation happens due to copyright status. All the 
Volumes represented in the EF dataset describe here are US public domain, leading to some some caveats 
about the dataset. Copyright determinations in the US vary depending on the circumstances of the work, 
but works published before 1923 are generally in the public domain. As a result of that year’s transition 
from universal to contextual rules, there is a drastic shift in collection coverage at that point. This is seen in 
the quantity of volumes represented (see 1923 in Figure 1) and in the genres that are seen. For example, 
volumes in the sciences and social sciences increase proportionally from pre-1923 to 1923-, while history falls 
and literature falls precipitously.
3 Related Work

The Google Books Ngrams Corpus (Lin et al., 2012; Michel et al., 2011) provides token counts for n-grams, from unigrams to 5-grams, that occur in volumes scanned by the Google Books project. It is comprised of a similar breadth of materials as EF; indeed, a notable portion of the corpus underlying the EF dataset is from Google Books. Where our dataset differs is in format, providing counts for each page of each volume, while Google’s dataset only provides corpus-level counts though with longer phrases. The NGrams corpus includes information on copyrighted volumes, something not yet released publicly for EF. Finally, the EF dataset includes useful cleaning, such as the header/footer extraction, and provides additional features beyond ngrams.

Data for Research (DfR) from JSTOR (Burns et al., 2009) is another notable historical resource. DfR provide document-level n-gram counts for 1-4 gram counts, as well as \( TF \times IDF \) weighted lists of discriminatory terms on the documents. These can be downloaded for up to one thousand documents freely, or more with permission. The JSTOR collection holds primarily academic materials with a strength in digitized articles. The EF dataset differs in the scope of the collection, spanning published work more general, and has a different access model, with open access to preprocessed features of the entire collection.

Since the EF dataset was publicly released in 2015, with a smaller demonstrative dataset a year earlier, it has seen some researcher use and redistribution of recombinant parts. Underwood (2014; 2015) inferred genre labels for fiction, poetry, and drama; Forster (2015) inferred gender for the authors of a selection of literature in the collection; Goodwin (2015) trained mixture models of commonly occurring themes in fiction; finally, Mimno (2014) calculated co-occurrence tables for terms that co-occur in each year from 1800 to 1923.

4 Features

The EF dataset contains extracted information about digitized volumes as well as a small amount of metadata. The metadata includes the publication date (\( pubDate \)), title (\( title \)), bibliographic language (\( language \)), imprint information about the publishing context (\( imprint \)), and a set of identifiers for different contexts (\( id \), \( htBibUrl \), \( handleUrl \), \( oclc \)). The primary purpose of the dataset is features, so additional metadata must be obtained from secondary sources like the HathiTrust Bibliographic API \(^1\).

At the level of the volume, the only feature provided is \( pageCount \), a count of pages in the volume. Other features are provided at the page-level.

At the level of the page, we provide information by section: header, body, and footer. Headers and footers often contain information that is paratextual – related more to the structure of the book rather than

\(^1\)https://www.hathitrust.org/bib_api
the core content—such as headings, titles, and page numbers. They also tend to repeat over multiple pages, resulting in skewed word distributions. Headers and footers are processed using a custom two-pass algorithm that looks for recurring text at the top and bottom of each page.

For each section of the pages, we provide counts for `tokenPosCount`, `tokenCount`, `sentenceCount`, `lineCount`, `emptyLineCount`, `beginLineChars`, `endLineChars`, while `capAlphaSeq` is provided exclusively for the body of each page.

`tokenPosCount` provides an unordered list of all occurring tokens in that section, with counts. Counts are provided by part of speech and are case-sensitive, so ‘Jaguar’ (proper noun), ‘jaguar’ (noun), and ‘Jaguar’ (noun) are disambiguated, as are ‘rose’ (verb) and ‘rose’ (noun). The tokenization and part-of-speech tagging is done by OpenNLP (Apache, 2005), with the part of speech tags following those of the Penn Treebank (Marcus, Marcinkiewicz, & Santorini, 1993). `tokenCount` further provides the total count of tokens in the given section of the page, for convenience.

`sentenceCount` and `lineCount` provide counts of sentences and lines, where the former refers to the textual content while the latter refers to the physical structure. Sentence segmentation is done by Apache OpenNLP (Apache, 2005). Sentences that started on a different page are still counted, meaning that a sentence spanning a page break will be counted once for each page. Lines refer to the vertical lines of text physically on the scanned page. Additionally, `emptyLineCount` describes the number of lines that do not contain any content. This is interpreted based on the OCR process for a scanned page, which may vary for volumes scanned by different sources. Multiple consecutive empty lines are not counted, so empty line count in many cases is a proxy toward inferring the number of paragraphs on a page (i.e. `count + 1`).

`beginLineChars` and `endLineChars` count the characters along the left-most and right-most margins of a page, respectively. This information is useful for identifying the type of text on the page (Underwood, 2014). For example, lines of poetry may start with capitalized characters and end with punctuation frequently, prose may have a varied distribution of characters, or a table of contents may have many numeric values at the end of a line.

Finally, `capAlphaSeq` counts the longest length of consecutive alphabetical characters in the given section. Again, this information provides hints as to what type of content is on the page, and lock sequences of capital letters suggest back of the book indexes or title pages.

Provided at the page-level but not separated by section is an inferred language field. Even though there is a bibliographic language classification for each volume, there are instances where it may not be correct, or a book may have multiple languages within it. For this reason, the `languages` feature provides language likelihoods for each page, inferred by software from Shuyo (Shuyo, 2010).

Figure 4 shows the ratios of different notable characters at the start and end of lines, through a book, demonstrating the ability of these features to discriminate between parts of books and eventually between text and paratext. The information shown here is unsupervised, without annotation of what front matter is, what a table of contents is, or what prose is. Still, we see indicators of when books look different at the start and end, where the majority of paratext tends to occur. In Figure 4, we see that uppercase characters on the left-most side of a page are much more common at the end of a book. At the same time, seeing punctuation
or digits at the right-side side of a page is an indicator of the type of content we see at the start and end of a book: perhaps table of contents or a back of the book index. This form of information can be used robustly for classification; for example, it has been used to infer book genre (Underwood, 2014).

5 Tools

Since it is publicly available via rsync, structured in JSON, and permissively licensed with the CC-BY Attribution license, the EF dataset can be accessed by anybody and used however they may desire. To aid usage of the dataset, the HTRC Feature Reader library has been released for Python (Organisciak & Capitanu, 2016). The primary goal of this library is to simplify in-memory use of the EF dataset and to provide scaffolding for using it efficiently within the popular Scipy Stack for scientific work using Python.

6 Demonstrative Use

To demonstrate use of the EF dataset, we performed a duplicate text alignment and selection of best copies, topic modeling, word list generation, and multifaceted interactive visualization. These are intended as potential but realistic uses.

6.1 Similarity Between Texts

The EF holds features for each digitized volume copy in the corpus, which includes duplicate copies as well as reprints of the same book. A scholar may want to identify these texts, either to connect a text through its reprints, make sense of what texts are in an anthology of works, or to filter an analysis to only one copy of each text. The information held in the EF is enough for this task, enabling candidates for duplicate or overlapping texts to be surfaced.

We performed a duplicate candidate evaluation on a set of literature texts ($n=101947$) (Underwood et al., 2015). Specifically, 33 works known from metadata records to have at least twenty duplicate copies were sampled, from which a ‘target’ text was randomly selected for each work. A similarity ranking was then performed to measure Precision at 20: how many of the twenty most similar texts are the same as the target text.

For this demonstration, we measured similarity by reducing each text to a smaller dimensional representation and performing a similarity measure to find the most similar texts to a target. Specifically, Latent Semantic Analysis was used to interpret dimensions against tf*idf weighted term-document matrices, and euclidean distance was used to measure similarity.

Since similarity was judged using only content from the books, the ground truth was exact metadata matches augmented with hand-checking. The hand-checking was necessary because sometimes the metadata is inconsistent, with typos or variant spellings. One target text was *Gulliver’s Travels*, for example, a recently popularized title for a book that was actually published as *Travels into Several Remote Nations of the World. In Four Parts. By Lemuel Gulliver, First a Surgeon, and then a Captain of Several Ships*. An exact title match would not catch that a book by the latter title is nonetheless the same work as a book with the former title.

Measured on 33 sample works with at least 20 known duplicates in the 102k volume test corpus, the average precision at twenty for finding identical texts is $P@20 = 0.748$. While this means that 74.8% of matches are completely the same book at the target, the others are not completely unrelated. 2.8% of the texts returned are published subsets of the target book, while 10.8% are different books by the same authors as the target book.

Since duplicate texts occur close together in an EF-trained reduced dimensional space, this can be utilized for a basic selection process for the best scanned copy of a text, which is to say the cleanest, with regards to the OCR quality. We attempted this by averaging a centroid between all the known duplicates of a book, and identifying the book closest to the centroid.

Against 663 works with duplicate copies (totalling 15084 copies), we evaluated the “book closest to the centroid” selection policy by vocabulary size. OCR errors lead to increased numbers of unique words, so we would expect that worse copies of a book would have more unique terms. For our best-copy selection
process, we find that the top candidate has a median 1403 unique words less than the bottom candidate. Figure 5 shows this relationship to vocabulary holds for the top and bottom candidates.

6.2 Topic Modeling

Topic modeling refers generally to mixed model clustering trained on term co-occurrences, most popularly built using Latent Dirichlet Allocation (LDA) (Blei, Ng, & Jordan, 2003). The dimensions of a well-trained model can show conceptual coherence, allowing them to be interpreted qualitatively as conceptual topics. Topic Modeling is possible with the bag-of-words information provided in the EF, and can benefit from the page-level granularity, stripped headers, and POS tagging in the dataset. We built an example model making use of these features and aided by randomized page sampling and asynchronous prior assumptions on individual topic probabilities.

As topic cohesion is desired with topic modeling, the training process is more discriminating than for dimensionality reduction, as done for our book similarity demonstration earlier. Dimensionality reduction aims to accounts for the most amount of variance in any way possible, but for topics we often strive for conceptually informative groupings of words. For our demonstration of topic modeling, we filtered out a number of word types that are not as interested for generalized topic models, including determiners, personal pronouns, modals, symbols and different forms of adverbs.

Our topic modeling demonstration was built with 400 topics. Topic models are trained on word co-occurrences, so it is better to train with many examples on a smaller document frame than on full books. We used pages as the training frame. After all, a word occurring on the first page of a text may not necessarily reflect a strong relationship to a word on the last page. Another technique we used was randomized, multi-pass sampling. Earlier training texts can exert outsized bias on a topic model, so we trained with multiple passes across the collection, each one sampling only a small part of a book. The first pass only used 1 or 2 pages from each book (specifically, 1/256) to avoid a convergence of topics too early, and subsequent passes gradually increased the per-book page sample size. The choice of training 400 topics was not made by any detailed method. It was intuitively chosen with the motivation of training excess topics, because it is easier to systematically ignore uninteresting extra topics than it is to recover interesting ones that never were trained. Finally, we trained with a set of asynchronous document-topic probabilities that put the majority of topical probability mass in the first few topics. This has the effect of serving as a ‘catch-all’ for words that are common across the language in general, allowing niche concepts a place for their own less common and often more interesting patterns, as well as a way to identify the less interesting ones.

In LDA, each model can be interpreted as a process that generates words, with different distributions for each term’s likelihood to be generated by that model. For example, topic 359 is most likely to generate the words ‘sin’, ‘Christ’, ‘grace’. Subsequently a text can be assigned a distribution of how likely each topic is to have generated the words in that text. The resulting topics can be used to interpreted in the contexts of
Figure 6: Top topics representing various terms.

<table>
<thead>
<tr>
<th>love</th>
<th>grief</th>
<th>forest</th>
<th>city</th>
<th>philosophy</th>
<th>cat</th>
<th>night</th>
</tr>
</thead>
<tbody>
<tr>
<td>passion</td>
<td>sorrow</td>
<td>sunshine</td>
<td>streets</td>
<td>philosopher</td>
<td>stuff</td>
<td>light</td>
</tr>
<tr>
<td>loving</td>
<td>dream</td>
<td>branches</td>
<td>arms</td>
<td>ideas</td>
<td>stick</td>
<td>day</td>
</tr>
<tr>
<td>lovers</td>
<td>breast</td>
<td>rays</td>
<td>death</td>
<td>exist</td>
<td>leg</td>
<td>morning</td>
</tr>
<tr>
<td>true</td>
<td>weep</td>
<td>danger</td>
<td>sorrow</td>
<td>cases</td>
<td>pockets</td>
<td>sky</td>
</tr>
<tr>
<td>hate</td>
<td>weeping</td>
<td>conception</td>
<td>conception</td>
<td>pockets</td>
<td>feet</td>
<td>stars</td>
</tr>
<tr>
<td>beauty</td>
<td>anguish</td>
<td>tricks</td>
<td>wish</td>
<td>desire</td>
<td>looting</td>
<td>dark</td>
</tr>
<tr>
<td>hearts</td>
<td>despair</td>
<td>groove</td>
<td>words</td>
<td>intellect</td>
<td>pull</td>
<td>earth</td>
</tr>
<tr>
<td>tender</td>
<td>bosom</td>
<td>atmosphere</td>
<td>tower</td>
<td>individual</td>
<td>bit</td>
<td>darkness</td>
</tr>
<tr>
<td>wisest</td>
<td>bitter</td>
<td>plant</td>
<td>phrase</td>
<td>consists</td>
<td>cage</td>
<td>bright</td>
</tr>
</tbody>
</table>

Figure 7: Word lists for a selection of emotional, topical, and setting-based seed words.

texts (e.g. “what are the concepts that make up Pride and Prejudice?”), in a global context (e.g. “what are the different types of topics that we see in literature?”), or at a word level (e.g. “what topics are likely to generate the word “love”?”).

Figure 6 shows the top topics for three terms: ‘love’, ‘sad’, and ‘queen’. A qualitative interpretation would suggest that the topics are exhibiting some manner of coherence, such as love and God, love and passion, love and sorrow. Since verbs were not filtered, the top topics for ‘love’ include a topic of less interesting generally distributed words like ‘is’ and ‘be’; however, since we trained a catch-all topic, it is possible to programmatically filter such topics by measuring their similarity to the catch-all topic using a probabilistic distance metric like Hellinger Distance.

### 6.2.1 Word List Generation

Topic modelling can be further leveraged for a use common in information science: expanding a seed word into a list of related terms. In information retrieval this is referred to query expansion, which is used to match search results with appropriate terms beyond the exact query string that an information-seeking user typed into their query. In other areas, words lists are used as a convenient way to track concepts across a text, easier to transfer across collections and easier to interpret than topic models. One popular set of word lists is LIWC (Pennebaker, Francis, & Booth, 2001).

By normalizing word probabilities across topics and comparing their distance using Hellinger Distance, a seed word can be exploded into a list by identifying the words that occur across all topics most similarly to the seed. Figure 7 shows word list for a selection of keywords.

Previous work has leveraged word embedding models for word list generation (Fast, Chen, & Bernstein, 2016), which learn words by their immediate contexts. Skip-gram word embeddings are more conceptually appropriate for word list generation as their goal is to predict context words from a seed word; however, the EF dataset does not provide the positional information necessary for training a clean word embedding model.

### 6.3 Multi-faceted Visualization

A challenge to working with text data at the scale of EF is that even preliminary exploration is a time-consuming process, presenting a challenge to inductive inquiry. There is a value to being able to quickly
explore trends throughout the collection to assess their tractability as a study topic. To support exploratory data analysis, we adapted the EF dataset for the collection visualization tool Bookworm, an evolution of work presented in (Michel et al., 2011). This implementation of Bookworm on the EF dataset is available publicly

On the EF implementation of Bookworm, it is possible to observe longitudinal trends across all 5 million texts, or subfacets such as publication country or class (Figure 8 top and middle). It is also possible to observe trends that are not year-based (Figure 8 bottom), or even to query the system backend for the raw numbers to visualize yourself.

While the metadata was augmented from additional sources, the data underlying this visualization tool is from EF. One of the limitations here is that the solely unigram token counts in the EF keep the visualization from allowing phrases as search queries.

7 Conclusion

The EF dataset is designed to support a breadth of different research questions by providing access to millions of books in an open and straightforward way. Its strengths lie in the coverage of its collection – multiple

\[\text{http://analytics.hathitrust.org/bookworm}\]
languages, varied domains, and spanning hundreds of years – as well as its preprocessed features format, which saves time and computational resources while also providing a standardized foundation for supporting different research needs. These circumstances make the EF dataset valuable for large-scale textual needs, such as topic modeling and similarity measurements between books.

The principle guiding the creation of the Extracted Features dataset is that of non-consumptive access, which seeks ways to nurture effective large-scale text research within the constraints of intellectual property laws. Recent work on the EF dataset has released the same features for 13.7 million volumes, including those that are in-copyright or of unknown status. This release was made possibly only by the non-consumptive structure of the texts.

We demonstrated the malleability of the EF dataset for text mining and analysis through a selection of example uses. Though the public release of this dataset, researchers in many domains can leverage it in pursuit of their own work.

References


76
Identifying Users’ Gender via Social Representations

Tieyun Qian¹, Peisong Zhu¹, Xuhui Li²*, Dewang Sun²
¹State Key Laboratory of Software Engineering, Wuhan University, Hubei, China
²School of Information Management, Wuhan University, Hubei, China
*Contact author

Abstract
Gender prediction has evoked great research interests due to its potential applications like targeted advertisement and personalized search. Most of existing studies rely on the content texts. However, the text information is hard to access. This makes it difficult to extract text features.

In this paper, we propose a novel framework which only involves the users’ ids for gender prediction. The key idea is to represent users in the embedding connection space. We present two strategies to modify the word embedding technique for user embedding. The first is to sequentialize users’ ids to get the order of social context. The second is to embed users into a large-sized sliding window of contexts. We conduct extensive experiments on two real data sets from Sina Weibo. Results show that our method is significantly better than the state-of-the-art graph embedding baselines. Its accuracy also outperforms that of the content based approaches.

Keywords: gender prediction; users in social media; social contexts; social representations


Copyright: Copyright is held by the authors.

Acknowledgements: The work described in this paper has been supported in part by the NSFC projects (No.61272275, No.61572376, No.61272110, No. 91646206), and the 111 project(B07037).

Contact: {qty,lixuhui}@whu.edu.cn

1 Introduction

Gender prediction has attracted a great deal of research attentions (Cheng, Chen, Chandramouli, & Subbalakshmi, 2009; Mukherjee & Liu, 2010; Otterbacher, 2010; Peersman, Daelemans, & Vaerenbergh, 2011; Filippova, 2012; Bergsma & Durme, 2013; Xiao, Zhou, & Wu, 2013) in recent years due to its potential applications like targeted advertising and personalization. Almost all existing methods use the content texts to build the feature vector for classification (Cheng et al., 2009; Peersman et al., 2011; Filippova, 2012; Bergsma & Durme, 2013). It is often difficult to access the text such as microblogs or reviews due to the restriction of the websites. More importantly, there are a large number of users in social media who register only for browsing, i.e., they do not have contents. For instance, a sample of 1 million users from Sina Weibo in China shows that about 7.4% users do not post any message. For this kind of users, it is impossible to get any content features for analysis.

In this paper, we present a novel approach for gender classification which uses no content features. The key idea is to learn the social representation from the relations among users. In social media, there are two fundamental social relations, i.e., following and being followed relations. For simplicity, we will use friendOf to represent both relations, and use friends to represent all users in one user’s connection list, including family members, shoolmates, etc. For example, a sequence of “2: 8 6 4 10” represents that user 2 has friendOf relations with four users 8, 6, 4, and 10.

Existing methods for representing social relations include the traditional graph based representation (TGR for short) (Culotta, Kumar, & Cutler, 2015) and the recently developed graph embeddings (GE for short) (Perozzi, Al-Rfou, & Skiena, 2014; J. Tang et al., 2015). Basically, each friendOf relation is represented as an edge in the graph. Hence for the above example, we will have four edges, 2–8, 2–6, 2–4, and 2–10, as illustrated in Figure 1 (a). We argue that some important information are missing from TGR and GE. Note that, besides the explicit social relations as direct friendOf connections, there are implicit social relations among the friends themselves. For example, user 2 follows users 6 and 10 because they are his/her classmates.
This infers that users 6 and 10 are also classmates. If users 6 and 10 do not follow each other, then there will be no edge between them in the graph.

![Graph representation](image)

**Figure 1:** A sample of (a) graph representation, (b) word embedding

Our proposed approach can capture both the implicit and explicit relations. It builds concepts on word embedding (Bengio, Ducharme, Vincent, & Jauvin, 2003; Mnih & Hinton, 2007; Mikolov, Sutskever, Chen, Corrado, & Dean, 2013). Figure 1 (b) shows an example for the process of word embedding. Supposing the window size is 2, when modeling a sequence of words *(this, book, is, very, interesting)*, each word is treated as the current word, and words within the window (left and right) as contexts, and we have five contexts *(book, is), (this, is, very), (this, book, very, interesting), (book, is, interesting), and (is, very)*. Clearly, compared to the graph representation which only captures the explicit relations, word embedding encodes richer information as it reflects all the implicit relations among words co-occurring in one sentence. Similarly, the implicit social relations among classmates 6 and 10, corresponding to the word *is* and *interesting*, respectively, can also be captured by this method.

The above example illustrates the improvement of word embedding over the graph representations. However, due to the wide gap between linguistic and social contexts, word embedding has limitations when it is used to encode social relations. In language, syntax governs the sequence of words in a sentence. In contrast, the ordering is missing from the users’ ids in social contexts. For example, two classmates 6 and 10 can appear at any position of social contexts. Furthermore, there are a lot of phrases or idioms, shown as local structure in sentences. Hence a small window size like 5 or 10 is usually good enough to capture the local structure in word embedding. However, the related users may be far away from each other in social contexts. To deal with these two problems, we propose two modifications. One is the node sequentializing and the other the large-sized sliding window. The node sequentializing is to map the users’ id into a fixed order so as to eliminate the randomness in the neighborhood. The large-sized sliding window aims to enclose users in a long distance yet from same community into one context.

To the best of our knowledge, we are the first to adapt the word embedding approach to exploiting social relations for social representations. Our proposed method has the following key properties.

- It presents a new model to encode social relations which involves only users ids, while most of existing approaches rely on the contents to build feature vectors.
- It captures all kinds of social relations among users, while graph embedding techniques consider only explicit relations between two users.

The rest of this paper is organized as follows. Section 2 reviews related work. Section 3 introduces our approach for learning social representation for gender prediction. Section 4 introduces the data sets used for evaluation. Section 5 provides experimental results. Section 6 concludes the paper.

## 2 Related Work

We review the literature in this section, organized by the feature set, the word embedding and graph embedding techniques, and the classification method.

### 2.1 Feature set

In the area of gender classification, the word or character n-grams are the most widely used features (Peersman et al., 2011; Filippova, 2012; Burger, Henderson, Kim, & Zarrella, 2011; Bergsma & Durme, 2013; Cheng
et al., 2009). There are also a number of stylistic features extracted from the content, including the ratio of punctuation, capital letters, unique words (Filippova, 2012), slang words (Goswami, Sarkar, & Rustagi, 2009), word or sentence length (Filippova, 2012; Goswami et al., 2009), conceptual class (Bergsma & Durme, 2013), and the part-of-speech (POS) sequence (Mukherjee & Liu, 2010).

Almost all existing studies rely on the content information. The only exception is the neighbor vector representations (NVR) approach in (Culotta et al., 2015) which directly utilizes the network information. We will use that as a baseline to show the improvements of our embedding method.

2.2 Word embedding and graph embedding

Word embedding has shed lights on many nature language processing (NLP) tasks with the development in deep neural network. The typical techniques include NNLM (Bengio et al., 2003), LBL (Mnih & Hinton, 2007), CBOW and SkipGram (Mikolov et al., 2013). We adapt SkipGram to our task because it performs better than CBOW and also because it significantly speeds up the training process of NNLM and LBL.

Graph embedding is a classic problem. Traditional approaches like Isomap and Laplacian EigenMap have a quadratic time complexity to the number of the nodes. In recent years, researchers proposed GF using stochastic gradient descent (Ahmed, Shervashidze, Narayanamurthy, Josifovski, & Smola, 2013), LINE using edge sampling (J. Tang et al., 2015), and DeepWalk using random walk (Perozzi et al., 2014) for large scale network embedding. Among these, the LINE method achieves the best performance, and hence we use it as one of the baselines.

2.3 Classification method

A number of machine learning approaches have been explored to solve the problem of gender prediction, for instance, SVM (Rao, Yarowsky, Shreevats, & Gupta, 2010; Cheng et al., 2009; Mukherjee & Liu, 2010; Peersman et al., 2011), decision trees (Pennacchiotti & Popescu, 2011; Cheng et al., 2009; Alowibdi, Buy, & Yu, 2013), Naive Bayes (Mukherjee & Liu, 2010; Goswami et al., 2009; C. Tang, Ross, Saxena, & Chen, 2011; Alowibdi et al., 2013), logistic regression (Bergsma & Durme, 2013; Bamman, Eisenstein, & Schneebelen, 2014), the Winnow algorithm (Burger et al., 2011; Schler, Koppel, Argamon, & Pennebaker, 2005), and the maximum entropy learner (Filippova, 2012).

The classification method is not the focus of this paper. In our study, we choose to use LR as our base classifier since LR is not as sensitive to parameters as SVM and also because it performs well.

3 Learning Social Representation

In social media, users are connected with their family members, friends, schoolmates, colleagues, or people with similar interests. All these connections (neighbors) form social contexts. Furthermore, being friendOf with a same user, these connections may belong to a same community. For example, a user a follows her classmates b and c. Then “b and c” are the user a’s social context, and users a, b, and c form a community “classmate”. We can further deduce that if there exists another user d in this class, then d may have a social context of “a, b, and c”. The more times users appear in the same social contexts, the stronger relations are there among these users, and the larger probability they belong to the same community. Such an observation inspires us to borrow ideas from SkipGram (Mikolov et al., 2013), a recently developed word embedding technique which captures the semantic and syntactic relations among words.

3.1 Preliminary on SkipGram

The objective of SkipGram is to maximize the co-occurrence probability among the words that appear within a window in a sentence. More formally, we can define the objective function as:

$$ L = \sum_{w \in C} \log p(\text{Context}(w) | w) = \sum_{w \in C} \log \prod_{u \in \text{Context}(w)} p(u | w) $$

(1)
where $C$ is the corpus used for training. By applying the Huffman tree based hierarchical softmax (Mikolov et al., 2013), we can rewrite $p(u|w)$ as:

$$p(u|w) = \prod_{j=2}^{d_w} p(d_j^u | v(w), \theta_{j-1}^u)$$

$$p(d_j^u | v(w), \theta_{j-1}^u) = [\sigma(v(w)^T \theta_{j-1}^u)]^{1-d_j^u} \cdot [1 - \sigma(v(w)^T \theta_{j-1}^u)]^{d_j^u}$$

(3)

where $v(w)$ is the $d$-dimension vector for the central word $w$, and $d_j$ and $\theta_j$ is the Huffman code (either 0 or 1) and the $d$-dimension vector for the $j_{th}$ node on the path $p^w$, respectively. Equation (1) can then be optimized using gradient descending technique. The procedure is shown in Algorithm 1.

**Algorithm 1: SkipGram**

$$\text{SkipGram}(\Phi, b, \text{Context}(u_i), s, \eta)$$

1. for each $v_j \in \text{Context}(u_i)$
2. for each $u \in \text{Context}(u_i)[b, b+s]$  
3. $L(\Phi) = \log \prod_{u \in \text{Context}(u_i)} p(u|\Phi(v_j))$ 
4. $\Phi = \Phi - \eta \frac{\partial L}{\partial \Phi}$
5. end for
6. end for

3.2 Adapting SkipGram to user embedding

SkipGram is designed for word embedding. A typical scenario to use SkipGram is upon a corpus, where each word is naturally embedded in a paragraph or document. However, as analyzed in the previous section, the users in social media do not occupy such an characteristic. Furthermore, the users with local structure or community may be far apart from each other. Hence we present two strategies to adapting SkipGram to user embedding, i.e., node sequentializing and the large-sized sliding window.

3.2.1 Node sequentializing

Node sequentializing is the process of identifying, for each social context, a sequence of nodes for which the neighbors of a node are created, like the syntax governing the sequence of words in a sentence. The simplest sequentializing is just to use the nature order of ids. For example, user 1 precedes user 2, and user 2 precedes user 3. This method can eliminate the randomness of neighbors. However, the id information is irrelevant to the inherent structure of the network. Hence we present the following degree-based sequentializing method.

Given any two nodes $i$ and $j$, and their $k$-th layer of neighbors $N_k(i)$ and $N_k(j)$ in Graph $G$, the degree-based sequentializing defines a total order $\succ$ on nodes in $G$ which uniquely determines the position $O$ of a node $i$ in a sequence, such that:

- $O(i) \succ O(j)$ iff. $d(i) > d(j)$;
- $O(i) \succ O(j)$ iff. $d(i) = d(j)$, and $d(N_1(i)) > d(N_1(j))$;
- $O(i) \succ O(j)$ iff. $d(i) = d(j)$, $d(N_k(i)) = d(N_k(j))$ ( $k = 1..m-1$ ), and $d(N_m(i)) > d(N_m(j))$.

The $d(i)$ function denotes mapping the node id to the degree of this node. The basic idea is to sort the ids by their degrees. If the degrees of two nodes are equal, then incrementally compare their k-th neighbors’ degrees until a total order is given.
3.2.2 large-sized sliding window

The large-sized sliding window strategy is presented to deal with the problem caused by the randomness of user id, which is automatically assigned by the system and normally correlated with the time users start to use the social media. This means that users’ connections may not have the adjacent neighbor ids. When SkipGram is used in the NLP applications, the window size is often set to 5 or 10. This is definitely not suitable for user embedding because the number of friends is usually quite large. Hence we set the window size $s$ in SkipGram to a large value.

3.3 Algorithm for learning social representation

We can now learn the social representation by applying the SkipGram to the above prepared social contexts. The entire procedure, called as User Embedding (UE), is shown in Algorithm 2.

**Algorithm 2: User Embedding (UE)**

**Input:** The set of users $U = \{u_i\}$ and the friends of users $\{u_i, F(u_i) = \{f_{i1}, \ldots, f_{in}\}\}$

**Parameters:** the window size $s$, the dimensionality of user embedding $d$, and the learning ratio $\eta$

**Output:** matrix of user representations $\Phi \in \mathbb{R}^{\vert U \vert \times d}$

**Steps:**

1. for each user $u_i$
2. $A(u_i) = \{u_i\} \cup F(u_i)$
3. end for
4. Sequentializing ids in $F(u_i)$ and $A(u_i)$
5. for each user $u_i$
6. for the $j$th friend ($j = i_{1:i_n}$) in $F(u_i)$
7. SkipGram($\Phi, j, F(u_i), s, \eta$)
8. end for
9. SkipGram($\Phi, 1, A(u_i), \vert A(u_i) \vert, \eta$)
10. end for

In Algorithm 2, lines 1-3 initiate the explicit social contexts. Line 4 sequentializes the ids. Lines 5-10 build the user embedding. Specifically, lines 6-8 iterate on the implicit social contexts and line 9 on the explicit social contexts.

There are three parameters tunable in the UE algorithm: $\eta$ the learning ratio, $d$ the dimensionality of embedding vector, and $s$ the size of contexts (also called the window size). Among which, $\eta$ is related to the training speed and we just use the default setting. We investigate the effects of $d$ and $s$ in the experimental part.

4 Data sets

4.1 Data collection

The data is collected from Sina Weibo, which is one of the largest micro-blogging services in China. Each user in Sina Weibo has a profile, which has several fields, such as userid, screen name, gender, tags, description, the number of followers, followees, and messages. Table 1 shows a sample profile of a celebrity in Sina Weibo.

Most of the fields in the user’s profile are optional. Many users choose to leave them blank, and many users do not post any microblogs either. However, due to the very nature of social media, the users
intend to connect with others. This results in a number of connections for each user. We can crawl the uids of friends through the API provided Sina, which can be used for our experiments.

We start from a public domain data set\(^1\) including the profile information of 1 million users. From which, we construct two data sets. One is the mute celebrities and the other the ordinary users.

### 4.2 Mute celebrities

The data set of mute celebrities (MC for short) contains 1280 users (640 female and 640 male users, respectively) who have at most 5 microblogs. We build such a data set due to the reasons below.

- At the beginning, we intend to choose the real mute celebrities from the original public data. However, we only find 21 celebrities meeting this requirement. This is too sparse for performing experiments. Hence we have to expand the data set by relaxing the minimum number of posts to 5.

- We select the mute celebrities rather than the mute ordinary users since it is extremely hard to examine their gender if there is no additional information (e.g., microblogs, photos) for lurkers. In contrast, the gender of celebrities has been verified by Sina and hence there is no need to manually check gender.

This data set is used to validate the effectiveness of our method on mute users, which is the initial objective of our study.

### 4.3 Ordinary users

The data set of ordinary users (OU for short) contains 400 users (200 female and 200 male users, respectively) who have at least 10 microblogs. The number of ordinary users is much less than that of mute celebrities because of the labor-intensive procedure of manually checking. In order to verify the users’ gender in this data set, we recruit three undergraduate students to manually check the data in order to ensure 1) the account is not a spammer or implicit enterprise users like the owner of micro-shops, 2) the user’s gender is real, which is done by keeping only those with more than agreed labels from two students.

This data set is used to evaluate how our method performs on ordinary users. Moreover, we wish to compare our user embedding method with the content based approaches using features extracted from microblogs. This is why we set the condition of the minimum number of 10 microblogs for this data set.

For both the MC and OU data sets, we download up to 500 uids of the friends through the friends API from Sina. By using these uids, we build the implicit and explicit social contexts and construct the embedded uid vectors for our approach. This also forms the first layer graph. We further collect two-hop friends (neighbors’ neighbors) using the same script to construct the second layer graph. This additional part is necessary for the graph-based baselines (see below). Table 2 shows the statistic for two data sets, where the edge and degree in the first and second layer of graph is represented as the superscript of 1 and 2, respectively.

![Table 1: A sample user profile in Sina Weibo](image)

\(^1\)http://www.nlpir.org/?action-viewnews-itemid-232
Table 2: The statistic for two data sets

<table>
<thead>
<tr>
<th>Data set</th>
<th>UserNum</th>
<th>EdgeNum</th>
<th>AvgDegree</th>
<th>UserNum</th>
<th>EdgeNum</th>
<th>AvgDegree</th>
</tr>
</thead>
<tbody>
<tr>
<td>OU</td>
<td>400</td>
<td>66686</td>
<td>166.72</td>
<td>51152</td>
<td>1184586</td>
<td>23.16</td>
</tr>
<tr>
<td>MC</td>
<td>1280</td>
<td>125340</td>
<td>97.92</td>
<td>87202</td>
<td>2722699</td>
<td>31.22</td>
</tr>
</tbody>
</table>

second layer graph is significantly smaller than that in the first layer. The reason is that due to the resource restriction, we only keep the edges in second layer which already have one node in the first layer. The users (nodes) of the remaining edges are two hops away from the original 400 and 1280 users and thus are out of our considerations.

5 Experiments

We conduct experiments on two real data sets as introduced in the previous section. The users are randomly divided into five parts. We perform 5-fold cross validation and the results are averaged over five folds. We use the accuracy as the evaluation metric since both data sets are balanced on two classes.

5.1 Effects of window size

The window size and dimensionality are two key parameters in user embedding. We first investigate the effects of window size and show the results in Figure 2.

![Figure 2: Effects of window size](image)

It is clear that the best performance is achieved at the window size of 150 and 250 for the OU and MC data set, respectively, which is much larger than the small window size of 5 or 10 for the language contexts. Indeed, the accuracy on window size 5 on two data sets is 52.00 and 50.25. Both are the worst. This shows that the large-sized window strategy is appropriate for the user embedding problem. We will use the best window size as our default setting in the following experiments.

5.2 Effects of dimensionality

We investigate the effects of dimensionality by scaling it from 50 to 300. The results are shown in Table 3.

<table>
<thead>
<tr>
<th></th>
<th>50</th>
<th>100</th>
<th>150</th>
<th>200</th>
<th>250</th>
<th>300</th>
</tr>
</thead>
<tbody>
<tr>
<td>OU</td>
<td>79.75</td>
<td>82.50</td>
<td>82.00</td>
<td>82.50</td>
<td><strong>83.00</strong></td>
<td>82.50</td>
</tr>
<tr>
<td>MC</td>
<td>81.88</td>
<td><strong>82.03</strong></td>
<td>81.48</td>
<td>81.95</td>
<td>81.25</td>
<td>81.88</td>
</tr>
</tbody>
</table>

Table 3: Effects of dimensionality

From Table 3, we find the fluctuation of accuracy on MC is less obvious than that on OU. The largest change is 0.78%, showing that it is not very sensitive to the dimensionality. In addition, it can be seen that the best performance is 83.00% and 82.03% on OU and MC, respectively, reached at the dimensionality of
and 100. However, for the fair comparison with the word embedding and graph embedding approaches, we set the dimensionality to 100 in the following experiments, which is the same as those used in (Mikolov et al., 2013; J. Tang et al., 2015).

5.3 Effects of sequentializing

We evaluate the effects of three types of node sequentializing, i.e., by the random order, by the nature order, and by the degree order. The results are shown in Table 4. Note that all the results are under the best window size for each method. It can be seen that the accuracy on MC by the random order is better than that on OU. This contradicts with that by the nature and degree order, showing that the random order is not stable. We can also see that the results by the degree order are the best. The reason may be that in social network, well-connected nodes tend to connect to each other, known as the rich-club phenomenon (Colizza, Flammini, Serrano, & Vespignani, 2006). Using the degree order helps finding the inherent structure in social contexts and thus improves the performance.

<table>
<thead>
<tr>
<th></th>
<th>random</th>
<th>nature</th>
<th>degree</th>
</tr>
</thead>
<tbody>
<tr>
<td>OU</td>
<td>78.00</td>
<td>81.00</td>
<td>82.50</td>
</tr>
<tr>
<td>MC</td>
<td>79.61</td>
<td>79.53</td>
<td>82.03</td>
</tr>
</tbody>
</table>

Table 4: Effects of sequentializing

5.4 Comparison with baselines

We conduct extensive comparison experiments on six baselines, including content-based, graph-based, graph embedding, and user embedding approaches. The description about the baselines are as follows.

- **Word Frequency (WF):** This baseline uses the contents from microblogs. Each word in the microblog is represented as a vector with the dimension as $x:y$, where $x$ is word id and $y$ its frequency. This is the most widely used representation in gender classification.

- **Word Embedding (WE):** This baseline uses the words in users and their friends as the corpus to get the distributed representations of words. Following the practice in (Mikolov et al., 2013), we set the dimensionality to 200, and use the hierarchical softmax for approximation.

- **Original Graph Representations (TGR):** This is the traditional one-hot representation with the entry in users’ vector as $x:y$, where $x$ is the user id and $y$ is 1 or 0, standing for whether $x$ appears in this user’s neighbor.

- **Neighbor Vector Representations (NVR):** This baseline represents each user as a neighbor vector with the dimension as $x:y$, where $x$ is the two-hop neighbor node and $y$ stands for the fraction of its followers that are friends with each of its neighbors. For classification, we strictly follow the settings in (Culotta et al., 2015).

- **Graph embedding method (LINE):** This contains two baselines $\text{LINE}_1$ and $\text{LINE}_{1+2}$, which uses the users, the one- and two-hop friends to construct the network for graph embedding, respectively. We use the default settings in (J. Tang et al., 2015), i.e., the dimension size is 100, the number of negative samples is $K=5$, and the total number of samples is $T=10$ billion.

The results are shown in Table 5. We have the following important notes.

a. Our proposed UE method is the best among all approaches. Its improvements over other baselines are all significant under the 0.05 significant test. Firstly, it achieves a huge enhancement over the graph based NVR baseline with a 16.50% and 23.59% improvement on OU and MC. Secondly, it outperforms two graph embedding approaches by a large margin. For example, the accuracy on OU grows from 73.00% ($\text{LINE}_1$) to 82.50%. The performances of $\text{LINE}_{1+2}$ on both OU and MC are much worse than those of our UE as well. This clearly demonstrates that UE is more effective and resource efficient than the state-of-the-art
graph embedding approach LINE. Thirdly, it is much better the content based approaches WF and WE. This is a very strong indication of the potential application scenario of UE. No matter whether there are text information, UE is a good choice for classification.

b. The traditional one-hot graph representation TGR is the second best. This suggests that a users gender is highly correlated with his/her friends. This finding is interesting, indicating that the majorities in female users’ social networks are female. The same conclusion also holds for the male users. Furthermore, TGR outperforms the graph embedding LINE, suggesting that graph embedding may incur information loss. In contrast, our user embedding approach UE models both the implicit and explicit social relations rather than building a graph for explicit connections and thus is a better representation. We notice that the other graph based method NVR is the worst. The reason may be that NVR only keeps the indirect relations between users.

c. The performance of graph embedding method \text{LINE}_{1+2} is worse than \text{LINE}_1, different from that in (J. Tang et al., 2015). The reason may be arisen from that we do not include the edges whose nodes are two-hops away from the original users. Remember that the average degree in second layer graph is much smaller than that in the first layer, especially on the OC data set. This finding is important in that the second proximity may hurt the performance unless enough nodes and edges are supplemented. However, this needs extra overheads. We also find that WF is better than WE, suggesting that the content based approach does not benefit from the word embedding technique. This can be due to the diverse topic and vocabulary in microblogs.

6 Conclusion

We present a novel social representation to capture both the explicit and implicit relations among users in social media. By modifying the word embedding technique to exploit the social contexts, the proposed approach achieves very accurate results on gender classification. We conduct extensive experiments on two real data sets from Sina Weibo. The results show that our UE method is significantly better than both the traditional graph based approaches and the state-of-the-art graph embedding algorithms. This clearly demonstrates that our approach is extremely useful when the texts are unavailable. Further more, our method also outperforms the content based approaches. This strongly indicates that our method has general applications to all types of users.

In the future, we plan to conduct more experiments on data sets from Twitter or Facebook. We also plan to investigate how our method works on other user profiling tasks.

References


Let’s Workout! Exploring Social Exercise in an Online Fitness Community

Li Zeng¹, Zack W. Almquist², Emma S. Spiro¹
¹University of Washington, USA
²University of Minnesota, USA

Abstract
Increasing attention has been paid to promoting certain healthy habits through social interaction in online communities. At the intersection of social media and activity tracking applications, these platforms capture information on physical activities as well as peer-to-peer interactions. Importantly, they also offer researchers a novel opportunity to understand health behaviors by utilizing the large-scale behavioral trace data they archive. In this study we explore the characteristics and dynamics of social exercise (i.e. fitness activities with at least one peer physically co-present) using data collected from an online fitness community popular with cyclists and runners. In particular, we ask if factors such as temporal seasonality, activity performance and social feedback vary by the number of people participating in an activity; we do so by comparing associations for both men and women. Our results indicate that when peers are physically co-present for fitness activities (i.e. group workouts), exercise tends to be more intense and receive more feedback from other users, across both genders. Findings also suggest gender differences in the observed tendency to complete activities with others. These results have important implications for health and wellness interventions.

Keywords: social exercise; health behaviors; online fitness communities; social media; behavioral traces


1 Introduction
Exercising is known to be associated with numerous physical and mental benefits such as controlling weight (Blair, 1993), decreasing the risks of cardiovascular diseases and reducing stress (Fletcher et al., 1996). Although aware that exercising is good for health, not everyone engages in physical activity on a regular basis. The Center for Disease Control in the United States estimates that only 20% of adults meet exercise guidelines (Jaslow, n.d.).¹ Indeed, the World Health Organization continues to combat the global obesity epidemic (Organization, 2000). Health, and more specifically physical activity, is a complex issue and as such has been studied in many fields. Beyond the medical community, issues of health are studied in the social sciences, where research has examined how social, racial, emotional, and socioeconomic factors influence health promotion and likewise health disparities (Walker, Sechrist, & Pender, 1987; Frankish, Milligan, & Reid, 1998; Lee, Wildeman, Wang, Matusko, & Jackson, 2014; Dishman, Sallis, & Orenstein, 1985).

A number of studies demonstrate that peer influence and social support have positive health-related effects, such as helping people lose weight and participate in more physical activities (Ahtinen et al., 2009; Chen & Pu, 2014; Wing & Jeffery, 1999; Kulik & Mahler, 1989; Dishman et al., 1985). Prior research also points out that exercising with others can improve psychological functioning (Plante, Coscarelli, & Ford, 2001). However, many of these early studies of peer effects are based on surveys or experiments involving a small number of participants. Recently, social media and online fitness communities are gaining scholarly attention as a new research environment in which to study health behaviors (Centola, 2013).

Social media have been used to explore health communication and promotion (Paul & Dredze, 2011; Morris et al., 2011; Vaterlaus, Patten, Roche, & Young, 2015; Pechmann, Pan, Delucchi, Lakon, & Prochaska, 2012).

¹http://www.cdc.gov/nchs/fastats/exercise.htm
Studies have even used randomized experiments to establish causal peer influence effects (Zhang, Brackbill, Yang, & Centola, 2015). Online fitness communities offer even more promising directions for work in this area (Centola & van de Rijt, 2015). These new platforms are specifically designed to provide participants with a group of peers and social support in reaching their fitness or health goals. Moreover, users are able to use wearable devices to track personal activities, including exact Global Positioning System (GPS) traces of routes, and upload them to their online profiles within the community.

For researchers, an important feature of online platforms is their ability to archive a large volume of behavioral trace data, including fitness statistics, user profile data, and potentially users' social networks, as part of their normal operation. This enables users to explore and compare their own activity efforts to others in their “fitness circles”, but it also presents novel opportunities to analyze health behaviors. Researchers now have the opportunity to observe not what people say they do, but what they actually do.²

In this study, we employ behavioral trace data from one such fitness community to study social exercise. In particular, this work analyzes how factors including gender, temporal seasonality, activity performance and social feedback may vary by number of people participating in exercise. Our work aims to answer the following research questions: (1) When do individuals choose to exercise alone and in groups, and are these dynamics gender dependent? (2) How do peoples' fitness behaviors differ, in terms of performance and social feedback, when exercising alone compared to activities with others physically co-present? In answering these questions, this study has implications for health promotion and social network-based health interventions.

2 Related Work

2.1 Social Support and Physical Activity

Previous work suggests that factors such as pleasant surroundings, an enthusiastic exercise leader, and sympathetic co-exercisers during leisure-time activities are all likely to relieve negative emotions associated with exercise (Haskell, Montoye, & Orenstein, 1985; Pelphrey et al., 2003; Flaherty, 2005). A laboratory-based study found that exercising with others helped to reduce stress and produce overall positive effects on energy, calmness and tiredness, compared with a control group exercising alone (Plante et al., 2001). However, this experiment was conducted in a laboratory setting and participants (recruited from a college student population) were fairly homogeneous in terms of age and fitness levels, making its applicability in real-world settings a open question. Despite limitations of prior studies, it is well-established that social support and physical activity are linked (Dishman et al., 1985; McAuley et al., 2000; Berkman & Glass, 2000).

More recently, the relationship between social support and physical activity has been studied using mobile fitness applications (Munson & Consolvo, 2012). Some new platforms explicitly include the element of social support, allowing users to exercise in a virtual group environment so as to motivate them to perform physical activities (Campbell, Ngo, & Fogarty, 2008; Consolvo, Everitt, Smith, & Landay, 2006; Chen & Pu, 2014). For example, (Chen & Pu, 2014) designed a mobile application with gamification settings of competition (i.e. two users compete to gain more virtual rewards by exercising), cooperation (i.e. two users contribute equally to win virtual rewards) and hybrid (i.e. weighting the cooperation and competition settings). Even though users are not required to exercise side by side in the physical environment, all three conditions of virtual group activities were found to increase users’ activity frequency and intensity.

2.2 Online Fitness Communities

The past few years have seen in explosion of new online fitness communities (e.g. RunKeeper, MapMyRun, Strava, etc.) where users’ natural, everyday activity can be tracked and explored with a rapidly expanding collection of tools and technologies (Centola, 2013). These fitness communities sit at the intersection of social media and activity tracking applications; users can not only track/log their activities, but also interact with a group of peers and posted activities. One example of such a platform is Strava. Promotional content on

²This is not to say behavior trace data from online fitness communities is not without limitations (as we discuss at the end of this paper).
the site’s home page\(^3\) says: “The social network for athletes. Connect with friends and make the most of every run and ride.” The proliferation of similar platforms, coupled with recent research indicating that just over 20% of adults use some form of technology to track their health data (Fox, 2011), signals new opportunities for understanding the social dimensions of health-oriented behaviors. In particular, behavioral traces of human behavior and interaction collected from these online sources offer novel data and strategies for understanding social dynamics and peer influence.

Online fitness communities have attracted researchers from many disciplines. Some scholars are interested in the technical potential of sensors and human-computer interaction aspects of these technologies (Consolvo et al., 2008), others have focused on play, incentives, and user engagement (Chen & Pu, 2014). A growing body of work concerns social media as a potential tool in medicine (Centola, 2013). All of these approaches promise insight into the social aspects of health, however, limited work has specifically explored the effects of peer co-presence, leaving a gap in our current understanding of the social dynamics in these settings.

3 Data

Data were collected for the online fitness community Strava. Strava continues to grow in popularity among cyclists and runners in recent years. Strava provides two main competitive, gamifying features to motivate users to reach their fitness goals. The one is the ability to compare users’ activity efforts against their history efforts or compete with other athletes. Another important feature is to accomplish challenges and earn achievement badges. As such, millions of people upload their rides and runs to Strava every week via their smartphones or GPS devices.

Strava was also chosen as the research environment because it has a number of desirable characteristics: (1) it attracts an increasing number of users around the world who upload millions of activities to the platform every week; (2) the Strava Application Programming Interface (API)\(^4\) provides access to public Strava behavioral trace data; and (3) data include activity characteristics, user characteristics and user social network characteristics, providing rich data for research that seeks to understand health behaviors.

To retrieve data about an activity posted to a user’s profile on Strava, a valid activityID is required. We randomly generate a list of activityIDs from a previously-built (and ideally exhaustive) ID space. We then query each activityID so as to check if this ID exists. If that activityID does not exist on the platform, we discard this one. If it exists, we retrieve the data associated with that activityID (i.e. information and metadata about a particular activity). As each activity is associated with a particular user, this data includes a summary representation of the posting athlete which allows us to query for a detailed representation of that user. Study procedures were reviewed by the Institutional Review Board at the authors’ university.

As stated previously that Strava API provides access to its rich metadata, our data include but not limit to the following main components: (1) activity characteristics such as activity type (e.g. cycling, running, swimming, etc.), activity location, activity names, activity-related stats (e.g. distance, moving time, elevation, etc.); (2) user characteristics such as user demographics, user location, physiological measurements (e.g. height, weight, etc.), equipment; (3) social interaction characteristics such as following and followed relationship, comments, kudos (or 'likes').

We collected 888,093 sampled activities posted during 2011-2016 from 514,362 unique users. 81.33% of users report their gender as male. 14.64% of users report their gender as female and 4.03% of users’ self-reported gender are unknown. 93.98% of data represent rides (cycling) and runs, but data also represents workout types such as swimming, walking, hiking, skiing, etc. It is important to note that this data does not represent a random sample of platform users, but is instead more likely to capture the behaviors of highly active users. While this limits the generalizability of any user-specific findings, it also means our analysis is conducted on regular users who are more likely to fully utilize platform features. This is important for this case, because many of our research questions focus on group activities - fitness events where users specify in the application that they are physically co-present with others. Moreover, we do obtain a random sample of activities, and much of the following analysis uses the activity itself as the unit of analysis.

---

\(^3\)Strava homepage is at [https://www.strava.com/](https://www.strava.com/)

\(^4\)Strava V3 API Documentation is at [https://strava.github.io/api/](https://strava.github.io/api/)
4 Methods

To explore the characteristics and dynamics of solo and group exercise, we begin by classifying observed data based on the number of peers co-present for the activity. Each activity record has an attribute that represents the number of athletes/users taking part in that activity. We identify activities involving only one athlete as solo activities and activities having more than one athletes as group activities. Our data have 696,856 (78.47%) solo activities and 191,237 (21.53%) group activities. In the analysis that follows, we evaluate the characteristics and dynamics of each type of activity, comparing solo exercise against group exercise. We do so for men and women separately, in order to tease apart any gender differences in these results. Comparing between genders is also important because the Strava platform is heavily male dominated and prior work has noted that individuals within this environment may have different experiences and social networks (Spiro & Almquist, 2016).

When exploring the temporal dynamics of exercise, we consider: (1) the day of the week (i.e. Monday to Sunday) and (2) the hour of the day (i.e. 0:00 to 23:00) during which activities are observed. Each activity has a local timestamp indicating the start time of the event. An analysis of temporal features could help understand the seasonal patterns of solo and group activities. We hypothesize that people prefer to do more solo activities on weekdays because it is possibly harder to coordinate time among multiple people for group activities during these days. Hence, we also hypothesize that group activities take place more frequently on weekends. We also want to find out the “busy” time periods for solo activities and group activities.

Next, we want to understand how performance, that is physical exertion, is related to peer co-presence. Prior work on peer influence and social support suggests that peer co-presence motivates individuals to engage in physical activity more regularly and more often, but it might also motivate them to work harder, exerting themselves more throughout their physical activity. Measuring performance is challenging and likely involves a multifaceted approach. Strava also applies a diversity of measurements for physical activities. Therefore we consider five distinct measures to operationalize activity performance. These include:

a. Distance: total distance of an activity (available on all data)
b. Elevation: total elevation gain of an activity (available on all data)
c. Duration: total moving time of an activity (available on all data)
d. Activity effort: average watts (available on cycling activities - around 60% data) and average speed (available on all data)\(^5\)
e. Physical challenge level: Strava provides its users with a computed “suffer score” which it calculates based on estimated heart rate intensity (available on premium users’ activities - around 10.5% of data)\(^6\)

Finally, our work aims to compare social feedback for activities that are complete solo versus in groups. Strava users can post comments and kudos (i.e. “likes” or +1s) on an activity. Related work suggests that these social interaction functions have motivating effects on physical activities (Chen & Pu, 2014). Hence, we are interested in analyzing whether workouts as a group receive more comments and/or kudos from activity participants or other Strava users.

5 Results

In this section, we present findings to address the research questions outlined previously. First, we discuss overall gender differences in posting behavior and exercising alone or with a group. We then continue, considering the temporal dynamics of when individuals choose to exercise alone and in groups, and how these dynamics are gender dependent. Next we evaluate how fitness behaviors differ, in terms of performance

\(^5\)Average watts measures the rate of energy conversion with respect to time. Since it is available on partial data, we use the measure of average speed that is available on activities of other types including running, hiking so as to avoid possible biased sampling merely from rides.

\(^6\)Solo activities and group activities account for 76.25% and 23.75% of premium users’ activities, respectively, roughly matching the proportions in the entire data sample. Therefore, we believe that the measure of suffer score is still representative and not likely to introduce large bias to the analysis.
5.1 Gender Differences in Activity

<table>
<thead>
<tr>
<th>Activity counts</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Female Alone</td>
<td>91,248</td>
</tr>
<tr>
<td>Female with Others</td>
<td>27,576</td>
</tr>
<tr>
<td>Male Alone</td>
<td>580,832</td>
</tr>
<tr>
<td>Male with Others</td>
<td>158,448</td>
</tr>
</tbody>
</table>

Table 1: Counts and Percentages of Solo/Group Activities by Gender

Strava is a male-dominated platform, where the large majority of platform users are males, and this is reflected in our dataset. To begin our analysis, we count the number of activities of both types for each gender, and calculate the corresponding percentages. Table 1 shows a greater percentage of activities posted by females are group activities than among males, indicating that there are proportionally more women who are involved in group activities.\(^7\)

Figure 1 shows how each gender is observed to participate in activities that involve a specified number of participants. For example, in the Figure 1(a), the largest rectangle represents the proportion of activities posted that include only one user across gender. The width (0.7857) and height (0.7679) of the rectangle correspond to the proportion for men and women, respectively. Since, in this case, the width is greater than the height, we know that solo activities enjoy a larger proportion among men than among women. However, for activities having two participants, we find a much larger proportion among women, indicating exercising with a single peer is observed more frequently among women. As the number of activity participants increases, rectangles gradually deviate from the dashed line, indicating that female users tend to engage in more activities in smaller groups (mostly in a group of size two); as participants continue to increase, we find that the trajectory of the rectangle position goes back, closer to the dashed line, indicating that men (in proportion) are involved in larger group activities.

\(^7\)A chi-squared test was performed to determine if significant difference exists in these counts. Unsurprisingly, given the dataset size, we find a highly significant \((p < 0.001)\) relationship between gender and exercising alone versus with others.
5.2 Temporal Dynamics of Activity

One of our primary research questions aims to identify seasonality patterns for solo and group exercise. Figure 2 shows the proportion of solo and group activities posted to Strava across the week. We observe that proportions of group activities are smaller on weekdays compared to that of solo activities. Moreover, about 50% group activities occur over weekends, greatly exceeding the proportion of solo activities. For solo activities, we see a relatively stable and consistent pattern throughout the week. Exercising on Mondays and Fridays appears to be less attractive to athletes, as fewer activities - both solo and group activities - are posted on these days.

We further examine the seasonality patterns of solo and group activities in terms of the hour of the day and the day of the week. Figure 3 visualizes solo/group activities occurring during a specific hour on a specific day of the week. Greater numbers of activities are represented by darker blue squares; lighter squares indicate smaller numbers of activities.

We observe that group activities, in general, are not frequent during regular work hours as we see clusters of dark blue squares positioned at the hours after work (nearly 5pm - 7pm) on weekdays or in the mornings (nearly 6am - 11am) on weekends. This pattern is consistent across gender, however males exercising with others tend to do so slightly early on weekends. We find that solo activities occur most often early in the morning or after work on weekdays, and in the mornings on weekends. Again, this pattern is consistent across gender with males exercising slightly earlier in the morning on weekdays and weekends.

5.3 Co-Present Peers and Activity Performance

Our second research question considers the relationship between the number of activity participants (i.e. peer co-presence) and activity performance. We measure performance in terms of multiple dimensions, including, distance, elevation, duration, activity effort, and physical challenge level. These measures were discussed in detail in Section 4. For each measure, we take the average of its value across all activities given a specific athlete count. Then, we examine trends as the number of activity participants increases. We find a consistent pattern across all measures: performance increases sharply as the number of participants increases up to five, when performance shows diminishing returns – increasing but at a much smaller rate.

Figure 4 shows this result for average watts, as well as moving time, as a function of athlete count. Other performance metrics show similar results, and are available in the appendix. It should be noted that we have few activities with very large group size, so we expect greater noise (e.g. greater uncertainty about estimates and larger confidence intervals) as group size goes up. In order to present readable visualizations, we truncate athlete count at 20. Our preliminary analysis shows some gender differences in baseline activity effort, such as average watts for solo activities, so visualizations show results by gender. However, the observed relationships between performance and group size are consistent across both gender groups, as seen. Moving
time is one of the few performance metrics that shows a negative relationship with group size, and only for groups larger than 5-10 athletes.

5.4 Social Feedback for Activities

Finally, we analyze how the number of athletes involved in a particular activity is associated with subsequent feedback received by the posting athlete by examining three main interaction types. Each of these types of social interaction enables Strava users to provide peers with social feedback about their posted activities. Users can make comments and give kudos (or “likes”). In addition, the original author of the activity can post associated photos.

Figure 5(a) shows what percentage of activities that were ever commented on by other platform
users for type of activity, across genders and then combined. We find that group activities are much more likely to be commented on; percentages of activities commented involving groups are twice as high as those of solo activities in across all cases. Additionally, we do not see significant gender differences here. We do not show results for kudoed activities because it has roughly the same pattern as commented activities.

Figure 5(b) shows the percentage of activities with accompanying photos posted in Strava or Instagram. We see that users are more likely to post photos to group activities. Moreover, we observe that female users tend to post more photos to both solo and group activities than male users.

Overall, group activities and solo activities differ in terms of the proportion of received social feedback. We find that group activities are much more likely to attract social feedback including comments and likes from peers. Moreover, group activities tend to motivate users’ behavior of content sharing by posting activity-related photos, which in turn is likely to gain more attention among users’ online social circle.

6 Discussion

This study compares activities posted by users in an online fitness community. In particular, we focus analysis on how the characteristics of these activities – when they occur, how intense they seem to be, and how much social feedback they receive – may be associated with the number of co-participants. To do so we make use of a unique dataset collected from the online community Strava, utilizing application features that allow users to specify who they are exercising with; data comprise not only the behavior of individual athletes (users), but also detailed records of who is physically co-present with these users. Our analysis demonstrates a number of significant findings.

First, we observe that female users tend to post activities that involve a single peer – exercise events where the number of total participants is two. Males, on the other hand, tend to post solo activities or activities that involve larger groups. These results hint at specific gender preferences in group exercise and have numerous implications for peer effects on motivation and health promotion. Importantly, for any network-based intervention or behavior change, the social networks (and as a consequence influential peers) for men and women look very different (Granovetter, 1973; Centola & Macy, 2007; Bakshy, Rosenn, Marlow, & Adamic, 2012; Lewis, Gonzalez, & Kaufman, 2012). Results suggest that females might have a single or small set of influential strong ties (i.e. exercise partners), while men may have a large, diverse set of peers who could be influential.

This study finds evidence for strong seasonal effects on group exercise. Group activities usually take place after work during the week or early in the day on weekends, whereas many solo activities also take place in the early mornings on weekdays. Building from the previous discussion, while strong diurnal patterns are unsurprising in human behavior (Golder & Macy, 2011), results demonstrate that opportunities for peer influence on health behaviors are likely to be restricted or constrained in systematic ways. For example,
designers of application features might suggest that exercise partners should take into account optimal times for group exercise and individuals preferences for when to work out. While outside the scope of this study, there are many interesting directions for future work that considers mechanisms to affect behavior change.

Group activities differ from solo activities, in terms of effort, exertion and performance, our analysis indicates. When exercising with others, even just a single peer, athletes see notable gains in workout intensity and energy expended – increases in average power output (measured in average watts for cycling events), moving time and distance. Interestingly, these gains continue to increase for every additional activity participant (though primary gains are seen for the first 5 additional co-present peers). Activities with co-present peers might be informal (organized by the participants themselves) or formal (group rides perhaps organized by a local club or other organizational entity). Further work might tease apart these different conditions to offer insight into peer effects and impact of institutional structure on exercise (Vilhjalmsson & Kristjansdottir, 2003).

Co-presence and social interaction are distinct but related concepts. In the final component of the analysis presented here, we consider observations of social exchange among athletes. In particular, we consider social feedback behaviors - platform users commenting and liking each others’ activities. Findings demonstrate that group activities are associated with higher levels of social feedback than solo activities. Moreover, group activities are significantly more likely to include multimedia (photos). In the latter case, data also reveals gender differences, where females are more likely to post group activities with photos than males. Increased engagement and social feedback may also be related to motivation and future activity, suggesting more promising directions for further work.

7 Limitations

While the study presented here offers novel insight into the characteristics of physical activities where peers are co-present, it is not without limitations. One notable concern is the ability of Strava application users to restrict their activities to be private, shared only with pre-screened peers. Private accounts, and likewise private activities, cannot be accessed from the Strava API, and therefore are excluded from the data used in this study. Athletes who choose to restrict access to their data may systematically differ from users who make their data public. Unfortunately, we are unable to assess the impact of this bias because of lack of data. Instead, one should be careful about generalizing these results beyond the population of study.

8 Conclusion

As social fitness mobile applications become widely used for personal activity tracking, social support and health promotion, opportunities for understanding the effects of social networks and peer influence on behavior change and health expand. Drawing on features of social media and activity tracking applications, many of these new platforms capture rich data about physical activities as well as peer-to-peer interactions. The behavioral trace data they archive have the potential to significantly alter understanding of health and well-being. In this study, we explore the characteristics and dynamics of social exercise, that is fitness activities with at least one peer physically co-present. Our research focuses on quantifying diurnal patterns, activity performance and social feedback as they vary by the number of people participating in an activity; we also compare associations by gender. Our results indicate that when peers are physically co-present exercise tends to be more intense and receive more feedback from other users. Findings also suggest gender differences in the observed tendency to engage in physical activity with others. The implications of these results for network-based health and wellness interventions are also discussed.

References


9 Appendix: Peer Co-presence versus Activity Performance

This appendix contains results demonstrating how all measurements of activity performance change as the number of participants increases.
Figure 6: Activity performance metrics as a function of number of participants. Colored bands around mean line represent bootstrapped 95% confidence intervals.
Welcome New Americans!
Investigating the role of hyper-local online communities in integration of immigrants

Rosta Farzan\textsuperscript{1}, Grace Ahmed \textsuperscript{2}, Yvette Espinoza\textsuperscript{3}, Natalie Euley\textsuperscript{4}, Marissa Villarreal\textsuperscript{5}
\textsuperscript{1}University of Pittsburgh
\textsuperscript{2}University of Vermont
\textsuperscript{3}California State University, Fresno
\textsuperscript{4}University of Alabama in Huntsville
\textsuperscript{5}University of Houston

Abstract
The United States has a continuously growing immigrant population. A problem many of these New Americans may face is adapting to the new culture. Researchers have been investigating ways technology can play a role in supporting acculturation of the immigrant population. In this work, we studied the role of a particular class of technological support, hyper-local online communities, designed to support individuals living in the same geographical boundaries. Through a survey of a 50 immigrants from two distinct areas in the US, we investigated whether utilization of hyper-local online communities can be associated with better integration with the local community demonstrated as increase sense of belonging and satisfaction. We also investigated potential factors contributing to utilization of such platforms. We report the results of our study; highlight potential implications for design of technology for immigrants and discuss future direction of research in this area.

Keywords: immigrants; hyper-local online communities; social networking sites; adaptation; sense of belonging


Copyright: Copyright is held by the authors.
Contact: rfarzan@pitt.edu

1 Introduction

The United States has a continuously growing immigrant population (Gibson & Yung, 2006; Grieco et al., 2012). A challenge faced by many immigrants as well as the host country is integration of the immigrant population into the new community (Scott & Scott, 2013; Su’arez-Orozco, Suárez-Orozco, & Qin-Hilliard, 2014). There have been attempts in studying the role of technology in tackling this challenge (Tsai, 2006). For example, immigrants have a substantial information need in adapting to the new society and it has been shown that technology can benefit immigrants by supporting their information and practical needs in adjusting to living in a new country (Shoham & Strauss, 2008; Fisher, Durrance, & Hinton, 2004).

With growing popularity of online communities and online social networking sites, there has been a growing interest in understanding whether such online platforms can play a role in strengthening local communities and supporting individuals living in the same geographical areas. As a result, over the last decade, there has been an emergence of a new category of online communities, known as hyper-local online communities, that are focused on supporting the local communities and which rely heavily on offline connections of their members. Nextdoor\textsuperscript{1}, a for-profit example of such platforms, has been very successful in attracting attention from residents and local governments (Masden, Grevet, Grinter, Gilbert, & Edwards, 2014). Nextdoor is a social networking site dedicated to neighborhoods and is designed to encourage its members to connect with people who live in the same or close-by neighborhood. As of July 2016, more than 50\% of neighborhoods in the US are represented on Nextdoor. There are various other examples of such systems supporting particular cities or neighborhoods. For example, E-Democracy, has been serving local

\textsuperscript{1}https://nextdoor.com/
communities in the city of Minneapolis and St. Paul in the US state of Minnesota to encourage community development (López & Farzan, 2015)

While there has been studies of the general impact of hyper-local online communities in community development, there has not been any special attention to immigrants as potential benefactors of such online systems. In this work, we have investigated the role of hyper-local online communities in the integration of immigrants within their local communities. We argue that hyper-local online communities can provide immigrants with a platform to learn more about their local communities, connect with their neighbors, and, as a result, develop a sense of belonging to the local community which can lead into the broader integration in the new country.

2 Hyperlocal online communities, sense of community, and immigrants

Hyper-local online communities are online platforms that are designed to support people in a bounded geographic region, such as a neighborhood or a town. They aim at enriching residents’ and visitors’ experiences with their local surrounding through a variety of services such as providing hyper-local maps to offer optimal walking, hiking or biking paths, creating a marketplace for residents to exchange goods and services, aggregating and sharing hyper-local news, encouraging local activism and civic participation, and social networking among residents.

Over the last decade, there has been a great deal of attention devoted to studying the impact of such systems on local communities. In particular, researchers have been considering the impact of hyper-local online communities with respect to social capital in the community, sense of community, individual empowerment, strong democracy, and economic development (Lopez & Farzan, 2015). For example, in a study of Netville, an online network to connect local residents, a group of residents were randomly assigned to receive access to high-speed Internet and a neighborhood mailing list. The results of the study suggest that online connections improved neighbors’ familiarity with each other and improved their communication (Ellison, Steinfield, & Lampe, 2007). Being “wired” especially supported increased contact with “weak-ties” in the neighborhood (Hampton & Wellman, 2003), and connections to weak ties are particularly important for accessing information and strengthening social cohesion and a sense of community within existing groups (Granovetter, 1973).

McMillan & Chavis defined a sense of community as the “feeling that members have of belonging, a feeling that members matter to one another and to the group, and shares faith that members’ needs will be met through their commitment together.” (Navarrete & Huerta, 2006) Individuals who experience a sense of community are more satisfied with and more committed to their communities. (Navarrete & Huerta, 2006).

While there have been a number of studies on the impact of hyper-local online communities on the local communities, most of those studies have focused on general population of the neighborhoods and cities. There has not been any work investigating at what level hyper-local online communities can help immigrants integrate in the host community. Computer technology can assist immigrant families in overcoming barriers posed by loss of social networks, social disconnection, and limited language proficiency (Tsai, 2006). There is some concern that immigrants’ use of internet technologies can increase their isolation in their new community. Computer-mediated forms of interpersonal communication have become the technologies of choice for immigrants to maintain ties to friends and family in their country of origin, and those who are more active in social engagement with co-ethnics tend to be less active with respect to interactions with native-born Americans. At the same time, those immigrants who are more engaged in interpersonal communication with native-born Americans have higher levels of psychological well-being surrounding American society and culture (Kim & McKay-Semmler, 2013). Immigrants who communicate more via Internet with local people have higher levels of sociocultural adaptation, including social, psychological and physical adaptation (Chen, 2010).

While knowledge and use of computer systems can be associated with higher levels of adaptation, specific sites such as hyper-local online communities have yet to be studied independently to see if their use leads to similar levels of adaptation. Dual forms of community, online and physical, in virtual transnational communities of immigrants facilitate the creation of identity and the production of trust processes (Navarrete & Huerta, 2006). Similarities between virtual transnational communities of immigrants and hyper-local networks in their physical and virtual forms may suggest that the online-offline nature of community networks can also contribute to an increased sense of community among their users.
3 Research Hypotheses

This study aims to answer how immigrants benefit from the use of hyper-local online communities and what barriers might prohibit their use of such platforms.

Supported by prior research on the impact of hyper-local online communities on strengthening individuals’ weak ties, we expect immigrants to benefit from such communities by (1) building new connections to local residents, (2) having more access to locally-oriented information, and (3) being able to share information with the local residents. While there has not been specific research on impact of hyper-local online communities on sociocultural adaptation of immigrants, some existing research suggest there can be positive effect. For example, a study of Chinese immigrant to Singapore shows those who used more Singaporean websites had better sociocultural adaptation (Chen, 2010). Therefore, our first and second hypothesis are targeted at investigating the relationship between utilizing hyper-local online communities and sociocultural adaptation of immigrants in terms of satisfaction with their local community and difficulties they face within the community. We posits our hypotheses as:

H1: Immigrants who use hyper-local online communities will exhibit higher level of satisfaction and sense of community with their local community.

H2: Immigrants who use hyper-local online communities exhibit lower levels of difficulties within their local community.

In addition to trying to investigate the relationship between utilizing hyper-local online communities and sociocultural adaptation, we were interested to study how immigrants generally feel about hyper-local online communities and how their perception of these sites influences the extent they make use of them. Therefore, our third and fourth hypotheses focus on assessing the relationship between participants’ perception of hyper-local online communities and whether they use them or not as stated below:

H3: Immigrants who perceive hyper-local online communities to be important are more likely to utilize them.

H4: Immigrants who perceive hyper-local online communities to contribute to their community life are more likely to utilize them.

We anticipate that several demographic characteristics of participants can influence both the utilization of hyper-local online communities and their impact. For example, longer residency in the U.S. may indicate more stable social relationships, work history, and English proficiency that in return can lead into or be correlated with more satisfaction, stronger sense of community, and higher level of sociocultural adaptation. Research has consistently shown that limited English proficiency is a barrier to adoption of any type of computer or Internet use, even when controlling for economic status. (Mirchandani, Ng, Sangha, Rawlings, & Coloma-Moya, 2005) In addition to years of residency, and English proficiency, we included other demographic related questions such as age, race, education, and income.

4 Methods

Fifty-four English speaking participants living in Houston, Texas and greater Burlington, Vermont, between the ages of 18 and 64, who were born outside the United States were recruited for this study. Participants completed an online survey containing questions regarding demographic information, sociocultural and psychological adaptation, and their use of hyper-local online communities. Flyers were posted at community organizations that serve the immigrant population, such as non-profit centers assisting refugees and public libraries. We encouraged participants to share the survey link with other first generation immigrants. Laptops were also made available at community events so participants could access the survey regardless of Internet access at home.

4.1 Instruments

In addition to demographic questions, the survey instrument included questions on sociocultural and psychological adaptation to measure participants’ integration within their community, their use of hyper-local online communities, and their perception of the impact and contribution of hyper-local online communities.
4.1.1 Sociocultural adaptation

Chen’s scales of sociocultural adaptation (Chen, 2010), based on Ward and Kennedy’s SCAS were used to measure social and cultural adaptation of the participants. We adapted the scale further to focus on eight of the twelve items found by Chen to significantly contribute to sociocultural adaptation. Social adaptation includes: making friends, understanding jokes and humor, making yourself understood, and communicating with people from different ethnic groups. Cultural adaptation includes: understanding the U.S. political system, understanding cultural differences between the United States and your country of origin, understanding Americans’ morals or value systems, and seeing things from Americans’ point of view. Each of the eight items was measured on a 5-point Likert Scale where 1 was no difficulty and 5 was extreme difficulty.

4.1.2 Psychological adaptation

A modification of Gao and Gudykunst’s (Gao & Gudykunst, 1990) adaptation scale was used to measure participants’ psychological adaptation. The items included: how comfortable do you feel living in America, how satisfied are you with your work or study performance in America, how comfortable are you interacting with Americans, how satisfied are you with your English language ability, how satisfied are you with living in American culture, and how much is life for you in America an enjoyable experience. As in Chen’s modification of Gao and Gudykunst’s psychological adaptation scale, we simplified the 7-point Likert scale to a 5-point scale where 1 was "not at all," and 5 was "very much."

4.1.3 Use of hyper-local online communities

Participants were asked if they had used at least one of the following hyper-local online communities: NextDoor, Every Block, Front Porch Forum, another online group for their neighborhood, or another online group for their city. NextDoor and Every Block are available in Houston, Texas. Front Porch Forum is available in Burlington, Vermont.

4.1.4 Perception of importance of hyper-local online communities

A set of questions asked the participants to rate the importance and contribution of hyper-local online communities with respect to: getting community news and local event announcements, learning about local businesses, resources and services, discussing or understanding others’ views on community issues and happenings, sharing information or ideas, getting involved in local initiatives or causes, meeting neighbors and other community members, and helping neighbors in need. The responses were recorded on a 5-point Likert scale (1-Not at all important, 5- Extremely important).

4.1.5 Perception of contribution of hyper-local online communities

Finally, the survey asked the extent the information or discussions happening on the online tools in the last twelve months had contributed to: I feel that my participation is welcomed; I feel my participation is valued by others; I have been introduced to new ideas or points of view; I have learned more about my neighbors of races, ethnicities, or home languages different from my own; I am more informed about issues that affect my community; I am more committed to my community; I have learned more about how to influence decisions in my community; I am more satisfied with my local community as a place to live or work; and I am more confident I could help from people in my community. The questions were recorded on a 5-point Likert scale (1- not at all, 5- very much).

5 Data consideration

Demographics information in our survey included age, gender, education, race, income, residency in the US, English proficiency, also zipcode of the participants that indicated whether they lived in Houston, Texas or Burlington, Vermont. The majority of the participants were under the age of 35 with most being between 18 to 24. Therefore, we discretized the age variable into the binary variable of below or above 35. Similarly, due to skewness of data, we discretized race into three categories of white, Hispanic, and others; education into
two categories of college degree or below; income into two categories of above or below 50K, US-residency into two categories of above or below 10 years; and English proficiency into two categories of excellent or not. Percentage of data within each category is represent in Table ??.

<table>
<thead>
<tr>
<th>Location:</th>
<th>Houston, Texas</th>
<th>56%</th>
<th>Burlington, Vermont</th>
<th>44%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age:</td>
<td>Under 35</td>
<td>55%</td>
<td>Over 35</td>
<td>45%</td>
</tr>
<tr>
<td>Race:</td>
<td>White</td>
<td>16%</td>
<td>Hispanic</td>
<td>50%</td>
</tr>
<tr>
<td>Degree:</td>
<td>College</td>
<td>44%</td>
<td>Below college</td>
<td>56%</td>
</tr>
<tr>
<td>Income:</td>
<td>Above $50k</td>
<td>45%</td>
<td>Income below $50k</td>
<td>55%</td>
</tr>
<tr>
<td>Residency:</td>
<td>Above 10 years</td>
<td>66%</td>
<td>Below 10 years</td>
<td>34%</td>
</tr>
<tr>
<td>English proficiency</td>
<td>Excellent</td>
<td>48%</td>
<td>Not excellent</td>
<td>52%</td>
</tr>
</tbody>
</table>

Table 1: Demographic Characteristics of Respondents

To ensure there is no correlation between independent variables and that there is no collinearity problem, we first conducted a factor analysis using Principal Component Analysis with Varimax rotation. The demographic features loaded into 3 components. The result of the factor analysis is shown in Table ???. For each component, we included the independent variables representing the highest degree of variation for that component. Component 1 consists of zipcode, gender, and education which indicate correlation between gender and education and location of residence in the US; component 2 consists of race and residency which indicates correlation between particular races having lived in these two cities longer than others, and component 3 consists of age and using hyper-local online networks which indicated certain age groups are more likely to use hyper-local online communities.

<table>
<thead>
<tr>
<th>Component 1</th>
<th>Component 2</th>
<th>Component 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Zipcode</td>
<td>.659</td>
<td>-.297</td>
</tr>
<tr>
<td>Gender</td>
<td>-.660</td>
<td>.137</td>
</tr>
<tr>
<td>Age</td>
<td>-.018</td>
<td>.022</td>
</tr>
<tr>
<td>Race</td>
<td>-.113</td>
<td>-.806</td>
</tr>
<tr>
<td>Education</td>
<td>.735</td>
<td>.191</td>
</tr>
<tr>
<td>Income</td>
<td>.565</td>
<td>.334</td>
</tr>
<tr>
<td>Residency in US</td>
<td>-.220</td>
<td>.827</td>
</tr>
<tr>
<td>English</td>
<td>.430</td>
<td>.516</td>
</tr>
<tr>
<td>Use of hyper-local online networks</td>
<td>.216</td>
<td>.144</td>
</tr>
</tbody>
</table>

Table 2: Principle Component Analysis for all demographics factors in the survey

6 Results

To test our hypotheses, we conducted a series of regression analyses as reported below.

6.1 H1: Relationship between using hyperlocal online communities and the feeling of comfort and satisfaction in the local community

All scales of psychological adaptation loaded into one factor, with the sum ranging from 13 to 30 with a mean of 25.35 and a total of 55 responses. A higher score indicates better psychological adaptation and more satisfaction with life in their community. Similar to the previous model, the dependent variable fit Gamma distribution the best. The model is significant ($\chi^2=39.31$, df=8, sig. <.0001). The results of the regression for the significant factors is presented in Table ???. The results shows that controlling for demographic
information, usage of hyper-local online communities does not influence participants’ satisfaction with their community. However, in terms of demographic information, those who have lived in the US longer than 10 years had higher psychological adaptation scores, i.e. they felt more comfortable and satisfied in the local community: $\mu = 25.61$ ($\sigma^2 = .71$) vs. $\mu = 22.36$ ($\sigma^2 = .91$). Male participants tend to feel more comfortable and satisfied in the local community: $\mu = 25.40$ ($\sigma^2 = .9$) vs. $\mu = 22.55$ ($\sigma^2 = .69$). Respondents in Houston, Texas reported higher levels of satisfaction than respondents in Vermont: $\mu = 25.83$ ($\sigma^2 = 1.11$) vs. $\mu = 22.17$ ($\sigma^2 = .7$).

### Table 3: Regression results predicting relationship between usage of hyper-local online communities and satisfaction with the community

<table>
<thead>
<tr>
<th>Parameter</th>
<th>B</th>
<th>Std. Error</th>
<th>Lower</th>
<th>Upper</th>
<th>Wald $\chi^2$</th>
<th>df</th>
<th>Sig.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lived over 10 years in US</td>
<td>-.136</td>
<td>.0461</td>
<td>-.226</td>
<td>-.045</td>
<td>8.653</td>
<td>1</td>
<td>.003</td>
</tr>
<tr>
<td>Male</td>
<td>.119</td>
<td>.0414</td>
<td>.038</td>
<td>.200</td>
<td>8.247</td>
<td>1</td>
<td>.004</td>
</tr>
<tr>
<td>Houston</td>
<td>.152</td>
<td>.0548</td>
<td>.045</td>
<td>.260</td>
<td>7.736</td>
<td>1</td>
<td>.005</td>
</tr>
</tbody>
</table>

Table 3: Regression results predicting relationship between usage of hyper-local online communities and satisfaction with the community

### 6.2 H2: Relationship between using hyperlocal online communities and the feeling of sense of difficulty in the local community

All questions on sociocultural adaptation scale focusing on feeling of sense of difficulty within the community loaded into one factor. We represented this measure as sum of all values for related questions. The response variable ranges from 8 to 38 with a mean of 18.06 and a total of 50 responses. A higher score indicates more difficulty adapting to the American life and is thus more negative while a lower score indicates less difficulty and is thus more positive. We conducted a generalized linear regression model and modeled the dependent variable of difficulty with Gamma distribution based on the fit of the model. The model is significant ($\chi^2 = 15.23$, df=8, Sig.=.05). The results of the regression for the significant factors is presented in Table 4. The results indicate that controlling for all other demographics factors, those who use hyperlocal online networks have better sociocultural adaptation, which is to say, less difficulty in communication and understanding the American culture: $\mu = 14.07$ ($\sigma^2 = 1.5$) vs. $\mu = 18.91$ ($\sigma^2 = 1.22$).

In terms of demographic differences, men tend to feel less difficulty in the local community: $\mu = 14.93$ ($\sigma^2 = 1.29$) vs. $\mu = 17.82$ ($\sigma^2 = 1.38$) and those who were under the age of thirty-five felt less difficulty in the local community: $\mu = 14.57$ ($\sigma^2 = 1.17$) vs. $\mu = 18.25$ ($\sigma^2 = 1.55$).

### Table 4: Regression results to assess the relationship between usage of hyper-local online communities and sense of difficulty in the community

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Coef.</th>
<th>Std. Error</th>
<th>Lower</th>
<th>Upper</th>
<th>Wald $\chi^2$</th>
<th>df</th>
<th>Sig.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Use of hyper-local online networks</td>
<td>.296</td>
<td>.1211</td>
<td>.059</td>
<td>.533</td>
<td>5.974</td>
<td>1</td>
<td>.015</td>
</tr>
<tr>
<td>Male</td>
<td>-.177</td>
<td>.1037</td>
<td>-.381</td>
<td>.026</td>
<td>2.928</td>
<td>1</td>
<td>.087</td>
</tr>
<tr>
<td>Below 35 years</td>
<td>-.225</td>
<td>.1053</td>
<td>-.431</td>
<td>-.019</td>
<td>4.566</td>
<td>1</td>
<td>.033</td>
</tr>
</tbody>
</table>

Table 4: Regression results to assess the relationship between usage of hyper-local online communities and sense of difficulty in the community

### 6.3 H3 & H4: Relationship between perception and utilization of hyper-local online communities

All scales of how important the ability to perform a variety of actions on a local community site loaded into one factor, with the sum ranging from 11 to 35 with a mean of 24.66 and a total of 50 responses. A higher perception of importance of such platforms. Similarly, all scales regarding the perception of contribution of hyper-local online communities loaded into one factor, with the sum ranging from 9 to 45 with a mean of
28.49 and a total of 51 responses. A higher score indicates higher perception of the impact of such platforms. In both cases, the regression model prediction the relationship was not significant. However, as presented in Figure 1a and 1b, there is a positive pattern suggesting that more positive perception can lead into higher chance of utilizing hyper-local online communities.

![Error bars: +/- 1 SE](image1.png)

(a) Importance

![Error bars: +/- 1 SE](image2.png)

(b) Impact and contribution

Figure 1: Perception of hyper-local online communities

7 Discussion

In this paper, we studied what factors relate to utilization of hyper-local online communities among immigrants and how the usage of such systems relates to immigrants’ integration within their communities. Our results provide preliminary evidence that usage of hyper-local online communities can be related to less sense of difficulty within the local community. We also observed that there are specific demographics that significantly correlate with sociocultural adaptation, such gender, years of residency, and age. We also observed a significant difference among the two locations we studied in terms of sociocultural adaptation. These results suggest that hyper-local online communities can be promoted as an approach to support immigrants population, especially the group of immigrants who have a more difficult time integrating within the new community. We acknowledge that this a very preliminary work and further research is necessary to explore the impact of hyper-local online communities in details across different communities, locations, and with relation to different features of those platforms. Furthermore, further research is required to understand the mechanisms by which hyper-local online communities assist immigrants and how they can be designed more effectively. Nevertheless, we hope to raise awareness among the information scientists on the impact of a specific information technology for increase a sense of belonging among immigrant population that are such an important part of the American society.

References


Personalized Community Detection in Scholarly Network

Zheng Gao, Xiaozhong Liu
1School of Informatics and Computing, Indiana University Bloomington

Abstract
Most graph clustering methods partition a network into communities based solely on the topology and structure of the network. Due to this, the means via which communities are detected on a network are insensitive to the preferences of a user who is searching the network with a specific, personalized information need. Such partition algorithms may be of diminished value for scholars exploring networks of research if these scholars possess prior preferences on what information they consider relevant. To better address this type of information seeking behavior, we introduce a personalized community detection algorithm that provides higher-resolution partitioning of areas of the network that are more relevant to a provided seed query. This algorithm utilizes the divisive Girvan-Newman approach but incorporates a user’s personal preferences as a prior. We show that this personalized algorithm can produce a more fine-tuned partition of a scholarly network when compared to existing prior-insensitive approaches.

Keywords: personalized community detection; graph mining; networkanalysis


Copyright: Copyright is held by the authors.

1 Introduction
A common information-seeking behavior in the academic domain is that of scholars searching for research relevant to their areas of expertise. An ideal user-centered community detection method would provide higher-resolution partitions in areas of the graph most relevant to the user, while the remaining areas of the graph are partitioned in a coarser manner that brings about no loss when considered with respect to the user’s information need. The result of such a method would be a graph where fine-grained detail is allocated to areas of interest while the rest of the graph is partitioned at a higher level of abstraction. Although traditional algorithms lead to reasonable network partitions, in their default implementations they operate globally and are insensitive to a user preference for a specific subregion of the graph. Therefore they cannot produce this result.

For instance, suppose some experts in information retrieval want to conduct an open-ended search in a network of scholarly research, one not guided by a search keyword but is limited in scope to their domain area. Since they are very familiar with the subdomains and specialties of their field, a graph representation that makes these distinctions is useful. IR experts may be interested in the relationship between their field and other disciplines; medical IR experts may be interested in biology, for example, or image IR experts may be interested in computer vision. Partitioning these related disciplines at the same level of granularity as the home discipline may be distracting at best, and confusing at worst.

To make scholarly network search sensitive to a user’s expertise and preferences, we propose an implementation of the Girvan-Newman divisive method to personalize clustering results. Given a user’s indication of their search preferences, we employ an edge-based PageRank method to calculate relevance scores on all relationships in the network. Then we combine edge relevance scores and edge betweenness scores together to create an integrated edge-importance metric that is subsequently used to cluster network nodes. Finally we calculate and choose the clustering result with highest modularity.

Section 2 is brief literature review for classic clustering algorithms. Section 3 describes our personalized community detection algorithm via modification of the Girvan-Newman method for community detection. In section 4, using information from a ACM dataset, we generate a homogenous paper network to test our algorithm and compare our result with original divisive method. Our conclusion and future research directions are presented in section 5.


2 Literature Review

The problem of finding community structure in graphs has long been a central research topic in network science. Existing community detection methods can be divided into three categories: partitioning algorithms, spectral algorithms, and dynamic algorithms (Fortunato, 2010; Leicht & Newman, 2008). Divisive partitioning algorithms start with the entire network and remove edges in some metric-determined order to find community structure, while agglomerative partitioning algorithms find structure by starting with all nodes as isolates and adding edges. Spectral algorithms make use of the spectrum (e.g. eigenvalues) of a similarity matrix derived from the original data to perform dimensionality reduction, thus allowing clustering to be performed in a lower-dimensional space. Dynamic algorithms detect communities based on the probability of the path of a random walk on the network. In 2002, Girvan and Newman first raised edge-betweenness-based divisive method which refers to the number of shortest paths passing through an edge in a graph (Newman & Girvan, 2004). Edge betweenness refers to the number of shortest paths passing through an edge in a graph. It serves as a representation of the edge’s importance in the network. The method iteratively removes the edge with highest edge betweenness score to naturally divide the graph into several subgraphs.

3 Methodology

In personalized community detection, our overall goal is to generate clusters with different resolution, dependent on the preferences indicated by the user. This means that clusters containing nodes relevant to the indicated preferences will be divided at a higher level of granularity. In contrast, clusters containing nodes less relevant to the indicated preferences will be larger and more broad in subject scope. If the same personalized algorithm is run with different prior preferences, the resulting partition will be different.

The Girvan-Newman method is the most widely-used divisive method for community detection. The core procedure in Girvan-Newman method is edge betweenness calculation. The edges with the highest edge betweenness scores are iteratively removed, generating isolated subgraphs that are subsequently understood as local communities. In order to modify this method for variable levels of granularity, it is necessary to guarantee that edges connecting nodes more relevant to the users’ indicated preferences are more likely to be removed in the earlier iterations of the Girvan-Newman method.

Our method for incorporating user preferences consists of six major steps. The pseudocode is shown in algorithm 1:

It is necessary to have a clear and operational definition of “prior preference” for this particular task. While the Girvan-Newman method operates on edges, our method receives the user’s prior preferences for nodes. For step 1, the user is given the ability to select one “seed” node to represent their preferences in the graph. The selected node is then given a prior score of 1, while all other nodes in the network are given a score of 0. To cast a wider net, we can modify this approach by applying a prior score of 1 to the selected seed node and all of its neighbors connected by an edge, that is, all nodes a distance of 1 from the seed. Overall, the prior preference is represented through a mapping function from the user to the nodes in the network.

Step 2 is the calculation of edge-based PageRank which represents edge relevanceto user preference (Csáji, Jungers, & Blondel, 2014). After translating the user’s indicated preferences into prior scores $P$ onto nodes, we use a transaction function to transfer the value from nodes to edges. We define the edge prior score $(P_{\text{start node}} + P_{\text{end node}})/2$.

However, because only a few nodes in the network can have prior score 1, most edges will be assigned a score of 0. To smooth the edge scores, we apply PageRank on the edges:

$$SC = \frac{SC + \alpha \cdot \sum_{e \in E} SC_e/T}{1 + \alpha}$$

Where $E$ refers to all incoming edges to the start node of edge and $T$ refers to out-degree of the start node. After multiple iterations of the PageRank algorithm, the edge relevance scores converge. Edges attached to more relevant nodes will have higher edge scores (Bae, Park, Ahn, & Park, 2016).

Step 3 involves calculating each edge’s overall importance score using both edge relevance and edge
Algorithm 1 My algorithm

1: procedure MyProcedure(Graph<V,E>)
2:   $P_{node} \leftarrow $ prior node score in Graph
3:   $SC_{edge} \leftarrow (P_{startnode} + P_{endnode})/2$
4:   while iteration > 0 do
5:     for all edge in Graph do
6:       node $\leftarrow$ startnode; $T \leftarrow$ out degree of node;
7:       $E \leftarrow$ incoming edges collection
8:       $SC_{edge} \leftarrow (SC_{edge} + \alpha \cdot \sum_{e \in E} SC_{e})/(1 + \alpha)$
9:     iteration $\leftarrow$ iteration $-$ 1
10:   end for
11: end while
12: for all edge in Graph do
13:   $SP_{edge} \leftarrow$ edge betweenness score for each edge
14:   $f(x) \leftarrow$ sigmoid($x$)
15:   $S_{edge} \leftarrow (f(SC_{edge}) + \beta \cdot f(SP_{edge}))/ (1 + \beta)$
16: end for
17: while $N(E) > 0$ do
18:   $Graph^t \leftarrow$ Remove max($S_{edge}$) from Graph
19:   if $Modularity(Graph^t) > Q$ then
20:     $Q \leftarrow$ Modularity(Graph$^t$)
21:   end if
22:   $N(E) \leftarrow N(E) - 1$
23: end while
24: end procedure

betweenness. We use sigmoid function to normalize both into (0,1) range and calculate edge importance as:

$$S = \frac{\text{sigmoid}(SC) + \beta \cdot \text{sigmoid}(SP)}{1 + \beta}$$

Step 4 involves removing edges iteratively in the manner of the Girvan-Newman method. After calculating edge importance score, we rank all edges based on edge importance score and at each iteration the highest ranking edge is removed from the graph.

Step 5 is finding the best partition through modularity calculation. After removing the most important edges, we calculate modularity score based on current partition result. Each isolated subgraph will be regarded as a cluster. We use the modularity to evaluate current partition. The formula of modularity for unweighted and undirected graph is

$$Q = \frac{1}{2m} \sum_{i,j} (A_{ij} - \frac{k_i k_j}{2m}) \cdot \delta(C_i, C_j)$$

Where $m$ is the total number of edges in the graph, and $A_{ij}$ indicates the presence of an undirected edge between node $i$ and node $j$. $k_{ij}$ represents the degree of the node. The expected number of edges between node $i$ and node $j$ is $\frac{k_i k_j}{2m}$. $\delta$ is the Kronecker delta, and has a value of 1 when $C_i = C_j$ and 0 otherwise. Modularity has a range from (-1,1). Traditionally, a higher modularity value is taken to represent a better partition of the graph.

Step 6 is iterating steps 4 and 5 until the partition with the highest modularity score is found.

The final outcome is a partition of the graph that is at a higher level of granularity in the areas given a higher prior preference score by the user. Figure 1 shows the difference between our method and the original method by Girvan and Newman. The gray node indicates the node selected by the user as a paper of interest. Using the Girvan-Newman method, the granularity of the detected communities is even throughout the whole graph, as seen on the left side of Figure 1. It is insensitive to the user’s preference. On the right, we see that the community selected by the user is divided into finer clusters, while those not pertinent are grouped on a coarser level. The user’s input causes one particular part of the graph to have a higher partition resolution.
4 Experiment

4.1 Dataset
The dataset we use is from ACM digital library. In its default format, the network formed by the ACM data is heterogeneous, with nodes representing authors, publications, and venues. Our method requires a homogenous graph, and so we extract an papers-only subgraph from the original data. Table 1 shows the description of the papers graph. We apply our method on only the largest connected component in the graph.

In the end, we decide to use the graph in our approach to test whether our approach can generate better result than original divisive method.

4.2 Experiment Setting
We set alpha in the PageRank equation to a value of 0.5 based on empirical testing result. In order to identify an optimal value for $\beta$, we first select the 10 most cited papers of the 15 most influential scholars in the information retrieval domain. These 10 papers serve as test seeds whose prior value are set to 1. We then follow the steps above to obtain edge prior scores and use edge-based PageRank to obtain edge relevance scores. The PageRank algorithm is run for 100 iterations. Edge-betweenness scores are calculated using a breadth-first search method. This allows the calculation of the edge importance scores. We calculate the highest modularity using values from the set (0.1, 0.5, 1, 5, 10, 20) for the hyperparameter $\beta$ and record the resulting partition. We then compare the results across different values of $\beta$, and with the partitions produced by the original Girvan-Newman method.

4.3 Result
To create a ground truth to evaluate the efficacy of our method, we manually allocated a set of top 15 conferences on the subject of information retrieval into one of the following 5 categories: human-computer interaction, core information retrieval, machine learning, evaluation, and multimedia. In the ACM papers graph, if a paper was published in one of these 15 conferences, it is assigned that conferences corresponding

---

1Microsoft Academic in https://academic.microsoft.com/
Figure 2: A toy sample for true positive & false negative case calculation

Figure 3: Precision distribution on $\beta$

category label. We use this ground truth dataset to determine if the personalized community detection method we have presented performs better in terms of precision, recall, Rand index, and F-value.

This evaluation ground-truth dataset is an assignment of papers to categories. Our method partitions the graph of papers into clusters or groupings that serve as proxies for these categories. Given the ground truth dataset, we consider each pair of nodes assigned to the same cluster as a positive case, and each pair of nodes assigned to different clusters to be a negative case. Each positive pair of nodes that were also in the same cluster by the partitioning method are true positives, while each positive pair of nodes that are in different clusters by the partitioning method are false negatives. A pair of nodes placed in the same cluster by the partitioning method but are not in the same category in the ground-truth dataset are false positives, while a pair of nodes placed in different clusters by the partitioning method and are also in the different clusters in the ground-truth dataset are true negatives. We illustrate our evaluation method using the example in Figure 2. The left side represents the ground truth, while the right side is the hypothetical partition found by a clustering method. In this example, the method has correctly partition all the nodes except one, where a node belonging to cluster 1 was erroneously assigned with the nodes of cluster 2. In the partitioning results, the first (uppermost) partition has two nodes in it. There are 1 pair in this first partition, and it is a correct pairing. There are 3 nodes in the second partition, and therefore there are 3 pairs associated with it. Of those 3 pairs, 1 is correct, but 2 are incorrect. Therefore, in this example case, the partition has 2 true positives, 2 false positives, 4 true negatives and 2 false negatives.

Our ground-truth dataset contains 10,548 nodes that represent a paper. For each candidate value of $\beta$, we calculated the average precision across the 10 seed papers (Figure 3). The precision peaks at $\beta = 1$. Based on the PageRank formula described above, this value of $\beta$ indicates that edge relevance is helpful, but if weighted too much will harm performance. Due to this, we chose $\beta = 1$ for the remainder of our evaluation trials. Overall evaluation statistics are presented in Table 2.

Based on these results, we conclude that our personalized community detection algorithm has higher precision, and therefore assigns nodes that are related to each other in the ground-truth dataset to the correct...
partition more frequently than the default Girvan-Newman method. The recall of our method is slightly lower than the Girvan-Newman, which we believe to be a result of our method’s tendency to cluster at a higher resolution. The Rand index is a measure of the similarity between the ground truth and the partitioning result. The result in our algorithm is slightly higher than in the Girvan-Newman method, meaning our method has better fitness when it comes to approximating the ground-truth. The F-value, being the harmonic mean of precision and recall, can be considered a reasonable metric of the efficacy of each algorithm. By F-value, our method slightly outperforms the Girvan-Newman method. The Girvan-Newman method does indeed produce reasonable clusters, but our algorithm, when taking into account prior preference, performs better overall on the papers graph.

5 Conclusion

In this paper, we propose a personalized community detection that is an extension of the Girvan-Newman divisive method. We test our approach on a papers network created from data in the ACM digital library. Our results indicate that our enhanced approach is able to incorporate user preference and cluster nodes around user preference with a higher resolution without losing the graph’s overall topological structure.

The Girvan-Newman method encounters some problems when applied to some directed graphs with special structure, and so the usual approach is to treat a directed graph as undirected when using this method for community detection. Our method, which is derived from the Girvan-Newman method, suffers from the same problem. Furthermore, the Girvan-Newman method runs with a time complexity of $O(nm)$ on an unweighted network, where $n$ is the number of nodes and $m$ is the number of edges. As such, it has a relatively high computational cost compared to other state-of-the-art partitioning methods like InfoMap. We intend for this personalized community detection algorithm to be able to respond quickly to different users with different preferences, but for each new user preference the method must be run again in its entirety. As such, we are interested in seeing our concept of incorporating our core idea of a user preference prior into other partitioning methods that overcome some of the shortcomings of the Girvan-Newman method. We are considering the use of genetic algorithms, which may allow the graph to be partitioned only once. In such a case, the partition for a new user preference can be calculated from a previous partition through the optimization of some well-defined objective function. We are also interested in seeing how our method performs on more complex heterogeneous graphs (Liu, Yu, Guo, & Sun, 2014).

References

Human-assisted OCR of Japanese Books with Different kinds of Microtasks

Kosetsu Ikeda¹, Ryota Hayashi², Kiyonori Nagasaki³, Atsuyuki Morishima²
¹Chiba University, Japan
²University of Tsukuba, Japan
³International Institute for Digital Humanities, Japan

Abstract

Human-assisted OCR is a common approach for transcribing books and has been used for many digital library projects. This paper reports our project for transcribing the book collections of National Diet Library in this approach. Our project is unique in two ways. First, we try to extend the human-assisted OCR approach by distributing microtasks in many ways other than just showing tasks in the specific Web page on PC screens. Second, we deal with Japanese books which have thousands of characters, some of which look similar to each other. This paper shows that we can expect high-quality results even if we transcribe Japanese texts with microtasks and the number of preformed microtasks to be stable if we distribute microtasks to equipment with which worker perform microtasks in their daily lives.

Keywords: Digital transcription; Crowdsourcing; Microtask


Copyright: Copyright is held by the authors.

Acknowledgements: The authors are grateful to the contributors to Crowd4U, whose names are partially listed at http://crowd4u.org. This research was partially supported by the Grant-in-Aid for Scientific Research (#25240012) from MEXT, Japan and by CREST, JST.

Contact: lumely@chiba-u.jp

1 Introduction

Recently, digital transcription with crowdsourcing attracts much attention, and there are a large number of crowdsourced digital transcription projects in the library and digital archive domains (Australian Newspaper Digitisation Program, n.d.; National Archives Transcription Pilot Project, n.d.; Terras, 2012; Causer, Tonra, & Wallace, 2012; Ishihara, Itoko, Sato, Tzadok, & Takagi, 2012). A common approach in digital transcription projects is the human-assisted optical character recognition (human-assisted OCR) approach, where workers correct the results of OCR software or directly transcribe characters that the OCR fails to transcribe (reCAPTCHA, n.d.).

This paper reports our project for transcribing the book collections of National Diet Library (NDL) (National Diet Library Digital Collections, n.d.) in this approach. Our project is unique in two ways. First, we try to extend the human-assisted OCR approach by distributing microtasks in many ways other than just showing tasks on PC screens. Most of existing projects ask crowd workers to perform such tasks on the specific Web page. In contrast, we try to extend the approach by distributing tasks in many other ways. For example, we distribute tasks to the task-on-the-floor (shortly TOF) system, where people walking perform tasks while walking on the floor, and to the smartphone lockscreen application, where people who want to unlock their smartphones perform tasks.

Second, we deal with Japanese books which have thousands of characters, some of which look similar to each other. This raises an interesting question because people performing microtasks may not pay a great attention to the task compared to the case where they perform tasks on the specific Web page.

The main contributions are as follows. First, we explain our ongoing project whose approach is novel. To the best of our knowledge, our project is the first to try to distribute tasks in many ways other than PC screens in the digital library domains. Second, we show our preliminary results that suggest that the microtask approach is effective to some extent although the Japanese language contains thousands of characters some of which are similar to each other.

In our project, we only extract text data from digital images and do not consider the layout of book pages.
Figure 1: Overview of Crowd4U

Figure 2: Microtasks used in the project

2 The Hondigi Project and Crowd4U

Hondigi is an academia-based project to try to transcribe books in the National Diet Library digital collections, whose copyrights are expired. As a subproject of Hondigi, we have been exploring the effective usages of microtasks for transcribing books.

In this project, we use Crowd4U (Morishima, 2013), a nonprofit open microvolunteering and crowdsourcing platform for academic and public purposes. Figure 1 shows the overview of Crowd4U. A prominent feature of Crowd4U is the ability to distribute the registered tasks in many ways. Currently, tasks can be distributed to (1) any Web pages with embedded javascripts, (2) Crowd4U terminals that are located six universities in the world, (3) lockscreen applications that can be downloaded from Google Play, and (4) the TOF system that are located in four universities in Japan.

3 Microtasks Design

This section explains two tasks currently used in the project.

Task A: Identifying Non-Textual Part of Books.

The purpose of this task is to identify non-textual part of book contents. The identified part will be published as a list of figures.

Figure 2 (a) shows an example of this task. This task asks workers whether the shown image contains an illustration, chart, or photo. Each worker chooses one of “Yes” “No” or “I have no idea.” We apply majority voting for obtaining the final decision.

Task B: Choosing Correct Characters

This task asks workers whether the result of OCR is correct or not. Figure 2 (b) shows an example. The OCR outputs a ranked list of character candidates with a confidence value for each character. The task
first shows the candidate with the highest confidence. The target character is surrounded by a red box, and each worker chooses one of “Yes,” “No,” “I don’t know” or “It is not a character,” or directly enter the shown character to a text field. We used two versions of Task B. The first version do not allow workers to directly enter the characters. As we will show in the next section, this caused a problem and we added the text field to the task in the second version.

We ask more than one worker to perform the task and changes the shown candidate in the following way.

- If three workers answer “yes,” the final result is the character.
- If two workers answer “no,” we change the shown character to the next candidate.
- If three workers answer “It is not a character,” the final result is null.
- If a worker enters a character, we increase the number of “no” and we use the entered character as the next candidate.

4 Preliminary Results

We registered the tasks to Crowd4U and volunteer workers performed the tasks with PC screens, lockscreen applications, and the TOF systems.

(1) Number of Performed Tasks

We started to distribute the tasks on Feb. 2, 2015. We had 153,201 results as of Apr. 10, 2016. Figure 3 shows how the monthly number of performed tasks in the period. The two solid lines represent the total numbers for Task A and B, respectively. The dotted lines are breakdowns of the number for Task A.

Some observations are as follows.

First, the number of tasks dramatically increases if we distribute tasks to the lockscreen applications. The numbers on the line with diamonds are 1.36 times larger than those with triangles on average. This is because we distributed only Task A to the lockscreen applications.

Similarly, there is a huge impact of the TOF system on the number. The numbers on the solid line with squares are twice larger than those on the dotted line with diamonds on average. The only difference between them is that the former includes the number of tasks performed with the TOF systems.

We observe that the number of tasks are affected by several factors. First, workers tend to perform many tasks when the task are announced for the first time, but eventually they are getting tired and only a few active users continue to perform tasks. Therefore, it is effective to use distribution frameworks that do not depend on the motivation of workers, such as the TOF systems.

We placed the systems on university campuses and most of workers are university students. Therefore, the seasonal factor exists. For example, the number of tasks performed on the TOF systems becomes small on Summer and Spring vacations of Japanese universities.

(2) Quality of the Task B Results

For this task, we used OCR results generated by the ABBYY FineReader (ABBYY FineReader, n.d.) as a trial. We started to distributed the tasks on Nov. 12, 2015. We obtained 20,644 results and 3,979 characters as of Apr. 10, 2016. We selected the 285 out of the 3,979 characters in the first two paragraphs of the book “Iwanami Köza Nihon Rekishi” (The Historical Science Society of Japan (Ed.), 1935), and we manually compared the characters with the direct OCR results (i.e., the top-ranked characters of the OCR outputs).

Table ?? shows the comparison result. The $F_1$ score of the extracted characters is 0.13 points higher than that of the direct OCR results. The direct OCR results contained two misidentified characters (i.e., the parts of books that are not characters but identified as characters) but the crowd successfully removed them. On the other hand, there are five characters that the workers could not output the correct characters. This is because the workers answered “No” to all candidates in the list of characters the OCR output. The problem happened only with the first version of Task B and we found that no such problems happened with the revised version of Task B.

Table ?? explains the classification of incorrectly recognized characters. The first and second categories represent minor errors. The first category “Form is different” means that the meaning of character
is correct but the form is different. Some Japanese characters have different forms with the same meaning. The second category “Size is different” means that the shape is correct, but the size is different. The third to fifth categories are serious errors. As we explained, the OCR misidentified two characters and the workers could not give answers to five characters.

Workers succeeded in removing completely different answers although they are performing tasks in ways where they cannot necessarily concentrate on the tasks. And workers failed to correct different size characters because the difference of sizes of characters in our task is hardly distinguishable.

<table>
<thead>
<tr>
<th>Number of characters</th>
<th>Precision</th>
<th>Recall</th>
<th>F₁</th>
<th>TP</th>
<th>FP</th>
<th>FN</th>
</tr>
</thead>
<tbody>
<tr>
<td>OCR results</td>
<td>287</td>
<td>0.75</td>
<td>0.76</td>
<td>0.76</td>
<td>213</td>
<td>69</td>
</tr>
<tr>
<td>Task results</td>
<td>280</td>
<td>0.90</td>
<td>0.88</td>
<td>0.89</td>
<td>253</td>
<td>27</td>
</tr>
</tbody>
</table>

Table 1: Digital transcription results (for 285 characters)

<table>
<thead>
<tr>
<th>OCR results</th>
<th>Task results</th>
</tr>
</thead>
<tbody>
<tr>
<td>Form is different</td>
<td>5</td>
</tr>
<tr>
<td>Size is different</td>
<td>14</td>
</tr>
<tr>
<td>Completely different</td>
<td>48</td>
</tr>
<tr>
<td>Not a character</td>
<td>2</td>
</tr>
<tr>
<td>Couldn’t give an answer</td>
<td>0</td>
</tr>
</tbody>
</table>

Table 2: Classification of errors

5 Discussions

A weakpoint of the volunteer-based crowdsourcing is that the number of performed tasks is not stable. The number tend to be the largest when the introduction of new tasks are announced and gradually becomes small. We found that distributing tasks to equipment placed in their everyday lives, such as lockscreen applications and the TOF system, is effective to keep the number large. Since the forms of microtasks that can be distributed in such ways are limited to simple ones, effectively mixing tasks with different granularities is desired.

OCR software is not good at dealing with text containing figures and tables. In particular, some Japanese books contains texts that are overlapped by official seals. We found that crowdsourcing is effective to deal with such cases.
Some Japanese characters can appear in text in different sizes and they have to be distinguished. We found that the naive task design cannot deal with such cases. Using a square to surround the candidate character in the task may help workers grasp its size.

So far, we used fixed-sized microtasks showing only one character to each worker. We noticed that the size is often too small since OCR results are generally good. We will consider dynamically changing the granularity of tasks depending on the quality of OCR results.

6 Summary

This paper reported our project for transcribing the book collections of NDL. In the project, we try to extend the human-assisted OCR approach by distributing microtasks in many ways other than just showing tasks on PC screens and apply the approach to Japanese books which have thousands of characters. This paper explained the overview and the initial findings in our first attempt. Our experience so far suggests that (1) we can expect high-quality results even if we transcribe Japanese texts with microtasks. (2) If we distribute microtasks to equipment with which worker perform microtasks in their daily lives, the number of performed microtasks becomes stable, and (3) The task design is important for the efficiency of the transcription. Future studies include introducing tasks to correct the direction of the character. And to dynamically change the task granularity for improving efficiency.

References

Enter a Job, Get Course Recommendations

Nan Li², Naren Suri¹, Zheng Gao¹, Tian Xin³, Katy Börner¹, Xiaozhong Liu¹
¹School of Informatics and Computing, Indiana University Bloomington
²College of Transportation Management, Dalian Maritime University
³School of Information Resource Management, Renmin University of China

Abstract
In an interdisciplinary learning environment, students are facing difficulties to locate the right education opportunities, e.g., campus courses or MOOCs, to achieve their career goals. In this paper, we propose a novel student program planning system. Using the system, students can enter job preferences, e.g., "Software Engineer at Google", and text-based and network graph-based recommendation algorithms will suggest education opportunities that help students achieve their career goals. Preliminary results show that the proposed solution is promising in recommending students a personalized education plan.

Keywords: Student Program Planning; Education; Text/Graph Mining


Copyright: Copyright is held by the authors.

1 Introduction

Personalized student program planning and course scheduling services have been identified as an important problem in the past few years. As students are learning in a more interdisciplinary environment, they face an increasing amount of pressure to locate the right courses and programs. In previous studies, Bendakir and Aïmeur (2006); Werghi and Kamoun (2009); Chu, Chang, and Hsia (2003) proposed a number of information recommendation and data mining methods, i.e., association rules and decision trees, in support of program planning and course recommendation by students. In most of those studies, courses were recommended to the target student based on his/her (computational) profiles.

However, students may have different career goals, and they may have different needs when choosing what courses to take. For instance, two iSchool students may take the very same required courses in the first academic year (because they come from the same academic program), but they may be interested in different courses in the second academic year because of different career goals.

In order to address this problem, we proposed a novel student programme planning system that takes current expertise and career goal information into consideration. Unlike prior efforts, we propose a new method to recommend education opportunities that meet student’s career goals, e.g., jobs. As Figure 1 shows, the proposed system lets users (actors) submit job information. Text-based and/or network graph-based search and recommendation methods are used to recommend high quality courses or MOOCs. The system represents information on jobs, university courses and MOOCs via a text index and a network graph index.

Figure 1: Course Recommender System Architecture.
In the heterogeneous knowledge graph index, each job, company, course, topic, etc. is interconnected with nodes of other types. There are 9 types of nodes and 10 types of edges, see Figure 2. The recommendation engine uses a random walk algorithm to recommend courses or MOOCs that match a student’s past courses and the target job (query) node. Preliminary results show that the proposed method recommends potential educational opportunities customized to the student’s profile.

2 Literature Review

In an interdisciplinary learning environment, the volume of education-related information available is rapidly increasing. For instance, MOOCs exist for many topics and the line between online and on campus education is blurring (Pappano, 2012). This abundance of educational information has created the need to help students find, organize, and use resources that match their individual goals, interests, and current knowledge (Farzan & Brusilovsky, 2006).

Over the past years, a number of information retrieval/recommendation as well as data mining techniques have been developed for student course planning. For instance, Farzan and Brusilovsky (2006) proposed a CourseAgent system to recommend courses by leveraging students’ assessment of courses. Similarly, the CourseRank system proposed by Parameswaran, Venetis, and Garcia-Molina (2011) integrates a number of different features for course recommendation, such as course requirement and student feedback. Meanwhile, a number of machine learning methods, e.g., association rules by Bendakir and Aimeur (2006), graph theory by Chu et al. (2003), and decision trees by Werghi and Kanoun (2009), have been employed to enhance the course recommendation performance and to better serve students.

However, to the best of our knowledge, the approach does not exist to utilize student career goals and comprehensive job market information to recommend education opportunities. Subsequently, we detail our proposed method to address this problem.
3 Methodology

In this section, we describe the proposed methodology in detail along with the preliminary experiment results.

3.1 Data Collection and Indexing

For this project, we collected various kinds of data, including university course data, MOOC data, and job posting data. Using information extraction algorithms, we extracted different named entities from the text data, i.e., course/MOOC description and job posting content. See legend in Figure 2 for sample node and edge types. A listing of all different nodes types and the number of exemplars for each type can be found in the top-right of the figure.

However, entities are isolated in the text index and relationships are not explicit. For example, from a course recommendation viewpoint, *Information Retrieval, Information Visualization* and *Bayesian Network* should be interconnected implicitly or explicitly. In order to address this problem, we index all the courses, MOOCs and jobs on a novel heterogeneous knowledge graph that interlinks all the jobs, courses, and MOOCs via semantically typed links collected from Wikipedia. For instance, by extracting Wikipedia concepts from the *Data Scientist* job postings, on average, they are linked to *Machine Learning*, *Matlab*, and *Unstructured Data* nodes with transitioning probabilities (0.072, 0.024 and 0.010 respectively). The concepts (keywords) extracted from Wikipedia are also interconnected on the graph via page incoming/outgoing links.

The complete network graph has a total 395,030 nodes and 993,526 edges. There are 8,350 jobs, 716 university courses, 750 MOOCs, 1,774 companies, 6,924 company specialities, 38 job functions, 954 locations, 375,208 related entities (keywords extracted from Wikipedia), and 316 instructors.

3.2 Text-based Approach

In the system, students can use a text query to represent their career goal, and the system recommends courses and MOOCs based on the job query together with the probability that the course matches the job query, $P(\text{course}|\text{job query})$. In the proposed system, we utilized a two-step approach. First, student inputs text query is sent to the job text indexation and relevant jobs are fetched. Then, we extracted the keyword information from each retrieved (and top-ranked) job posting as *job query*. For instance, when student inputs *Soft Engineer* in the system, we first retrieve a number of job postings from the job index, and then extract keyword list from those postings as the *job query* to represent student information need. Note that, the extracted keywords associate with the weight, e.g., frequency or probability in the target job postings, which can be translated to the query vector or query language model for the next step.

Meanwhile, a pseudo relevance feedback approach is used to further enhance the recommendation performance. For instance, the most important words/entities are extracted from the top ranked course/MOOC descriptions to enhance the query quality and ranking results. More detailed pseudo relevance feedback algorithm can be found in (Yu, Cai, Wen, & Ma, 2003).

3.3 Graph-based Approach

As aforementioned, a graph with job, course, MOOC and keyword nodes is constructed for graphical recommendation. On a graph $G$, when we query for a job from job node to retrieve the course or MOOC node, the traversed path over the edges with probabilistic weights results in ranking the highly relevant education opportunities. The results are calculated using the random walk algorithm to recommend and rank the candidate’s educational opportunities. On the graph, if we use $N_j$ to represent the query job, and $N_o$ for a candidate’s course/MOOC, the ranking score can be represented by:

$$P(N_j \rightarrow N_o) = \sum_{i} \prod_{N_x \in I_{N_j \rightarrow N_o}} P(N_{x+1}|N_x)$$

\(^{1}\text{Wikipedia 2015 Dump}\)
Table 1: Preliminary Result for Different Ranking Features

<table>
<thead>
<tr>
<th>Feature</th>
<th>Precision</th>
<th>MAP</th>
<th>Precision@5</th>
<th>Precision@10</th>
<th>MAP@5</th>
<th>MAP@10</th>
</tr>
</thead>
<tbody>
<tr>
<td>Vector Space</td>
<td>0.6113</td>
<td>0.7589</td>
<td>0.6250</td>
<td>0.6113</td>
<td>0.7487</td>
<td>0.7589</td>
</tr>
<tr>
<td>Language Model</td>
<td>0.7050</td>
<td>0.8275</td>
<td>0.6750</td>
<td>0.7050</td>
<td>0.7835</td>
<td>0.8275</td>
</tr>
<tr>
<td>Relevance Feedback</td>
<td>0.6550</td>
<td>0.7463</td>
<td>0.7000</td>
<td>0.6550</td>
<td>0.8198</td>
<td>0.7463</td>
</tr>
<tr>
<td>Graph (J-K-C)</td>
<td>0.5465</td>
<td>0.6345</td>
<td>0.5165</td>
<td>0.5465</td>
<td>0.6486</td>
<td>0.6345</td>
</tr>
<tr>
<td>Graph (J-K-K-C)</td>
<td>0.5425</td>
<td>0.6104</td>
<td>0.5150</td>
<td>0.5425</td>
<td>0.5953</td>
<td>0.6104</td>
</tr>
<tr>
<td>Graph (J-K-J-K-C)</td>
<td>0.4200</td>
<td>0.5945</td>
<td>0.3750</td>
<td>0.2333</td>
<td>0.4632</td>
<td>0.3303</td>
</tr>
</tbody>
</table>

where \( I_{N_j \rightarrow N_o}^k \cdot F(path_u) \) is a path instance belonging to a path function \( F(path_u) \). The random walk probability from \( N_j \) to \( N_o \) on this path can be calculated by \( \prod_{N_x \in I_{N_j \rightarrow N_o}^k} P(N_{x+1}|N_x) \), with \( P(N_{x+1}|N_x) \) being the transitioning probability between the nodes on the graph.

For this method, we propose a number of different random walk based path functions on the graph, \( F(path_u) \). For instance, on the proposed graph schema, the job \((N_{\text{job}})\) and course/MOOC nodes \((N_{\text{opportunity}})\) can be interconnected via three important path functions, \( N_{\text{job}} \rightarrow N_{\text{keyword}} \rightarrow N_{\text{opportunity}}; N_{\text{job}} \rightarrow N_{\text{keyword}} \rightarrow N_{\text{keyword}} \rightarrow N_{\text{opportunity}}; N_{\text{job}} \rightarrow N_{\text{keyword}} \rightarrow N_{\text{job}} \rightarrow N_{\text{keyword}} \rightarrow N_{\text{opportunity}}. \) The first one uses the direct relations between course/MOOC and job via entity information, and the second and third ones use the relationship between entities (e.g., Information Retrieval and Machine Learning are interconnected on the Wikipedia graph) and the relationship between job and entities.

### 3.4 Informal Evaluation

Two graduate students were asked to use the course recommendation system prototype. Each of them entered 10 text queries (e.g., job titles) and rated each recommended course/MOOC as ‘useful’, ‘just OK,’ or ‘not useful’. In Table 1, we report the performance of different recommendation functions (overall ranking performance and top recommended education opportunities accuracy). Precision, MAP (Mean Average Precision), Precision@5, Precision@10, MAP@5 and MAP@10 are reported as the evaluation metrics. In the experiment, we examine three different graphical ranking functions, \( N_{\text{job}} \rightarrow N_{\text{keyword}} \rightarrow N_{\text{opportunity}} \) (J-K-C); \( N_{\text{job}} \rightarrow N_{\text{keyword}} \rightarrow N_{\text{keyword}} \rightarrow N_{\text{opportunity}} \) (J-K-K-C); and \( N_{\text{job}} \rightarrow N_{\text{keyword}} \rightarrow N_{\text{job}} \rightarrow N_{\text{keyword}} \rightarrow N_{\text{opportunity}} \) (J-K-J-K-C).

We find that the proposed method, both text and graph ranking functions, can be useful for student program planning with career information. Meanwhile, when using each individual ranking feature, the text-based approach outperforms the network graph ones.

### 4 Conclusion

In this study, we introduced a novel method and prototype system that recommends courses based on job queries. Unlike prior efforts, student can enter their career goal, and the text and graph-based recommendation algorithms can recommended optimized education opportunities, MOOCs and local courses, to the user. Even though we find the text ranking features outperform graph ones in the preliminary result, the graph recommendation features can be significant in the next ranking fusion stage (as Figure 1 shows). For instance, based on the learning to rank studies, studies (Liu, Yu, Guo, & Sun, 2014; Liu, Xia, Yu, Guo, & Sun, 2016) showed that graph-based approaches can provide more distinct ranking information, which can significantly enhance the recommendation performance (e.g., from learning to rank perspective, language model plus PageRank can outperform language model plus vector space).

In the future, we plan to integrate different ranking features (supervised ranking fusion) to further enhance the recommendation performance. In addition, we will run a formal user study to identify task accuracy and performance by different user groups that might be interested to use the system, e.g., university vs. MOOC students; full-time vs. part-time students.
References


Anachronism in Global Information Systems: the cases of Catalogue of Life and Unicode

Robert D. Montoya¹, Seth R. Erickson¹
¹University of California, Los Angeles

Abstract
The Catalogue of Life (CoL) and the Unicode Standard are examples of information systems that aim toward universals: the goal of the CoL is to provide a “single integrated species checklist and taxonomic hierarchy”; the goal of Unicode is to be a “universal character set” covering the world’s writing systems. In this preliminary research paper we present anachronism as a key obstacle in the design, expansion, and evolution of such systems. We highlight the preservation of concepts (of species and of writing systems) through their inclusion in these systems as an example of how such anachronisms materialize. The goal in this piece is to present a more nuanced understanding of how information and documentary systems (viz-á-viz, indexes, taxonomies, knowledge organization systems, etc.) create new, multiplicitous temporal spaces as part of their construction—knowledge that can then be applied as information professionals build these systems and subsequently evaluate their functionality and efficacy.

Keywords: Temporality; Knowledge Organization; Standardization; Catalogue of Life; Unicode


1 Introduction

The central concern of this paper is to examine how global information systems constitute new temporal spaces that typically reformulate the temporal identities of the documents they contain. Knowledge organization systems and technical standards radiate deeply in academic and professional practices, often facilitating the global communication of data and knowledge across scientific labs, scholarly domains, country boundaries, and societies. Global information collaboration now depends on distributed computational systems that facilitate what Elizabeth Eisenstein called the “cumulative change” Eisenstein (2009, p. 113) that is essential to the progression of, not only scientific discovery, but the evolution of humanity’s collective production of knowledge in all domains. Information systems establish standards that are fundamental to the successful collection of information and its subsequent browsability and usefulness. In particular, for the purposes of this paper, the standardization of time as both an impetus for and emergent principle of the systems will be the analytic focus.

One problem with information aggregation is identifying how systems are used to stabilize temporally emergent processes (e.g., evolution of species and elements of writing systems) in order make useable the entities constituted within those processes. Such stabilization involves a degree of temporal flattening that is problematic because some of the attributes relevant to the ‘use’ of these entities concern their temporal position (e.g., phylogenetic relations and whether they are extinct or not matters). This creates a tension that we understand as a kind of anachronism. We present two case studies, The Catalogue of Life (CoL) and Unicode to illustrate how temporality functions in anachronistic ways in knowledge organization systems and standards, melding numerous qualities of time into one “pleated” (Bowker, 2009) interwoven fabric. We use approaches from infrastructure studies—specifically the analytical method of infrastructural inversion—to foreground these systems’ temporal effects (Bowker, Baker, Millerand, & Ribes, 2009; Bowker & Star, 2000).

This paper isn’t merely asserting that information systems need to (or, as entities, do) evolve over time (we take this to be a given quality that defines them). This paper is speaking to the temporal effects—composed and unavoidable—inherent in the construction of these systems (i.e., by describing or standardizing entities through them). We describe such effects as anachronisms in the sense used by Christopher S. Wood and Alexander Nagel (2010) in their monograph, Anachronic Renaissance. Our framing of anachronism is
meant to highlight how we can study and disentangle these temporal effects (which are to be understood as both positive and negative, as much as they are inevitable). Working with and around anachronism both defines a fundamental purpose of these systems, as well as one of the key challenges in creating and implementing them.

2 Temporality and Anachronism: A Framework

In their study of temporality in Renaissance art, Nagel and Wood (2010) present the Ship of Theseus as a paradigmatic example of the conflict between temporal modes. This paradox from classical literature hinges on the question of whether a ship used by Athenians in an annual ritual continues to be the same ship even as its original planks and timbers, rotting over time, are replaced with new ones. Is the ship’s “ontological stability across time” rooted in material continuity or structural relationships? Nagel and Wood argue that these twin modes of conceiving sameness over time form a dialectic. For some objects at some times, “substitution”—replacing rotting planks with new ones—is grounds for ontological rupture; in other cases it is not. In a third category are objects that fluctuate between these poles, manifesting what the authors describe as a dual historicity (ibid., p. 31). Anachronisms are fluctuations within this third category—occasions in which an object’s relationship to the past oscillates between material and structural concerns.

Continuing in this vein, global universal information systems also manifest a kind of double historicity, in that they are generally seen as both fabricated artifacts (in the sense that Tennis, below, uses the term), as well as artifacts that are meant to somehow retain the historical and narrative qualities of the documents they mean to organize and fix within the frame of an information structure. Like the new planks on the Ship of Theseus, we understand the “preservation” of endangered entities through classification and standardization (in order to retain some capacity to use them) as a temporal reframing in which those entities enter into new, potentially anachronistic relationships with other entities/concepts. Metatheoretical examinations of knowledge organization (KO) systems have been especially sensitive to the temporal dimensions of information systems. Joseph Tennis understands classification systems to be artifacts that evolve as an “intellectual pursuit” and offers “subject ontogeny” as a method of tracing this temporal change as classification schedules change punctuatedly over time (Tennis, 2002). Tennis has more recently articulated the notion of “second-order classification” (2015, p. 246), a metatheoretical “contour” of KO systems concerned with the changes that classification systems undergo over time. The temporal anachronisms described here are an inherent and inevitable byproduct of the “work of maintaining” these infrastructures “over time” (2015, p. 246). It is within this discourse that this study situates itself, but broadens the theory to encompass more general standardizing infrastructures, and focuses more directly on the temporal change instantiated by including concepts in the system at all.

3 Case 1: Catalogue of Life

The CoL is a global initiative with the primary goals of creating the most comprehensive species checklist and taxonomic hierarchy of the world’s biological organisms, estimated to be upwards of 1.7 million species (Species 2000, 2015). The CoL merges biodiversity data from nearly 160 Global Species Databases (GSDs), representing all major taxon groups (including animals, microorganisms, fungi, plants, and viruses). The CoL’s “management hierarchy” serves as the organizing mechanism for contributed GSD data. The hierarchy includes taxon relationships, distribution information, and nomenclatural variants that effectively can be used to centralize the knowledge output of the numerous taxonomies produced by the biodiversity community (Ruggiero et al., 2015). This structure facilitates the ready use of species data within other online infrastructures, as well as browsing of the established taxonomic tree. The CoL checklist and taxonomy is integrated in many online data aggregators, including the Global Biodiversity Information Facility, Encyclopedia of Life, and the Barcode of Life Database, to map data to accepted species nomenclature. As David Remsen indicates, “scientific names have become the primary means for referencing a taxon whenever a piece of information is intended to refer unambiguously to a particular type of organism or group of organisms” (2010). Given the importance of nomenclature as the “label” for taxa (Remsen, 2010), the CoL has dedicated itself to providing the most authoritative listing of species names for use in the globe’s biodiversity data ecology. Quite separate from its function as an authoritative communication and aggregative tool, however, this
examination critiques the CoL taxonomy’s internal ontological “consistency” (Furner, 2009, p. 12), and asks how, as a practical knowledge organization system, it transforms the temporal qualities of the concepts it seeks to document.

3.1 Preservation of Concepts in Catalogue of Life

Comprehensive species lists are closely aligned with biodiversity conservation activities. Stemming in part from early meetings of the Convention on Biological Diversity (CBD)—an international body dedicated to promoting sustainable development—nomenclature and taxonomy have been articulated as a core and necessary practice to understanding what species exist and how such identification information can be used “for effective decision-making about conservation and sustainable use” (Convention on Biological Diversity, 2016). The CoL strives to meet this pressing need by creating a “universal and complete reference” in order to “monitor, manage and protect biodiversity resources” (Species 2000, 2015).

Mapping data about species to names in the CoL, however, is no trivial matter. Despite rules dictating the application of names to taxa, nomenclature is hardly simple or consistent in practice. What defines a species (its circumscription) changes over time—one species can be given many names independently, multiple species can be merged under one species name after evaluation, etc. Each circumscription is, by definition, an approximation “equivalent to generating a new hypothesis in other branches of biology,” and such a hypothesis is always open to new interpretations as new forms of evidence or new modes of data analysis (computational or otherwise) are introduced (Gaston & Mound, 1993, p. 139). Without a direct link to the species concept it represents at a given place and time, it is difficult to know exactly what species concept a name refers to. Franz, Peet, and Weakley (2008, p. 64), highlight this disjoint between “names and taxonomy” and the problems it causes when a “name and its meaning evolve independently.”

One of the goals of the CoL is to impose a sense of order within a nomenclatural and taxonomic landscape that is defined by such conceptual fluctuation and divergent practices. Names (and thus the species concepts that these names are associated with) are preserved and fixed within the system and mapped accordingly to nomenclatural synonyms, misspellings, and other variants. Names are then embedded within a management taxonomy that is used to both map species concepts together in some manner consistent with taxonomic opinion.

3.2 Temporality and Anachronism in the Catalogue of Life

How then does temporality function as an emergent property of the CoL, and how do such temporalities perform anachronistically within the system? The management hierarchy can serve as one example, particularly since it is the primary vehicle by which users navigate and browse the system’s nomenclature. Once the CoL stabilizes nomenclature, those names are then embedded within a “consensus higher level classification” (Ruggiero et al., 2015, p. 2). With no “consensus among the world’s taxonomists concerning which classification scheme to use for the overall hierarchy of life” (Ruggiero et al., 2015, p. 2), the management hierarchy presents an alternative taxonomic schema agreed upon by a group of taxonomic experts. The management hierarchy alphabetizes taxa “below the rank of infra-kingdom” for “easier searching by those not familiar with the phylogenies of the many taxa therein” (Ruggiero et al., 2015, pp. 9,65). This rearrangement is due to the CoL function as a standard more focused on discovery rather than representing a cohesive taxonomic argument. Individual GSDs (each covering different taxonomic groups) are then attached to this taxonomic framework to facilitate a shareable ‘tree of life’ that is ‘glued’ together by this central taxonomic backbone (Species 2000, 2016).

The CoL becomes an amalgam of multiple taxonomic scales and analytic approaches (phylogenetic, evolutionary, etc.), exhibiting what Nagel and Wood term a “clash of temporalities” (Nagel & Wood, 2010, p. 37). Internally consistent temporal scales, however, are foundational to biological taxonomies used in practice, such as those represented by GSDs brought into the CoL. “Phylogenetic frameworks” are the typical “basis for … biological classifications” (Ruggiero et al., 2015), and inherent in these estimations of the “evolutionary past” (Baldauf, 2003, p. 345) is the understanding that the species on the tree of life evolve according to some quantifiable temporal timeframe. As an example, for trees formulated with molecular data, the “lengths of the branches correspond to the amount of evolution” (Baldauf, 2003, p. 346) between two species nodes. As a temporal melange, the CoL is not an internally consistent system, limiting its use...
in practice. Further, alphabetizing taxa in the CoL compresses the phylogeny-specific information that is a unique aspect of interpreting a GSDs ontology. In the CoL’s case, anachronisms are concealed by the management hierarchy implemented across the system to form a functional whole. Something extremely valuable is gained in this process of temporal compression, however, and that is the effective management of data that is essential to building biodiversity knowledge as part of a global, collective endeavor.

4 Case 2: Unicode

The Unicode Standard defines a “universal character set”—a database of letters, symbols, ideograms, and other types of characters that can be used in plain text data (Unicode Consortium, 2016). Unicode is a critical technical infrastructure as it enables consistent handling and interpretation of textual data across disparate computer systems. Over the last decade the adoption of Unicode has steadily increased; as of September 2016, close to 88% of all websites are encoded with UTF-8, a character encoding format that implements Unicode (W3Techs, 2016). Unicode is an ongoing project in the sense that the Unicode Consortium—the standard’s governing body—continues to accept proposals for new characters and scripts (i.e., collections of characters forming whole writing systems). While much of the current public interest in the standardization process concerns the addition of emojis (the pictographic characters frequently used in mobile messaging apps), most additions to Unicode begin as proposals from a small but dedicated group of linguists, typographic experts, and historians committed to the project of encoding the world’s writing systems (Warzel, 2016). A notable project in this respect is the Script Encoding Initiative (SEI) at the University of California, Berkeley, which facilitates the preparation of proposals for as-yet unencoded scripts (SEI, n.d.). Although Unicode version 9.0 (the most recent) includes 128,172 characters covering 135 modern and historical scripts (West, 2016; Unicode Consortium, 2016), the SEI website notes over a hundred scripts that are not yet part of the standard (SEI, n.d.).

4.1 Preservation of Concepts in Unicode

In what sense are Unicode characters concepts and how might such concepts become endangered and deserving of preservation? The answers to these questions are perhaps less apparent than in the case of the Catalogue of Life. The conceptual nature of Unicode characters is best demonstrated in the standard’s formal distinction between characters and glyphs. Characters are defined as “abstract representations of the smallest components of written language that have semantic value” (emphasis added), whereas glyphs “represent the shapes that characters can have when they are rendered or displayed” (Unicode Consortium, 2016, p. 15). Unicode is concerned exclusively with abstract characters, not glyphs. Although the published standard includes character glyphs, they are provided for reference only: Unicode characters are not defined by their graphical properties. Formally, each Unicode character—or code point—is defined by a unique numerical value (expressed in hexadecimal) and a canonical name meant to convey the character’s “semantic value” (e.g., the code point for the letter “i” is: U+0069 → LATIN SMALL LETTER I).

With the distinction between characters and glyphs in mind, Walsh and Hopper aptly describe Unicode as “an idealist, abstract, Platonic standard” (2012, p. 68). As the authors note, however, Unicode’s idealism isn’t an ontological argument about characters or the “significance or insignificance of material aspects of a document” (p. 69). Unicode defines characters as it does—as code points—in order to stabilize them as non-ambiguous, fixed entities, well insulated from the messy business of font rendering and bit-level representation. It creates a space of usable abstractions, not true abstractions. This insight is nicely put in one of the standard’s accompanying technical notes: “Because the Unicode Standard is a character encoding standard and not the Universal Encyclopedia of Writing Systems and Character Identity, the stability and uniqueness of published character names [e.g., LATIN SMALL LETTER I] is far more important than the correctness of the name” (Freytag et al., 2006).

Preservation in Unicode is, thus, about stabilizing the elements of writing as computationally tractable entities—a project that becomes more urgent as our relationship to text is increasingly digitally mediated. Deborah Anderson, head of the SEI, equates standardization with preservation as follows: “While the popular media has focused on the effort to save biological diversity and endangered languages, the case for preserving the writing systems of languages is largely unnoticed. Saving scripts by including them in
Unicode will help document the variety of writing systems while also enabling their study, appreciation, and use” (2005, p. 27).

4.2 Temporality and Anachronism in Unicode

The temporal rifts and foldings—the anachronisms—of Unicode become apparent by adapting an analytic approach from infrastructure studies, by asking when is a character? (Ruhleder & Star, 1996). On the one hand, the formal definition of a Unicode character can be dated precisely to the publication of the Unicode version in which it first appeared. (These dates are important because, once assigned, code points cannot be re- or un-assigned). On the other hand, characters are enmeshed in historical processes that both precede and follow from the standardization process. The inclusion of a character in Unicode does not bring an end to its graphical or semantic evolution; rather, it reconstitutes the structural and material conditions within which this evolution occurs.

Consider two quite distinct cases: the Han ideograph ([chinese character]) meaning “tool” in Chinese and Japanese was introduced in version 1.0.1, in June, 1992; similarly, the pistol emoji () was officially adopted in version 6.0, in October, 2010. The temporal flattening introduced by the standardization process belies these characters’ historical complexity. In the first instance, regionally-specific variants of Han ideographs like ([chinese character])—variants that are the product of centuries of use and evolution in Chinese, Japanese, Korean, and Vietnamese writing systems—are often merged in Unicode through a process known as Han Unification (Lunde, 2008). The character ([chinese character]) is seen here in a simplified Japanese form, an effect of the font rather than the underlying code point, which has no regional specificity. On a much more compressed time scale, Apple’s recent decision to change the way the pistol emoji appears on its operating systems—from an image of a lethal weapon () to that of a non-lethal water pistol ()—demonstrates the fragility of Unicode’s standardization efforts. Apple’s change generated significant outcry as an unwelcome intervention in the emoji’s semantic value (Zittrain, 2016). These brief examples illustrate that the standardization process is constitutive, limited, and a reconfiguration of the relationships within which writing systems evolve.

5 Conclusion

In this discussion we have introduced anachronism as a potentially useful way of understanding how the standardization and fixation of concepts within information systems creates new temporal spaces that are infrastructurally specific. We are moving toward an argument that temporal changes are not just challenges for designers and users of information systems; information systems also constitute temporal structures that merit closer examination. This project suggests several avenues for future work. Most pressingly, the themes of materiality and structural relationships suggested by Nagel and Wood need to be more thoroughly explored in relation to the Catalog of Life and Unicode. Within Information Studies (and Knowledge Organization, more specifically), work is often (rightly) focused on interconcept and intraconcept entity relationships, but such work must be more attuned to how temporality functions within these ontological spaces, and how such activity redefines concepts by their alignment with other concepts articulated under discreet and separate circumstances. Furthermore there are important distinctions between forms of “stability” suggested in these systems that should be elaborated. Such lessons can then be used to create an analytic framework that can support similar examinations of temporality within the field of Information Studies and the STS community.

References


Internal/External Information Access and Information Diffusion in Social Media

Tian Xia¹, Xing Yu², Zheng Gao³, Yijun Gu⁴, Xiaozhong Liu⁵
¹School of Information Resource Management, Renmin University of China
²Dept. of Human-Centered Computing, Indiana University Indianapolis
³Dept. of Information and Library Science, Indiana University Bloomington
⁴School of Cyber Security, People’s Public Security University of China

Abstract
As social media platform not only provide infrastructure but also actively perform algorithmic curation for profit and user experience, it leads to an information filter bubble phenomenon: users are trapped in their own personalized bubble and are exposed only to the opinions that conform their beliefs and interests, thus potentially creating social polarization and information islands. However, filter bubbles hardly restrict all the users in a large social network, some information explorers can break the bubble and bring external global knowledge back to the internal network. In this paper, we investigate this assumption via hashtag adoption prediction. First, we construct a heterogeneous graph and extract 17 features to describe the event of hashtag adoption. Then, we generate learning instances and train a lasso regression model to do prediction. Preliminary results show that information explorers are more likely to adopt new hashtags than others, thereby more internal and external information can be diffused via these special users.

Keywords: Social Media; Information Adoption; Information Diffusion; Information Access; Weibo

1 Introduction
The proliferation of social media is bringing about significant changes in how people perceive and make sense of their world (Pak & Paroubek, 2010; Shuai, Liu, Xia, Wu, & Guo, 2014). Millions of individuals communicate with each other through a variety of social media platforms, sharing pertinent information about the world as well as the most minute details of their social lives, thereby collectively shaping each others’ culture and worldview. However, in addition to allowing information to travel freely through social ties, many social media platforms perform an information language/policy/network/algorithmic barrier. This curation raises an important concern, often referred “filter bubbles,” where people are increasingly trapped in their own information “bubble”—being exposed only to information that conforms to their existing beliefs and political positions, potentially creating information “islands” and potentially social polarization. Note that, in a large social network, in most cases, filter bubbles hardly restrict all the users, and some information explorers can always break the bubble, while bringing some global knowledge back to the local network.

In this study, we investigate this interesting problem by leveraging massive Weibo data. Unlike most popular microblogging systems, most Weibo users (in China) are restricted in a local information network because of different local law reasons (Zhu, Phipps, Pridgen, Crandall, & Wallach, 2013). When Weibo users trying to access global information liek Facebook or Youtube, they usually have two alternatives, 1. using VPN or proxy servers; 2. access Weibo service outside mainland China. We define them as “Information Explorers”, and call other Weibo users as common users. In this paper, we will study whether information explorers, who have global information access, tend to adopt and broadcast topics in social network, thereby resulting in some global knowledge back to the local network.

2 Related Work
Information diffusion has been widely studied in recent years, it can be defined as the process by which a piece of information (knowledge) is spread and reaches individuals through interactions (Zafarani, Abbasi, &
Liu, 2014). So far, a lot of effort have been made to modeling how information spreads in social network, and until recently, some researchers put more attentions to the filter bubble problems, and propose new method like cross social media recommendation to break the bubbles (Liu, Xia, Yu, Guo, & Sun, 2016; Liu, Yu, Gao, Xia, & Bollen, 2016).

Because of technique, culture and different country policy reasons, Twitter and Weibo, the most popular microblogging systems, are isolated from each other. Therefore, current research work about filter bubble isolation problem focused on comparing the difference of Twitter and Weibo, and try to find some way to connect these two social networks together (Shuaal et al., 2014; Liu, Xia, et al., 2016). Besides above efforts for fighting with the filter bubbles, here we first propose another assumption that filter bubbles can be broken by some special users, who can access external information and bring it back to the internal network, and we verify this assumption by topic adoption prediction method.

3 Methodology

Are Information Explorers more likely to adopt and broadcast new topics in social media? To answer this question, we extract comprehensive features from Weibo retweet diffusion graph to predict information adoption behavior, and investigate whether the key features are strong related to these special users or not. If that’s true, we can say that information explorers are more important for information access and adoption, thereby, they can introduce external information to the internal network. To verify this assumption, we do analysis via the following steps as shown in figure 1.

First, we construct a heterogeneous graph to describe how users retweet messages and use hashtags. Hashtag can be treated as a topic, when a user introduces a new hashtag in his or her message, we say that the user adopt a topic. Formally, we define a topic adoption event as a user $u_i$, adopts a topic $h_j$, at time $t$ (Liu, Yu, et al., 2016), denoted as a triple $event_t = (u_i, h_j, t)$.

Let $G = (V, E)$ represents the graph for user-retweeting and hashtag adoption hybrid network, where each node $v \in V$ represents a Weibo user $u$ or a hashtag $h$, and edge $(u_i \rightarrow u_j) \in E$ represents user $u_i$ has retweeted a message posted by $u_j$, edge $(u_i \rightarrow h_j) \in E$ represents user $u_i$ has used hashtag $h_j$ in the past. After $G$ was created, all isolated nodes are removed to reduce the time and space cost. Figure 2 is an snippet of graph $G$, where $h_1$ and $h_2$ are hashtag nodes, and $u_i (i \in [1, 7])$ represents the user node, in particularly, $u_3, u_5$ and $u_6$ are information explorers in this figure.

The topic adoption model predicts if user $u_i$ will adopt hashtag $h_j$ in the future. Previous studies show that some features extracted from retweeting network are very useful for prediction (Yang, Sun, Zhang, & Mei, 2012), include: in-degree/out-degree, hashtag numbers that $u_i$ used and prestige like PageRank score. In our study, we extract more features related with information explorers from above graph $G$, all the features we used are listed in table 1.

where $f_i$ in table 1 is the $i^{th}$ feature we extracted, $u_i$ represents any user node of graph $G$, and $h_j$ represents a hashtag. All neighboring explorers of current node $u_i$ is denoted as set $N_1$, while $N_2$ is the set of neighboring explorers of all $u_i$’s neighbors. Take user node $u_1$ in figure 2 as an example, its neighboring explorers set $N_1(u_1) = \{u_5\}$, and $N_2 = N_1(u_2) \cup N_1(u_5) = \{u_3, u_6\}$.

Given two time periods $T = [t_1, t_2]$ and $\Delta T = [t_2, t_3]$, we first construct graph $G$ by using the data of period $T$, and then extract all features for each user from the graph $G$, these features are assumed as...
latent variables. Next, we collect all \((u_i, h_j)\) pairs which meet: \(u_i\) do not use hashtag \(h_j\) in the time period \(T\) and do adopt \(h_j\) in the period \(\Delta T\), these pairs constitute the positive instances. Otherwise, if \(u_i\) do not use \(h_j\) in both period \(T\) and \(\Delta T\), it means a negative instance. Therefore, the dependent variable is boolean: 1 indicates \(u_i\) adopt \(h_j\) and 0 means \(u_i\) does not.

Once we get the learning instances, we use the lasso method (Tibshirani, 1996) to train the classification model, based on which we interpret the relationship between the extracted features and the topic adoption prediction. The lasso approach allows us to carry out feature selection while training the model by adding a \(l_1\) penalty to the loss function of the logistic regression. It can help us select a most effective subset of all the candidate features. The penalized loss function is defined as:

\[
L = \sum_l (y_l - \sum_p \beta_p x_{lp})^2 + \lambda \sum_p ||\beta_p||_1
\]

where \(x_{lp}\) denotes the \(p^{th}\) feature in the \(l^{th}\) datum, \(y_l\) is the value of corresponding response, and \(\beta_p\) denotes the regression coefficient of the \(p^{th}\) feature. The last part \(\sum_p ||\beta_p||_1\) is the \(l_1\) penalty, and parameter \(\lambda\) controls the penalty strength.

By adding the \(l_1\) penalty, the important features will have high regression coefficients, and irrelevant/redundant features coefficient will be shrunk to zero. Therefore, the coefficients of explorer-related features would be more important in information adoption and diffusion.

## 4 Experiment

We extracted Weibo users, hashtags, and various kinds of relationships from 12,362,489 Weibo messages. The data covered the time period \(T\) from September 17, 2012 to September 23, 2012 (7 days) and \(\Delta T\) from September 24 to 25, 2012. We find 50836 information explorers by their geography locations, and the final graph \(G\) contained 328,065 nodes and 783,811 edges. From above dataset, we extracted 4,502 positive instances and randomly sampled the same number of negative instances to make the learning instances balanced.

We split all 9,004 instances into two groups, 6,303 instances (70%) for training the model, and 2,701 instances (30%) for testing. We employed a 10-folds cross validation to tune and find the optimized parameter \(\lambda\) for the shrinkage penalty. And the experiment was implemented and carried out using the R language.

<table>
<thead>
<tr>
<th>Id</th>
<th>Feature Description</th>
<th>Id</th>
<th>Feature Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>(f_1)</td>
<td>In-degree of (u_i)</td>
<td>(f_6)</td>
<td>(f_{12}) The size of (N_1) (for (f_6)), (N_2) (for (f_{12}))</td>
</tr>
<tr>
<td>(f_2)</td>
<td>Out-degree of (u_i)</td>
<td>(f_7)</td>
<td>(f_{13}) Average PageRank of (N_1) (for (f_7)), (N_2) (for (f_{13}))</td>
</tr>
<tr>
<td>(f_3)</td>
<td>PageRank of (u_i)</td>
<td>(f_8)</td>
<td>(f_{14}) Maximum PageRank of (N_1) (for (f_8)), (N_2) (for (f_{14}))</td>
</tr>
<tr>
<td>(f_4)</td>
<td>Number of (u_i) use (h_j)</td>
<td>(f_9)</td>
<td>(f_{15}) Average in-degree of (N_1) (for (f_9)), (N_2) (for (f_{15}))</td>
</tr>
<tr>
<td>(f_5)</td>
<td>(u_i) is an info-explorer or not</td>
<td>(f_{10})</td>
<td>(f_{16}) Average out-degree of (N_1) (for (f_{10})), (N_2) (for (f_{16}))</td>
</tr>
<tr>
<td>—</td>
<td>—</td>
<td>(f_{11})</td>
<td>(f_{17}) Total number of (u \in N_1/ u \in N_2) use (h_j)</td>
</tr>
</tbody>
</table>

Table 1: Features for Predicting Information Adoption

<table>
<thead>
<tr>
<th>Prediction</th>
<th>true</th>
<th>false</th>
</tr>
</thead>
<tbody>
<tr>
<td>false</td>
<td>1214</td>
<td>108</td>
</tr>
<tr>
<td>true</td>
<td>136</td>
<td>1243</td>
</tr>
</tbody>
</table>

Table 2: Confusion Matrix

<table>
<thead>
<tr>
<th>Id</th>
<th>Coefficient</th>
<th>Id</th>
<th>Coefficient</th>
<th>Id</th>
<th>Coefficient</th>
</tr>
</thead>
<tbody>
<tr>
<td>(f_1)</td>
<td>0.0670</td>
<td>(f_6)</td>
<td>-0.4136</td>
<td>(f_{12})</td>
<td>1.0</td>
</tr>
<tr>
<td>(f_2)</td>
<td>-0.0006</td>
<td>(f_7)</td>
<td>-2.315E-5</td>
<td>(f_{13})</td>
<td>1.0</td>
</tr>
<tr>
<td>(f_3)</td>
<td>-2.869E-5</td>
<td>(f_8)</td>
<td>4.715E-6</td>
<td>(f_{14})</td>
<td>4.163E-8</td>
</tr>
<tr>
<td>(f_4)</td>
<td>0.0007</td>
<td>(f_9)</td>
<td>0.0167</td>
<td>(f_{15})</td>
<td>-0.0108</td>
</tr>
<tr>
<td>(f_5)</td>
<td>0.9318</td>
<td>(f_{10})</td>
<td>-0.0067</td>
<td>(f_{16})</td>
<td>-0.0015</td>
</tr>
<tr>
<td>—</td>
<td>—</td>
<td>(f_{11})</td>
<td>0.5305</td>
<td>(f_{17})</td>
<td>1.0</td>
</tr>
</tbody>
</table>

Table 3: Features Coefficient List Obtained by Lasso Regression Model

The confusion matrix from our validation process was shown in table 2. For the final model, the prediction accuracy has achieved 90.97% in total with a sensitivity of 89.93% and a specificity of 92.01%.
For feature interpretation, we presented the coefficients of all features in Table 3, the table cells filled by dot symbols represent irrelevant features, while the most important features were highlighted by yellow background color. We also presented a plot in Figure 3 to describe the lasso model changes as the feature added gradually.

![Figure 3: Changes of Lasso Regression Model When Features Are Added Gradually.](image)

In Figure 3, we presented how the coefficient of each feature changes with deviance explained in the logistic regression model. This process can be better described as a process of adding predictors gradually into the model. In the figure, we saw a NULL model (to the left end) where no features were included and a full model (to the right end) where all the features were included. From the left end to the right end, we can see each feature was gradually added to the model. The slope of each feature represented the relationship between the feature’s coefficient and the fraction deviance explained in the model. We noticed that at first the lasso result in a model contains only feature $f_5$. Then the rest of the feature entered the model gradually until all features were included. We found that feature $f_5$, feature $f_{11}$, and feature $f_6$ have steep slopes comparing with other features. In other words, these three features were considered to be the most important features in the model. Their coefficients were shown in Table 3.

According to above evaluation results and interpretation, we draw the following points:

1. Due to the strong positive correlation of feature $f_5$, information explorers have greater possibilities to adopt new hashtags, in another word, explorers were more likely to spread new topics in Weibo.

2. The prediction performance has strong and positive correlation with the number of hashtags used by neighboring explorers, and negative correlation with the number of neighboring explorers. It means that user adoption behavior can be significantly affected by the user’s neighboring explorers.

In summary, information explorers played an important role in topic adoption and diffusion, their information behaviors can help us to bridge the information gap between the inner social network and the outside.

5 Conclusion

In this study, we divide the Weibo users into two categories, i.e., common users and information explorers, and investigate the positive contributions of information explorers via predicting the topic adoption in Weibo social media. Based on feature analysis, we find that explorers have greater possibility than common users to adopt new hashtags, which means explorers tend to spread new information on Weibo. Therefore, they bring more information to others, and to a certain extent reduced the filter bubble problem.

We also find that neighboring explorers significantly affect the information adoption of users. If hashtag is exposed more times by the user’s neighboring explorers, this user is more likely to adopt the hashtag later. However, if a user has more neighboring explorers, he/she has less possibility to use the hashtag.
One possible reason is that the user has already obtained some kind of information from the neighbors, and has low motivation to diffuse these redundant information.

References

Yang, L., Sun, T., Zhang, M., & Mei, Q. (2012). We know what@ you# tag: does the dual role affect hashtag adoption? In Proceedings of the 21st international conference on world wide web (pp. 261–270).
Toward Understanding Causes of Anomaly in Dynamic Restaurant Rating

Lei Li\textsuperscript{1}, Zhao Lu\textsuperscript{2}, Danchen Zhang\textsuperscript{3}, Sanqiang Zhao\textsuperscript{3}, Ke Zhang\textsuperscript{3},
\textsuperscript{1}Nanjing University of Science and Technology
\textsuperscript{2}East China Normal University
\textsuperscript{3}University of Pittsburgh

Abstract
Rating score and text review are the most common features provided in online review systems to gather the opinions shared by users. Product rating distributions usually evolve dynamically over time and potentially accompany with some unusual changes, namely anomalies, which might be caused by product quality change or spamming attacks. In this preliminary study, we analyze the time-series of rating score distributions by using the data collected from Yelp restaurants, and we apply Principal Component Analysis (PCA) to detect anomalous time points. Through manually checking the corresponding review texts, we further investigate the underlying reasons leading to anomalous rating scores. The potential reasons we identified include food/service quality change, user preference, and review spam. Our study is envisioned to help business owners respond timely to unusual feedbacks and manage their business more efficiently.

Keywords: rating score; anomaly detection; PCA; online restaurant review


Copyright: Copyright is held by the authors.
Contact: lileiwelldone@gmail.com; zlu@cs.ecnu.edu.cn; daz45@pitt.edu; saz31@pitt.edu; kez11@pitt.edu

1 Introduction

Online review systems, such as Yelp and Amazon, allow users to provide the rating scores associated with text reviews about products and services. On one hand, the ratings and reviews play an important role in shaping the decisions of the potential customers. On the other hand, they provide the business with timely feedbacks from their customers. User behaviors in online review systems usually rely on personal preferences (X. Li & Hitt, 2008; Feng & Qian, 2013), word-of-mouth effects (Cheung, Luo, Sia, & Chen, 2009), and even peer influence via their social connections (Zhang & Pelechrinis, 2014; L. Li, Zhang, Zhou, & Zhang, 2016).

Online ratings are usually represented as a score, typically ranging from 1 to 5. The rating score distribution in the different time periods varies dynamically. Figure 1 shows the distribution of rating scores over time for a restaurant named "Lombardi’s Pizza". Different colors represent the percentages of ratings with scores from 1 (blue) to 5 (pink). We can see that the distribution sometimes suffers unusual changes. For example, at the very beginning, i.e., from time point $t_1$ to $t_2$, there is a sharp increase in the number of low rating scores. Such unusual changes are defined as anomalies (N. Günnemann, Günnemann, & Faloutsos, 2014).

In terms of anomaly detection, there are many related topics such as detecting spam reviews (Mukherjee, Liu, & Glance, 2012; Lim, Nguyen, Jindal, Liu, & Lauw, 2010) and fake reviews (Mukherjee et al., 2013; Lin et al., 2014). These studies mainly focus on detecting outliers, which are often attributed to random data corruptions (such as the spam or the measurement errors). However, not only outliers, but also anomalies exist in the ratings. Anomalies are the irregular data but follow a specific pattern, e.g., a restaurant receives consistently a lower rating scores due to a temporal decline of its service quality. Previous studies (N. Günnemann et al., 2014; S. Günnemann, Günemann, & Faloutsos, 2014) have focused on anomaly detection by analyzing the temporal dynamics of rating scores. In particular, they found that rating scores cannot accurately reflect the real quality of products. They extracted the base rating that reflects the regular quality of products, and discovered the time points at which the product’s rating shows anomalies (N. Günnemann et al., 2014). However, the hidden reasons underneath the occurrence of these rating anomalies are still not well examined.
In this work, we collect the rating and review data for restaurants in Yelp, and then we apply Principal Component Analysis (PCA) to detect time periods with anomalous rating distributions. We further analyze the content of corresponding reviews to examine various potential causes of anomalous ratings. Our study is envisioned to help business owners respond timely to unusual feedbacks and manage their business more efficiently.

![Figure 1: The Rating Distribution of a Restaurant 'Lombardi’s Pizza' Over Time](image)

2 Research Design

2.1 Dataset

We collect review data from Yelp for restaurants in New York City. In particular, we query the webpage of a restaurant iteratively to get the full list of its historical reviews, and each review has a tuple format \( \langle \text{time}, \text{user id, rating score, review text} \rangle \). For example, the pizza restaurants shown in Figure 1 in our data has 3,492 reviews starting from 2009/1/10 to 2014/11/8.

2.2 Method

In this section, we introduce how to use PCA to identify abnormal rating time points and how we recognize the causes of these anomalies.

Following the method from previous work (N. Günnemann et al., 2014), we first discretize the whole collection time period into epochs, with epoch size as 2 weeks, and we calculate the percentages of rating scores using all reviews in each epoch. Eventually, we organize the time-series of rating distributions as a matrix \( X \), where each column \( i \) represents the rating distribution in time epoch \( t_i \).

We then apply PCA on matrix \( X \). We first calculate the covariance matrix \( S \), that is \( S = \frac{1}{r-1} X^T X \). Actually, there is an interesting connection between Singular Value Decomposition (SVD) and PCA. In particular, let the SVD of matrix \( X \) be \( X = U \Sigma V^T \). Then we calculate the eigenvalue decomposition for \( S \):

\[
S = \frac{1}{r-1} V \Sigma^T U^T U \Sigma V^T = V \Lambda V^T
\]

where \( \Lambda \) is a diagonal matrix containing the eigenvalues \( \lambda_i \) of \( S \) in descending order.

In summary, PCA captures the dominant patterns in the original data by construct a \( k \)-dimensional normal subspace \( X^k \) in the original \( t \)-dimensional space. The remaining \( t-k \) dimensions form the anomalous subspace \( X^a \), where we have \( X^a = X - X^k \). We select \( k \) using Equation 2, such that \( X^k \) can explain at least 95% of the variance in the original data. If we need all the components to explain the 95% variance, it means there is no potential anomalous ratings in the data.

\[
\frac{\sum_{i=1}^{k} \lambda_i}{\sum_{i=1}^{t} \lambda_i} \geq 95\%
\]
For the remaining $t - k$ components, i.e., with $\lambda_i > 0$, we calculate the root mean square of each column in $X^a$ as the anomalous deviation $d$ of the rating distribution in each epoch from the normal status.

3 Results

Figure 2 shows a ramen restaurant’s abnormal deviations at different time epochs, and the mean anomaly value of this restaurant is set as the threshold. Time points 2, 5, 9, and 10, circled in the Figure 3, are the identified anomalies time points. Table 1 shows the four rating scores’ anomalous values. The time points 2, 5, and 10 show an increase in low-star ratings and a decrease in high-star ratings. Then we read the related reviews and find the following facts: most of the customers complain about waiting time, for example: “The main negatives about Totto ramen is the wait time (about 30 minutes).” They also complain about the small size of the restaurant: “This place was pretty small” and the menu: “the menu selections were limited.” We also find another interesting fact: the user’s personal preference affects the rating score, as in: “the miso ramen had thin white noodles which I’m less a fan of.” Moreover, we find there are users who try to advertise other competitive ramen restaurants: “Not as good as hide chan or ippudo.” Meanwhile, time point 9 shows the increase of high rating values because of low prices: “the prices are great.” The ratings also increase because of the staff’s good performance: “staff is amazing.”

Based on each anomaly time points’ five rating scores anomalous value, the anomalous time points can be divided into two categories: high-to-low anomaly time points, and low-to-high anomaly time points.

For the high-to-low category, we examine the reviews with the low rating scores to identify the negative reasons mentioned by the customers. High price, long waiting time, low food quality, bad taste, too crowd, bad surroundings, inconvenient payment method, and a limited menu are basically all reasons for anomalies related to the restaurant quality. From the review text, we can further check the detailed reasons for the anomalous rating: for example, customers might complain that the food has few sauces. Spam and personal preference are the other two reasons that can lead to a low-to-high rating anomaly.

For the low-to-high category, some positive reasons overlap with the negative reasons, because some drawbacks of the restaurant have improved over time. We also found the positive reasons that related to the restaurant quality, such as great delivery, nice service, and good location, good taste, short waiting time, low price, good surroundings, and high food quality. Also, the detailed positive reasons can be detected. For example, a customer report that they enjoyed the perfectly boiled dumplings. Personal preferences are another reason that results in a positive rating anomaly.

4 Discussion and Conclusion

We employ PCA model to identify the anomalies rating time points. Then through reading the reviews written in the recognized anomalies time points, we explore the reasons that lead to the anomalies. Three restaurants’ rating and reviews on Yelp are used to discuss the various causes of positive and negative anomalous ratings, including the reasons related to the restaurants quality, personal preference, and spam. Other time points that are not anomalous may have the same causes of getting high rating or low rating scores. However, the
Table 1: The absolute values of anomalous deviations over time

<table>
<thead>
<tr>
<th>Rating Scores</th>
<th>Time Points</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>2</td>
</tr>
<tr>
<td>1</td>
<td>-0.028</td>
</tr>
<tr>
<td>2</td>
<td>0.216</td>
</tr>
<tr>
<td>3</td>
<td>0.134</td>
</tr>
<tr>
<td>4</td>
<td>0.187</td>
</tr>
<tr>
<td>5</td>
<td>-0.284</td>
</tr>
</tbody>
</table>

anomalous time points make these causes easier to be noticed. The restaurant’s management can improve the quality of their service by simply doing a deep survey at the time of the anomaly’s occurrence.

This is a preliminary study on anomaly detection by only using sample datasets from Yelp. As a result, the reasons we have detected may not include all the causes of anomalous rating. In future work, we will use large-scale rating datasets to detect the various causes of anomalies for different kinds of entities by combining the textural mining method on reviews. Another area that we will further study is using other anomaly detection methods to identify anomalous time points more effectively.

References


Types of Tags for Annotating Academic Blogs

Lei Li¹, Daqing He², Danchen Zhang², Yu Chi², Chengzhi Zhang¹
¹Nanjing University of Science and Technology
²University of Pittsburgh

Abstract
Academic blog sites are popular academic information exchange platforms, and they have been widely used in recent years. Blogs in those sites are often annotated with tags, and the tags can help to describe, organize and retrieve these blogs. However, it is still unknown what types of tags are frequently adopted for annotating academic blogs. In this poster, we present survey results for detecting the usage of tag types, and its changes with the bloggers’ demographic information. We believe that our study can benefit users in their access to academic blogs and help the academic blog websites improve their services.

Keywords: tag types; blog tagging; social tag; academic blog; academic social media


Copyright: Copyright is held by the authors.

Acknowledgements: This work is supported by the Major Projects of National Social Science Fund (No. 16ZDA224).

Contact: lileiwelldone@gmail.com; dah44@pitt.edu; daz45@pitt.edu; yuc73@pitt.edu; zhangcz@njust.edu.cn (corresponding author. +86-84315963)

1 Introduction
Social tags have been widely used for annotating various kinds of online resources, such as blogs, pictures, books, videos, and music. The tags can help users in online resources retrieval and organization. Previous work showed that tags can be divided into many types, and they proposed plenty of tag type taxonomies for online resources of different kinds, e.g., book (Golder & Huberman, 2006; Wu, He, Qiu, Lin, & Liu, 2012), movie (Sen et al., 2006), TV program (Melenhorst & Setten, 2007) and URL (Xu, Fu, Mao, & Su, 2006). And these studies also confirmed that knowing tag types can enhance the effect of the tag recommendation (Xu et al., 2006) and resources search (Bischoff, Firan, Nejdl, & Paiu, 2008).

Academic blog website, which supports scholars to share and to acquire academic related information through the posted blogs, has emerged in recent years (Li & Zhang, 2016). Few studies, however, have investigated the tag types of the academic blogs, which are created by academic bloggers. Moreover, the tag type taxonomy is context-sensitive, which makes it inappropriate to apply the existing generic tag type taxonomy to the tags of academic blogs. Furthermore, it is still unknown what kind of information the academic blog tags convey. In this study, we combine the tag type taxonomies proposed by previous studies and the unique characteristics of blog tags to propose the tag types that are applicable to academic blogs. To achieve this goal, a survey instrument is adopted to explore the types of tags that are preferred by the bloggers for annotating academic blogs.

Through understanding the tag types used by bloggers for academic blogs, a tagging system could recommend proper tags to the blogs that still need tags. The users of the blog site thus has opportunities to more effectively access the relevant blogs, which could be a significant factor that influences scholars to be more actively engaging in academic social networks (Bik & Goldstein, 2013).

2 Method

2.1 Research site
We chose blog.sciencenet.cn¹ to send the questionnaires. Blog.sciencenet.cn is one of the most popular academic blog sites in China, and its registered users all have to provide real names, research institutions or

¹https://en.wikipedia.org/wiki/ScienceNet.cn
universities, and research fields. The majority of its users posted academic blogs to communicate with others, and the author of each blog assign tags to his/her blogs.

2.2 Sampling

Before sending out our questionnaire, we conducted a sampling on the users. Our target users are those with many tagged blogs because they have rich tagging experience, and they log in the website more frequently thus have a greater chance to reply our questionnaire. We crawled each blogger’s webpage, which includes the blog visited times, number of friends (#friend), number of blogs (#blog), registration time, last visited time, and last published time. In total, we collected 44,509 bloggers’ information. Then based on the bloggers’ number of published blogs, we selected top 5,000 bloggers. After further manual checking on whether these bloggers indeed tagged academic blogs, we finally identified 4,111 bloggers to send our questionnaire.

2.3 Questionnaire design

The questionnaire aims to investigate bloggers’ demographic information and their blog tagging behaviors. The demographic information is used to detect the association between the blogger’s basic information and the types of tags she uses.

Guided by the tag type taxonomies in previous work (Xu et al., 2006; Golder & Huberman, 2006; Melenhorst & Setten, 2007; Sen et al., 2006; Bischoff et al., 2008; Heymann, Paepcke, & Garcia-Molina, 2010), we classified the tags into four types, namely content-based tags, context-based tags, subjective tags and organizational tags. Then according to the characteristics of the blog, survey items of each tag type were proposed. Table 1 shows the survey items belonging to the above four tag types.

In the questionnaire, firstly, bloggers were asked to answer the question about their demographic information which includes the blogger’s discipline, education, gender and age. Then base on their tagging experience, the bloggers were required to rate their preference of using the tag types (showed in table 1) for annotating the blog. They can rate each tag type with a Likert scale from 1 to 5.

<table>
<thead>
<tr>
<th>Tag types</th>
<th>ID</th>
<th>Items</th>
</tr>
</thead>
<tbody>
<tr>
<td>Content-based tags</td>
<td>T1.1</td>
<td>Describing the blog topic, and existing in the blog’s content</td>
</tr>
<tr>
<td></td>
<td>T1.2</td>
<td>Describing the blog topic, but not existing in the blog’s content</td>
</tr>
<tr>
<td></td>
<td>T1.3</td>
<td>Describing the blog topic, and existing in the blog’s title</td>
</tr>
<tr>
<td></td>
<td>T1.4</td>
<td>Describing the categories of the blog</td>
</tr>
<tr>
<td>Context-based tags</td>
<td>T2.1</td>
<td>The blog’s publisher</td>
</tr>
<tr>
<td></td>
<td>T2.2</td>
<td>The blog’s published time</td>
</tr>
<tr>
<td></td>
<td>T2.3</td>
<td>The blog’s published location</td>
</tr>
<tr>
<td></td>
<td>T2.4</td>
<td>The source of the blog</td>
</tr>
<tr>
<td>Subjective tags</td>
<td>T3.1</td>
<td>The opinion to the blog</td>
</tr>
<tr>
<td>Organizational tags</td>
<td>T4.1</td>
<td>Self-organization tags</td>
</tr>
</tbody>
</table>

Table 1: Survey Items for the Tag Types

3 Results

3.1 Participants

Launched from 09/16/2013 to 05/30/2014, 499 questionnaire responses were received, and of which 444 questionnaires were completed. The following analysis is based on these 444 questionnaires. Table 2 summarizes the demographic information of the participants.
### Demographic Information of the Participants

<table>
<thead>
<tr>
<th>Demographic Information</th>
<th>Options</th>
<th>N/%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gender</td>
<td>Male</td>
<td>387 (87.2%)</td>
</tr>
<tr>
<td></td>
<td>Female</td>
<td>57 (12.8%)</td>
</tr>
<tr>
<td>Age</td>
<td>21-30</td>
<td>211 (47.5%)</td>
</tr>
<tr>
<td></td>
<td>31-40</td>
<td>172 (38.7%)</td>
</tr>
<tr>
<td></td>
<td>41-50</td>
<td>51 (11.5%)</td>
</tr>
<tr>
<td></td>
<td>Above 50</td>
<td>10 (2.3%)</td>
</tr>
<tr>
<td>Disciplines</td>
<td>Engineering and technology science</td>
<td>178 (40.1%)</td>
</tr>
<tr>
<td></td>
<td>Natural science</td>
<td>178 (40.1%)</td>
</tr>
<tr>
<td></td>
<td>Humanities and Social Sciences</td>
<td>43 (9.7%)</td>
</tr>
<tr>
<td></td>
<td>Medical Science</td>
<td>28 (6.3%)</td>
</tr>
<tr>
<td></td>
<td>Agricultural science</td>
<td>17 (3.8%)</td>
</tr>
<tr>
<td>Education</td>
<td>Master candidate</td>
<td>55 (12.4%)</td>
</tr>
<tr>
<td></td>
<td>Master</td>
<td>94 (21.2%)</td>
</tr>
<tr>
<td></td>
<td>Doctor candidate</td>
<td>102 (23.0%)</td>
</tr>
<tr>
<td></td>
<td>Doctor</td>
<td>193 (43.5%)</td>
</tr>
</tbody>
</table>

Table 2: Demographic Information of the Participants

### 3.2 Usage of tag types

Table 3 shows the 444 participants’ usage of tag types for academic blogs. The top three tag types all belong to the content-based category, which are the tags (1) existing in the blog’s content, (2) existing in the blog’s title and (3) describing the categories of the blog. Meanwhile, the context-based tags that annotate the blog’s author, published time and published location are less used.

<table>
<thead>
<tr>
<th>Tag types</th>
<th>Items</th>
<th>1(low)</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5(high)</th>
<th>M</th>
</tr>
</thead>
<tbody>
<tr>
<td>Content-based tags</td>
<td>T1.1</td>
<td>8.6%</td>
<td>14.9%</td>
<td>21.4%</td>
<td>32.9%</td>
<td>22.3%</td>
<td>3.45</td>
</tr>
<tr>
<td></td>
<td>T1.2</td>
<td>18.9%</td>
<td>28.8%</td>
<td>32.2%</td>
<td>16.2%</td>
<td>3.8%</td>
<td>2.57</td>
</tr>
<tr>
<td></td>
<td>T1.3</td>
<td>5.9%</td>
<td>14.6%</td>
<td>31.5%</td>
<td>35.6%</td>
<td>12.4%</td>
<td>3.34</td>
</tr>
<tr>
<td></td>
<td>T1.4</td>
<td>7.7%</td>
<td>16.4%</td>
<td>34.2%</td>
<td>33.3%</td>
<td>8.3%</td>
<td>3.18</td>
</tr>
<tr>
<td>Context-based tags</td>
<td>T2.1</td>
<td>35.6%</td>
<td>27.7%</td>
<td>23.0%</td>
<td>9.9%</td>
<td>3.8%</td>
<td>2.19</td>
</tr>
<tr>
<td></td>
<td>T2.2</td>
<td>44.1%</td>
<td>22.7%</td>
<td>20.3%</td>
<td>9.2%</td>
<td>3.6%</td>
<td>2.05</td>
</tr>
<tr>
<td></td>
<td>T2.3</td>
<td>43.9%</td>
<td>27.3%</td>
<td>17.3%</td>
<td>9.7%</td>
<td>1.8%</td>
<td>1.98</td>
</tr>
<tr>
<td></td>
<td>T2.4</td>
<td>18%</td>
<td>20.0%</td>
<td>22.7%</td>
<td>23.9%</td>
<td>15.3%</td>
<td>2.98</td>
</tr>
<tr>
<td>Subjective tags</td>
<td>T3.1</td>
<td>20.3%</td>
<td>22.1%</td>
<td>30.0%</td>
<td>21.6%</td>
<td>6.1%</td>
<td>2.71</td>
</tr>
<tr>
<td>Organizational tags</td>
<td>T4.1</td>
<td>12.4%</td>
<td>20.0%</td>
<td>30.4%</td>
<td>28.6%</td>
<td>8.6%</td>
<td>3.01</td>
</tr>
</tbody>
</table>

Table 3: The Usage of Tag Types

### 3.3 Comparison the usage of tag types on different bloggers

The association between the blogger’s demographic information and the usages of the tag types is reported in this section.

To find the gender’s effect, ANOVA test was conducted. We find that female bloggers have the significantly stronger will to use the self-organizational tags to annotate academic blogs (F=7.098, p<.01). But there is no significant difference on the usage of other tag types between genders.

ANOVA test also shows that the younger bloggers with under 41 years old, in comparison with older bloggers, have the significantly higher preference to use the tags that describe the blog topic but do not exist in the blog’s content (F=8.782, p<.01). Meanwhile, these younger bloggers also have a significantly higher probability to use the tags for annotating the blog’s published time (F=7.652, p<.01). Other types show no significant difference among age groups.
Furthermore, through conducting ANOVA test, we find that bloggers with master degree or being master candidates, in comparison to those with doctor degree or being doctor candidates, are more likely to use subjective tags ($F=10.009$, $p<.01$) and the tags for annotating the blog’s published time ($F=9.268$, $p<.01$). No other statistical difference was found across user’s education levels. Similarly, ANOVA test shows no significant difference across the user’s disciplines.

4 Discussion and Conclusion

In this work, we explored what types of tags bloggers use for annotating academic blogs. To our knowledge, this is the first study that detects the tag types employed in academic blog platform. The results imply that tag types that describe the blog content, including those existing in the content and the title of a blog are more preferred by the blogger to annotate the academic blog. Tags that annotate the blog context information, such as the blog’s author, published time and published location, are less used. Based on the tag types’ preference for academic blogs, the academic blog website can improve the bloggers’ satisfactory with the tags recommendation.

This study also give us the evidence that not all tags are equally useful for different users. For example, the female bloggers are more willing to use self-organizational tags. The bloggers under 41 years old have the significant higher preference to use the tags that describe the blog topic but do not exist in the blog’s content, and the tags for annotating the blog’s published time. These findings imply that the tag recommendation should distinguish the bloggers demographic information to be conducted.

Future work includes studying the other factors that influence the academic blog tag usage, such as the time effect and the platform effect. Finally, we will try to apply our findings to the academic tagging systems to provide the bloggers with better service.

References


141
Automatic Course Website Discovery from Search Engine Results

Rui Meng¹, Zexin Zhao¹, Yu Chi¹ and Daqing He¹
¹School of Information Sciences, University of Pittsburgh, USA

Abstract

With the rapid development of Internet Technology, the forms of education have been undergoing drastic changes. Instructors are used to posting teaching materials on course websites and setting them publicly accessible. Thus large amounts of course resources have been well organized and shared, which also provide possibilities for building knowledge graphs for a specific domain. However, so far no specific method has been developed for collecting online course resources. In this paper, we propose a method to identify course websites by filtering search results from a general search engine. Experiment results show that the proposed method could achieve good performances on both within-domain and cross-domain tasks, which lays a solid foundation for further work on mining and integrating the online educational resources.

Keywords: course website; website identification; search results filtering; web mining; resource discovery; online educational resource


Copyright: Copyright is held by the authors.

1 Introduction

The dramatic increase of online educational resources has provided great opportunities to innovate new ways for learning and teaching. For example, students can enhance their understanding of course knowledge after class by reading the online resources from other similar courses. When designing a new course, an instructor can also come out useful ideas if he/she is referred to the course syllabus from other instructors. More importantly, the contributors of online courses, most of whom are instructors, are usually experts of one domain. When presenting a certain knowledge concept, they usually provide various related materials including textbooks, slides, readings, videos along with syllabus to augment students’ comprehensive understanding of the concept (Meng, Han, Huang, He, & Brusilovsky, 2016). As a result, a proper linking of online educational resources is of great important, for both learning and teaching purposes (Huang, Yudelson, Han, He, & Brusilovsky, 2016; Wang et al., 2016; Land & Greene, 2000; Liu, Jiang, & Gao, 2015).

Among the given variety of educational resources, this paper focuses on the online course websites since they provide more comprehensive organizations of knowledge within a domain. To achieve this goal, our first step is to build a data collection for online courses. However, to the best of our knowledge, there is no existing work regarding this topic in academia. Though the general search engines like Google can serve as an effective channel for collecting such resource at a low cost, their results usually consist of too much noise. This motivates us to invent new automated approaches to re-rank or filter the noised return results. Specifically, we first collect a large amount of course web pages from search engine results; then, we identify whether a result page is a course page or not based on classification (Käki & Aula, 2005; Brin & Page, 2012). Our classification approach utilizes both content features and layout features. The content feature attempts to capture the way how the content of a web page is written. We assume a course web page is more likely to include course-dependent words such as syllabus, week, schedule, reading and etc. than other webpages. The layout feature, on the other hand, tries to analyze how a webpage is presented to users. We observe that the course webpages usually contain several tables and listed items.

The remaining of the paper is organized as follows: in Section 2, we present the definition of our task and the feature design. The experiment setup and its corresponding results are provided in Section 3. At last, we provide our conclusion, as well as limitations and future directions.
2 Methodology

2.1 Task Definition

We define the course website as a site that presents detailed content regarding a specific course. The content includes the course description, syllabus, schedule, textbook, reading, and so on. We do not consider a pure course introduction page as a course website in this study since it only provides very limited course content. Specifically, our approach includes two steps for building online course data collection, as listed below.

- Step I: Retrieves a set of course website candidates $C = \{c_1, c_2, c_3, \ldots, c_m\}$ ($m$ is a predefined number indicating the depth of search results we went through) from a search engine using a query $q$. We will explain the ways of choosing query $q$ in Section 3.1.
- Step II: As for each website candidate $c_i$ in $C$, a binary classifier is adopted to determine whether $c_i$ is a course website or not. The classifier is trained on a manually-labeled dataset.

2.2 Feature Design

In the classification step (i.e., Step II), we extract a set of n features $F = f_1, \ldots, f_n$. According to a previous study (Dong, Qi, & Gu, 2005), both structure layout features and content features play important roles for identifying domain-specific websites. Inspired by their work, both of these two features are employed in our study, as listed below.

- For the content feature, we adopted the Bag-of-Word (BOW) feature on the textual content. The BOW feature is a widely-adopted feature for variety of text classification tasks (Sriram, Fuhr, Demir, Ferhatosmanoglu, & Demirbas, 2010; Lodhi, Saunders, Shawe-Taylor, Cristianini, & Watkins, 2002). Note that for each feature (one word), we adopt the TF-IDF (Term Frequency - Inverse Document Frequency) to indicate the feature importance, expecting that it helps remove the domain-dependent features and makes our method be generalized on cross-domain scenarios.
- We expect the structural feature would provide additional information for classification. For example, in the context of course website, HTML tags like $<table>$ and $<li>$ appear frequently in syllabus pages.

3 Experiment

3.1 Data Collection

Four courses were selected in this paper: Database, Java Programming Language, Information Retrieval, and Human Computer Interaction. These four courses cover important topics in Information Science domain, which widely-recognized as important research topics and taught in most iSchools. The first two courses are relatively basic and the latter two are advanced topics.

We collected 200 search results for each course from Google search (800 website pages in total). We examine three different ways of composing search queries: CourseName, CourseName + course and CourseName + syllabus, and finally choose CourseName + syllabus because of its relative good performance of retrieving relevant results from Google.

To train the classifier, we need a manually-labeled ground truth. This was obtained from manual labeling of two authors. The labeling on Java Programming Language is used for examining the agreement, in which we find a Kappa coefficient of 0.7465, indicating a fairly high agreement. The other three courses are then labeled separately.

Eventually, 800 course website candidates are labeled and the descriptive statistics is summarized in Table 1. Figure 1 shows the distribution of course websites in top 100 search results, from which we can see that the course websites only represent a small share of all results, and may occur in all different ranges of ranking positions. This further indicates the importance of putting extra efforts on result filtering.

The noise pages of each course are quite different. In Human-Computer Interaction, a large share of noises are about HCI degree introduction instead of course page. As for Java Programming Language, most noises are company advertisements and tutorials, but some companies also offer high-quality course websites.
<table>
<thead>
<tr>
<th>Course Name</th>
<th>Total Number</th>
<th>Number of Course Websites</th>
</tr>
</thead>
<tbody>
<tr>
<td>Database (DB)</td>
<td>200</td>
<td>44</td>
</tr>
<tr>
<td>Human Computer Interaction (HCI)</td>
<td>200</td>
<td>50</td>
</tr>
<tr>
<td>Information Retrieval (IR)</td>
<td>200</td>
<td>65</td>
</tr>
<tr>
<td>Java Programming Language (Java)</td>
<td>200</td>
<td>51</td>
</tr>
</tbody>
</table>

Table 1: The Statistics of labeled ground truth

![Course website distribution in top 100 Google search results. X-axis: the ranking positions of course websites on the Google search result page. Y-axis: number of course websites.](image)

3.2 Experiment Setting

The classification performance reported in this paper was based on the average of 5-fold cross-validation. Training and testing data was randomly generated on the dataset by 50 times. The Support Vector Machine with L2 regularization was employed (Yu, Ho, Juan, & Lin, 2013). We mainly report the F-1 score of positive class (positive indicates that a given website is a course website). And the reported performance is based on the average of 50 testing runs on optimal classifier parameters.

As mentioned above, we extracted both BOW and HTML structure features. Since a pure HTML structure feature based classifier does not perform well on most of the data collection. Here, we report the performances based on two feature groups: BOW and BOW + HTML. In the following sections, we consider two tasks. First, we built one classifier for each course, i.e., we split the dataset for one course into training and testing. Therefore, we can learn four classifiers and predict each course separately.

3.3 Within-course Website Identification

The performance of within-course prediction can be illustrated in Figure 2. The classifier achieves a fairly good performance on all the courses, especially on the course Java Programming Language. This might because of the relatively clear difference between its noises and the real course websites. It is easier for the classifier to differentiate websites of interest from others. Affected by the degree introduction websites under the similar titles, the Human-Computer Interaction fails in making a good result. As for the feature comparison, by adding additional HTML tags into the features, results of Information Retrieval and Java Programming Language improve significantly than the results based on only textual content features. Surprisingly, the performance of Database drops drastically by adding HTML features, and details show that the accuracy goes from 0.7428 up to 0.7928 but the recall goes from 0.9030 down to 0.7394. The HTML features help rule out irrelevant results at the cost of decreasing recall.
3.4 Cross-course Website Identification

Besides the regular within-course scenario, we are highly interested in the performance on the cross-course task. As we cannot label data for every course, whether our model can generalize well on other courses becomes a critical issue. In this set of experiments, we fix the same test sets which are used in within-course evaluations. This means we can compare the performance of the cross-course task to the within-course. But here we change the training sets to other courses to implement the transfer learning. Here we only report the results based on Whole HTML feature set. The experiment on another feature set shows similar results.

The experiment results are shown in Figure 3. In each course cluster, By Self is the performance same to 3.3, which is based on 5-fold cross-validation on one course. The By Courserename means the model is trained on the whole dataset of corresponding course, and By All Three means model trained on all three courses except the tested one. We can see that for most cases the model can be applied to another domain without losing to much performance, which indicates the generalization ability of our features. The high score of By All Three shows the benefits from abundant data. However the Java cluster presents a different pattern. Due to the characteristic of Java websites (unique content and many course websites maintained by businesses), the By All Three achieves a similar score to By Java.
4 Conclusion

The numerous of course websites provide massive amounts of valuable educational resources, but it remains challenging to differentiate the course websites from others. In order to solve the problem of collecting course websites, we propose a method to identify course website by filtering search results from a general search engine. Experiment results show that the proposed method could achieve good performances on both within-domain and cross-domain tasks. The robust performance and the strong generalization ability indicate the effectiveness of our proposed methods. There are several interesting directions to explore in the future, including exploring our approach on a larger scale and looking for effective methods for capturing domain-independent features. In addition, we are interested in working on potential applications such as online educational resources integration and recommendation.

References


Analyzing Figures of Brain Images from Alzheimer’s Disease Papers

Satoshi Tsutsui¹, Guilin Meng², Xiaohui Yao³, David Crandall¹, Ying Ding¹,4,5

¹School of Informatics and Computing, Indiana University, Indiana, USA
²Department of Neurology, Tenth People’s Hospital, Tongji University, Shanghai, China
³Department of Neurology, Indiana University Center for Neuroimaging, Indiana, USA
⁴Library, Tongji University, Shanghai, China
⁵School of Information Management, Wuhan University, Wuhan, China

Abstract

Which papers focusing on Alzheimer’s disease (AD) include MRI scans of human brains? These images play an important role in clinical detection of AD, but finding them currently requires manual inspection of papers after a keyword search. In order to provide AD researchers with a more efficient way of finding relevant papers, here we focus on three preliminary problems involving automatically identifying figures containing brain images, and solve them as automatic image classification tasks. This is a first step towards efficiently allowing AD researchers to retrieve papers containing a particular type of brain image (e.g. of a patient). We report preliminary results from a larger project, in collaboration with AD researchers.

Keywords: Alzheimer’s Disease; Figure Mining; Viziometrics


Copyright: Copyright is held by the authors.

Contact: stsutsui@indiana.edu.

Figure 1: In this paper, we show how to extract figures from PDFs of AD papers, and answer three questions:
Is it a brain image or not? What kind of imaging technology is used? Is it a human brain or not?

1 Introduction

The exponential rate of growth of biomedical literature (Larsen & Von Ins, 2010) makes it hard for researchers to follow the latest progress even in their specialized domain. Brain imaging researchers, for example, must manually inspect large numbers of papers, even after performing a text-based search, to find those containing brain images. With that issue in mind, we currently collaborate with researchers on Alzheimer’s disease (AD), aiming to contribute to AD research from computer and information science perspective.

Brain imaging plays a key role for clinical detection of AD. The most widely used biomarkers of AD include focal atrophy and metabolic dysfunction, which can be detected by brain imaging. In fact, a definite diagnosis of AD requires the examination of brain tissue, which is usually performed after the death of a patient (Ballard et al., 2011). This fact shows that brain imaging, imaging, a way of inspecting a brain without surgery, plays a key role for clinical detection of AD.
In this paper, we focus on analyzing brain images in AD papers. Currently, the purpose is to assist AD researchers to find papers containing a brain image of their interest, but it also has a potential for assisting clinical diagnosis because we could retrieve papers containing a brain image that is similar to that of a patient of interest.

This paper reports preliminary results from a larger project collaborating with AD researchers. We investigate figures of brain images, and identify three specific problems to solve (Section 3). Then we regard these as image classification problems, and approach them using computer vision techniques (Section 4). We also discuss several future challenges (Section 5).

2 Related Work

Analyzing figures in scientific papers has recently become an interesting area of research. Several works (Choudhury, Mitra, & Giles, 2015; Clark & Divvala, 2015; Kuhn, Luong, & Krauthammer, 2012; Clark & Divvala, 2016) extract figures directly from PDFs. In addition to extraction, there are lines of works (Savva et al., 2011; Choudhury & Giles, 2015; Siegel, Horvitz, Levin, Divvala, & Farhadi, 2016) that classify figures. However, they focus on classification into general categories such as diagrams, photos, tables or data plots. In contrast, we are interested particularly in brain imaging in the AD domain. In addition to focusing on individual figures, Viziometrics (Lee, West, & Howe, 2016a, 2016b) focus on analysis of figures from collections of papers. It has recently been proposed emphasizing the visual content analysis of papers in contrast to bibliometrics, which has been focused only on textual content of papers.

3 Problem Specifications

We consulted with an AD expert to investigate what kinds of information are important to them when looking at figures of brain images. We manually checked 30 random AD papers, examined the figures in them, and defined the following problems as preliminary steps:

1. Does the figure contain an image of a brain?
2. If so, which imaging methods was used to capture it? (e.g. Magnetic Resonance Imaging (MRI), Computed Tomography (CT), Positron Emission Tomography (PET) and Single Photon Emission Computed Tomography (SPECT) )
3. Is it a human brain or not (e.g. a rat)?

4 Computer Vision Approach

The three questions above can be posed in terms of image classification, which is well-studied in computer vision. The corresponding classification problems are:

Task 1 Given a figure, classify whether it contains brain images or not.
Task 2 Given a figure containing brain images, classify into four imaging types: MRI, CT, PET, SPECT
Task 3 Given a figure containing brain images, classify whether it is a human brain or not.

It is possible to encounter figures that have multi-labels, but in this preliminary study, we excluded such cases.

4.1 Dataset Preparation

The state-of-the-art image classification approach is to use deep neural networks (LeCun, Bengio, & Hinton, 2015; Razavian, Azizpour, Sullivan, & Carlsson, 2014; He, Zhang, Ren, & Sun, 2015). Hence we need to prepare training data. We manually downloaded PDF files of 100 articles via PubMed search. We first searched with the keywords Alzheimer and Brain Imaging, and randomly downloaded 50 articles which we expected to contain a variety of brain images. Then, in order to get negative examples (i.e. figures not containing brain images), we also downloaded another 50 articles by just searching Alzheimer.
After obtaining 100 PDFs, we extracted 302 figures using PDFFigures 2.0 (Clark & Divvala, 2016). We manually selected 66 figures containing brain images, for use as positive examples for task 1. For negative examples, we randomly extracted 66 figures from the remaining ones that do not contain brain images. This yielded 132 figures in total. Some examples are shown in Figure 2 and 3.

For the 66 figures containing brain images, we investigated their types of brain images. In fact, the figures are often composed of multiple brain images of different types, and even non-brain imaging figures such as plots. We first manually (multi-)labeled each figure into MRI, PET, CT, and SPECT. This labeling requires some expertise, and took a few hours for an AD expert because sometimes the images are post-processed (e.g. coloring or annotation), which makes it hard to confidently distinguish the imaging method and requires the expert to read the contents of the paper. This labeling resulted in the data for task 2 (38 MRI images, 27 PET images, 4 CT images, and 4 SPECT images). When training a classifier, we only used images having a single label. Also, we annotated whether the brain is human or not, forming the dataset for task 3. Of the 66 brain images, 12 figures are non-human brains. We did not find figures containing brain images of both humans and animals.

When training classifiers, we use equal numbers of data for each label to avoid biasing the classifiers. In other words, we limited the training data by the least frequent label. As a result, we only have 132 images for task 1, 16 for task 2, and 24 for task 3. The dataset is available on the web 1.

4.2 Figure Classification Approach

The state-of-the-art approach in image classification uses deep learning (LeCun et al., 2015), specifically convolutional neural networks. However, it requires much more training instances than we collected if we train from scratch, so we used a transfer learning approach (Razavian et al., 2014). We used a pre-trained 50 layer deep residual network (ResNet50) (He et al., 2015) on 1.2 million images from ImageNet (Russakovsky et al., 2015), and re-trained only the last classification layer. We note that it might be possible to use more complicated approach to obtain a better performance, but the focus on this preliminary paper is not optimizing the performance.

4.3 Classification Results

Since we only have a limited amount of data (especially for task 2), we evaluated the accuracy using 4-fold cross validation. Recognizing brain images works relatively well (93 % and 83 % accuracy on task 1 and 3, respectively, compared to a 50% baseline of random guessing). However, recognizing imaging type does not work as well (56 % accuracy on task 2), although this is a more difficult task (25% random baseline). In order to understand the difficulties, we compared an aggregated confusion matrix shown in Table 1. It indicates that CT is always classified correctly, however, PET and SPECT are difficult to distinguish. We note that the size of data is a limitation of our study, especially for task 2 with only 16 examples. Hence it might not be valid to generalize our results. We discuss this limitation in the next section.

<table>
<thead>
<tr>
<th>Truth</th>
<th>Predicted</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>MRI</td>
</tr>
<tr>
<td>MRI</td>
<td>1</td>
</tr>
<tr>
<td>PET</td>
<td>0</td>
</tr>
<tr>
<td>CT</td>
<td>0</td>
</tr>
<tr>
<td>SPECT</td>
<td>0</td>
</tr>
</tbody>
</table>

Table 1: Confusion matrix for brain image type recognition (task 2)

1http://homes.soic.indiana.edu/stsutsui/ad_brain_images/
5 Challenges and Future Work

We noticed two challenges. First, the cost of collecting training data is quite high. State-of-the-art computer vision techniques require large amount of data. Crowdsourcing (e.g. Amazon Mechanical Turk) is often used in computer vision research to annotate images with ground truth labels, but annotating medical figures requires domain expertise, so crowdsourcing is difficult to use. It is not practical to ask many AD researchers to annotate. A possible solution is to annotate examples in a semi-automatic manner using some heuristics (Mintz, Bills, Snow, & Jurafsky, 2009). It might be possible to use figure captions to automatically; however, these annotations always have noise, so we need techniques with noise robustness.

Another challenge is to localize brain images within a figure because many figures have multiple types of brain images. In computer vision, techniques based on Region-based Convolutional Neural Networks (R-CNN) (Girshick, Donahue, Darrell, & Malik, 2014) are the state-of-the-art. However, they again require large amounts of large training data. Some previous work in figure mining (Siegel et al., 2016; Lee & Howe, 2015) parse figures into sub-figures, but they assume plots or charts with a white background, so these techniques cannot be directly applied to brain images whose background is sometimes black (e.g. right bottom of Figure 2).

6 Conclusion

This study presents preliminary results from a project collaborating with AD researchers. In this paper, we focused on brain image analysis from AD papers. As an initial step, we identified three specific problems, and discussed future challenges. However, of course, many more problems remain to be solved other than these three, such as recognizing which region of a brain is highlighted in a brain image, or segmenting brain images out of complex multi-part figures.
Figure 3: Example figures not containing brain images. (Figure sources PMID: 11906265, 27355515)

References


Kuhn, T., Luong, T., & Krauthammer, M. (2012). Finding and accessing diagrams in biomedical publications. In *Association for the Advancement of Artificial Intelligence (AAAI)*.

Larsen, P. O., & Von Ins, M. (2010). The rate of growth in scientific publication and the decline in coverage provided by Science Citation Index. *Scientometrics, 84*(3), 575–603.


A Twitter-based Recommendation System for MOOCs based on Spatiotemporal Event Detection

Yuanyuan Wang¹, Naoki Maruyama², Goki Yasui², Yukiko Kawai², Toyokazu Akiyama²
¹Yamaguchi University, Japan ²Kyoto Sangyo University, Japan

Abstract
Nowadays, students utilize MOOCs (e.g., Coursera, edX) and SNS services (e.g., LINE, Twitter, Facebook, Tumblr) in courses for learning. This paper presents a Twitter-based recommendation system to search and communication, and it is associated with a web page by detecting spatiotemporal events such as opinions, questions, or impressions about courses on Twitter. Through it, users can grasp popular courses or avoid crowded courses referring to time periods while they browse any web pages. Moreover, the system also enables users to communicate with others browsing the similar pages or users' locations about the similar pages. For this, the system extracts relevance between different pages by detecting tweets of each page in each time period with machine learning algorithms and the number of unique Twitter users. Thus, the system presents a ranking of recommended pages, a tag cloud of tweets and a list of tweets which are related to recommended pages to help users obtain the latest information about recommended pages. In this paper, we propose that students utilize the system to enhance interaction among and with others in actual classrooms. This promises to enlarge the learning effects of students and improve the student collaboration.

Keywords: Twitter; MOOC; spatiotemporal events; recommendation; communication


Copyright: Copyright is held by the authors.

Acknowledgements: This work was partially supported by MIC SCOPE (150201013), and JSPS KAKENHI Grant Numbers 15K00162, 16H01722.

Contact: y.wang@yamaguchi-u.ac.jp, i1458076@cse.kyoto-su.ac.jp, gxyasui@gmail.com, {kawai, akiyama}@cc.kyoto-su.ac.jp

1 Introduction

With the proliferation of massive open online courses (MOOCs) in e-learning, enormous amounts of various courses from top universities and institutions are freely shared on the Web, such as Coursera¹ and edX². In order to promote the learning effects, students utilize not only MOOCs but also SNS services in courses such as LINE³, Twitter⁴, Facebook⁵, and Tumblr⁶. In the real world, students participated in classrooms often share personal opinions, questions, or impressions about courses through location-based SNS services. Several studies have examined the use of SNS in MOOCs (Van Treeck & Ebner, 2013; Liu, McKelroy, Kang, Harron, & Liu, 2016), the findings indicated that the SNS can augment the learning experience by providing an environment to enhance communications. In addition, the SNS provides a space to post personal feelings or reflections of learning in an informal and quick manner. However, there is not enough to provide a summary of SNS data or recommendation for MOOCs based on the spatiotemporal information of SNS data, it is difficult for aggregating the students’ opinions about courses. Therefore, it is important to detect the spatiotemporal events such as opinions, questions, or impressions about courses from SNS data, to help users to better understand the courses while they browse course web pages, and to support communications between users with others in classrooms.

¹https://www.coursera.org/
²https://www.edx.org/
³https://line.me/en/
⁴https://twitter.com/
⁵https://www.facebook.com/
⁶https://www.tumblr.com
In this paper, we have developed a novel system for recommending relevant course web pages associated with the current browsing page based on detected spatiotemporal events. The system can present a quick overview of the latest information about each course web page by extracting feature words from tweets and each course web page. In order to recommend courses, the system can detect relevant course web pages by calculating the similarities between feature words from tweets related to course web pages.

The system has three main features:

1. **Tweet classification.** Detecting related tweets about web pages in different time frames of a day based on the content of tweets and web pages by adopting machine learning algorithms (e.g., $k$-NN, SVM).

2. **Course recommendation.** Ranking course web pages by calculating the similarities between TF-based feature words from tweets about course web pages and counting the number of unique Twitter users who attend in each actual classroom of each page.

3. **User communication.** Utilizing a communication function as TWinChat (Wang et al., 2014) supports web users to chat with Twitter users who attend in actual classrooms of the similar pages.

As depicted in Figure 1, we propose use cases in MOOCs that illustrate how the system can enhance the interaction among and with students in actual classrooms. A ranking of recommended courses to help a user grasp popular courses or avoid crowded courses referring to time periods while the user browses a course page of a MOOC site. Moreover, a tag cloud of tweets and a list of tweets to help the user gain a quick overview and details of the latest information about recommended courses.
2 Twitter-based Recommendation System by Spatio-Temporal Analysis

To use this system, which is on the basis of existing Web services, Twitter users are required to follow an account\(^7\) of our system, as followers of our service, and web users are required to simply install a toolbar (a browser plug-in). Once a user browses a web page with the installed toolbar, the system records the information into a database, which is used for mapping tweets to the web page based on feature words (classroom names) detected from the tweets and the web page, and classifying the tweets of the web page in different time frames of a day by adopting machine learning algorithms. In our system, anonymous of all messages (tweets) can be maintained through Twitter services and a WebSocket server\(^8\). In our previous work, we acquire a total amount \(n\) of tweets based on a given location, and calculate the average frequency of each word \(i\) that appears in each tweet \(t\) with a standard sigmoid function \(1/(1 + e^{-x})\) for weighting each word \(i\) related to location names by the following formulae.

\[
\sum_{i=1}^{m} \left( x_i \times \frac{1}{1 + e^{-x_i}} \right) \times \frac{1}{m}
\]

\[
x_i = \frac{\# \text{tweets with } i}{n}
\]

The flow of our system is described as follows:

- After a user selects a web page to browse, the system then returns a ranking of recommended pages, a tag cloud of tweets, and a list of tweets are associated with the web page (see Figure 1).
- When the user checks a time period, the ranking of recommended pages, the tag cloud and the list of tweets can be changed by the user’s specified time period.
- When the user checks a recommended page, the tag cloud and the list of tweets can be changed by the user’s specified recommended page.
- When the user clicks a tag, the system then presents a list of tweets about it, in which most related tweets are presented.
- When the user sends a message through a chat box of our system, the system presents it in the tweet list, Twitter users or other users can receive it.
- When a Twitter user replies the message of the user through Twitter service; the system presents the reply relating to the web page in the tweet list.

3 Use Cases of Utilizing Twitter-based Recommendation System in MOOCs

3.1 Individual Learning

For an individual learning assignment accomplished by searching courses through the MOOC site, there are three interesting use cases.

a. Self-learners can choose courses according to study time based on current situations of actual classrooms.

b. Beginning students can make course plans referring to others’ learning experiences on relevant courses.

c. Students are motivated in friendly competition by grasping others’ learning progresses in similar courses.

\(^7\)https://Twitter.com/@RtQAService

\(^8\)http://gihyo.jp/dev/feature/01/websocket/0001
3.2 Collaborative Learning

For a collaborative learning assignment accomplished by searching courses through the MOOC site, there are three interesting use cases.

a. Students can share the attending courses through Twitter-based recommendation system.

b. Experienced students can immediately teach or reply beginning students’ questions about similar courses.

c. Online students browsing the course web page only, can easily chat with others in actual classrooms.

4 Conclusion and Future Work

In this paper, we propose that online students utilize the Twitter-based recommendation system for MOOCs to enhance the interaction among and with students in actual classrooms. Through it, users can grasp popular courses or avoid crowded courses referring to their study time while they browse course pages. We also stated some use cases of individual learning and collaborative learning through our system. These promise to enlarge the learning effects of students and improve student collaboration.

For future work, we need to improve our tweet classification method by considering the locations of tweets that are not actual classrooms. In addition, we will enhance our system to recommend not only courses but also appropriate users. Furthermore, we plan to expand our system to support not only web pages, but also other learning contents (e.g., lecture slides, lecture videos, etc.).

References


Automatic ICD Code Assignment to Medical Text with Semantic Relational Tuples

Sanqiang Zhao¹, Daqing He¹, Danchen Zhang¹, Lei Li², Rui Meng¹
¹University of Pittsburgh ²Nanjing University of Science and Technology

Abstract
Mining the Electronic Medical Record (EMR henceforth) is growing in popularity but still lacks good methods for better understanding the text in EMR. One important task is assigning proper International Classification of Diseases (ICD henceforth, which is the code schema for EMR) code based on the narrative text of EMR document. For the task, we propose an automatic feature extraction method by means of capturing semantic relational tuples. We proved the semantic relational tuple is able to capture information at semantic level and it contribute to ICD-9 classification task in two aspects, negation identification and feature generation.

Keywords: ICD-9 Classification; text mining; EHR mining


Copyright: Copyright is held by the authors.

Contact: saz31@pitt.edu, dah44@pitt.edu, daz45@pitt.edu, lileiwelldone@gmail.com, rum20@pitt.edu

1 Introduction

Electronic Medical Record (EMR) has become widely available in the medical domain, but mining such information with state-of-the-art natural language processing and machine-learning algorithms is still lacking. One important task is assigning proper International Classification of Diseases (ICD) code based on the narrative text in EMR.

ICD code is a standard classification schema and provides standard diagnoses and procedural tools for the collecting and reporting health information. Specifically, this paper focuses on the ninth revision of ICD code, clinical modification (ICD-9-CM henceforth).

Generally, ICD-9-CM is manually decided by human based on the narrative text of the EMR document. Concretely, nurses interview patients and write down complaints. Then a registration clerk or a trained human coder assigns the ICD-9-CM code(Tsui, Wagner, Dato, & Chang, 2001) by viewing the entire narrative text of the EMR document. However, due to the large number of patients and complaints, manually assignment of the ICD-9-CM code is a time-consuming and error-prone task. Therefore, automatic assignment can help in two aspects: (1) reducing error by supporting decision-making for human coders, and (2) speeding up code assignment and reducing cost(Ira Goldstein & Anna Arzumtsyan, 2007).

The goal of our study is to automatically assign an ICD-9-CM code to the given EMR document. We adopted a standard supervised machine-learning approach, which extracts a bunch of features from the narrative text and learns the weight of each feature from the training data set. Previous studies have examined the performance on several simple feature-extraction methods, such as bag-of-word and rule-based features. However, as the medical text in EMR is relatively simple and straightforward, it is possible to better understand the text by applying a state-of-the-art natural language technique.

Specifically, we propose an automatic feature extraction method for capturing semantic relational tuples, like <rib,pain> and <cough, persistent>. This kind of tuples are widely distributed in narrative text of EMR documents and can be used to identify their ICD-9-CM codes effectively. Compared to features in prior studies, features from our approach delves into the semantic level of medical text.

2 Approach for Semantic Tuple Extraction

Though previous studies (Crammer, Dredze, Ganchev, Talukdar, & Carroll, 2007; Farkas & Szarvas, 2008; Ira Goldstein & Anna Arzumtsyan, 2007) have demonstrated the effectiveness of different features for the
ICD-9-CM classification task, most of these features require predefined extracting rules and also unable to
discover enough semantic features. Instead, we need a method for extracting more meaningful features and
reducing manual effort as much as possible. Kavuluru et al. (Kavuluru, Rios, & Lu, 2015; Kavuluru, Han, &
Harris, 2013) developed an automatic approach, in which they applied the part-of-speech tagger and extracted
word sequence following certain part-of-speech tag patterns, such as adjective + noun, noun + noun, etc.,
However, the part-of-speech tagger cannot extract syntactic relationships and cannot handle modification
relationships from texts.

Inspired by the idea of extracting semantic relational tuples for the product review (Qiu, Liu, Bu, & Chen, 2011),
we extract the <aspect, status> tuples, in which an aspect can be a symptom, procedure, etc., and a status can be
the modifier for this aspect, such as <cough, recurrent> (cough being one type of symptom while recurrent
being status), <rib, pain> (rib being body part while pain being status).

Here, we discuss preliminary reasons why our method is supposed to be useful.

- Narrative text in EMR documents is simple and straightforward, few implicit opinions or rhetoric in
  it. Therefore, the <aspect, status> tuple is reachable and effective expression for understanding the
  narrative text.

- Narrative text in EMR documents is formal and the majority of expressions is standard.

3 Experiment

3.1 Experiment Setup

We evaluated effectiveness of our semantic relational tuples through classification tasks. We utilize Logistic
Regression as our classifier on the CMC data set, which contains 978 radiological reports. Due to data set is
skewed (data sizes for different codes are imbalanced), previous researchers and us adopt the micro-averaged
f1-score as the metric.

The 10-fold cross-validation was employed in our experiment. Also, in order to reduce the random
effect, we run the 10-fold cross-validation multiple times (100 times in our case) and provide the mean of
micro-averaged f1-scores.

3.2 Identify Negation

Based on prior studies, negation expression, such as “no focal pneumonia“, should not be considered when
assigning ICD code (Ira Goldstein & Anna Arzumtsyan, 2007). Traditional approach for identifying negation
expression is through Negex algorithm, which is rule-based algorithm and incorporates many hand-built rules.
However, our method automatically identify negations without predefined rules but some opinion words, such
as “no”. In this section, we compare classification performance by removing negation detected by different
negation detection methods.

<table>
<thead>
<tr>
<th>Negation Detection Method</th>
<th>Nothing Applied</th>
<th>Negex Algorithm</th>
<th>Our method</th>
</tr>
</thead>
<tbody>
<tr>
<td>f1-score</td>
<td>0.827</td>
<td>0.842</td>
<td>0.844</td>
</tr>
</tbody>
</table>

Table 1: Classification Performance based on different Negation Detection Algorithm

Based on Table 1, both negation detection algorithm would detect negation for improving classification
performance. Our method performs similar as Negex algorithm. However, Negex algorithm benefits from
many hand-built rules and major inaccuracy comes from the lack and conflict of hand-built rules. Our method
keeps away from the limitations.

3.3 Selected Features

By further proving effectiveness of our semantic feature tuple, feature selection is employed. Note that we
have totally 2700 semantic features generated by our approach and we rank them by inverse document
frequency (IDF) (Yang & Pedersen, 1997) and add them into model gradually.
Table 2: Classification Performance by Adding Feature Gradually

Table 2 shows trend of the performance by adding 500, 1000, 1500, 2000, 2500 and 2700 semantic features respectively. Initially the improvement on classification performance is little but last 700 features contributes to 1 percentage of improvement.

4 Conclusion

Our feature extraction approach for capturing semantic relational tuple significantly improves performance of ICD-9-CM code assignment task. The most important reason is our approach delve into semantic level of text. The semantic relational tuple is widely distributed in narrative text of EMR documents and can be used to identify the ICD-9-CM codes effectively. Experiment result shows that our approach could achieve significant improvement on f1-score by more than 1 percentage. Another contribution is our method can also identify negation expression without hand-built rules.

References

Kavuluru, R., Han, S., & Harris, D. (2013). Unsupervised extraction of diagnosis codes from emrs using knowledge-based and extractive text summarization techniques. In Canadian conference on artificial intelligence (pp. 77–88).
Learning Semantic Representation from Restaurant Reviews: A Study of Yelp Dataset

Sanqiang Zhao¹, Shuguang Han¹, Rui Meng¹, Daqing He¹, Danchen Zhang¹
¹University of Pittsburgh

Abstract
Users’ preference such as rating only provides uni-dimension information, but reasons behind users’ preference may be related to various aspects of an item, such as the types, certain attributes. By observing user-generated review always provides such rich information, we proposed an item representation based on review data. This approach supports semantic operation, which could potentially enables more recommendation scenarios. Our experiments further demonstrated that this approach gained much better performance than classical item representation methods.

Keywords: Semantic representation; contextual information modeling; recommender System


Copyright: Copyright is held by the authors.
Contact: saz31@pitt.edu, shh69@pitt.edu, rum20@pitt.edu, dah44@pitt.edu, yuc73@pitt.edu, daz45@pitt.edu

1 Introduction

Recommender systems have been widely developed in recent years for eliminating the problem of information overload and providing personalized recommendations of items (e.g., books, accommodations, musics and news articles) (Koren, Bell, Volinsky, et al., 2009). User and Item (the to-be-recommended products) are two important concepts in a recommender system. Modern recommendation techniques usually attempt to learn users’ tastes based on their item preference histories (often utilizing users’ ratings on items), and make recommendations based on a group of like-minded users who share the preference on the same items. However, the preference information may suffer from data shallowness — users’ preference such as rating to an item only provides a one-dimension assessment of the item, but the reasons behind the users’ preference may be related to any combinations of various aspects of the item, such as the types, certain attributes, its relationship with other items, and even surrounding contextual environment. Therefore, external data resources that provide additional information and full-coverage of preference aspects are favored to enhance the current recommendation performance.

We think that the user-generated review for items is one such resource. Particularly, since the number of users in a commercial recommender system usually far exceeds the number of items, there are more information for items than users. In terms of reviews, we do find that many items consist of tens or hundreds of reviews in many large-scale datasets, and each piece of review further contains tens or hundreds of words. More importantly, content modeling for review texts can provide a better and complementary understanding of why and how users like the items. Therefore, we focus on developing a better content-based modeling approach in this paper. Previous studies have tried both the simple bag-of-word approach (Pazzani & Billsus, 2007) and the latent topic modeling approach that tries to capture the semantic meaning of content (Wang & Blei, 2011). However, the latent semantic analysis for short text such as reviews usually does not yield a good performance (Hong & Davison, 2010), which motivates us to find alternative solutions.

Recent efforts building on top of the modeling of the contextual information for content have achieved substantial performance boosts in many text modeling tasks (Mikolov, Chen, Corrado, & Dean, 2013). One successful example is the development of Word2Vec (Mikolov et al., 2013), which models each word into a low “semantic” dimension, with a dense vector based on the word’s usage contexts. Inspired by this idea, we thought that each item in a recommender system can also be represented by a lower “semantic” dimension based on its context. Here, we use item reviews for representing its context since it directly related to how people assess the item. The new representation of items enables us to perform a set of semantic operations. In our later analysis, for example, we find that by aggregating the representations of a Chinese restaurant in Scottsdale, Arizona and a Casino in Las Vegas, we could locate a restaurant in Las Vegas and serves Chinese...
food. We will provide more detailed explanations about such representation in §2. We believe such operations serve more recommendation scenarios.

To further understand how and whether this modeling approach would work in real-world scenarios, we follow the standard recommendation experiment protocol and attempt to predict user ratings. Specifically, we hold out a certain amount of user ratings for testing, and the rest for training. By comparing our new representation approach with the traditional representation approaches, we can then examine the effectiveness of our model. Again, due to the relatively sparsity of user information in the commercial recommender system, we focused on the item-based recommendation in our paper (Lee & Seung, 2001; Salakhutdinov & Mnih, 2011; Hoyer, 2004).

2 Our Approach: Contextual Representation of Items through Reviews

As mentioned above, the bag-of-word representation often encounters the vocabulary mismatch problem, whereas latent topic modeling can, to some extent, solve this problem by mapping each word into a lower “semantic” dimension. However, latent topic modeling does relatively poorly on handling word contextual information. Therefore, our approach aims to handle both semantic representation and contextual information of an item.

Specifically, our model can be illustrated by Figure 1. Suppose that we have two items: item \( i \) and item \( j \), and each of them receives one piece of review, \( R_i \) and \( R_j \), respectively. \( R_i \) consists of three words \((w_1, w_2, w_3)\), and \( R_j \) contains two words \((w_4, w_5)\). In the beginning, we represent each word using a vector. The vector is obtained through pre-training a large data corpus based on Word2Vec (Mikolov et al., 2013). In this paper, we set the vector dimension as 200. On top of this word vector representation, we then aggregate word vectors to represent an item if a word has appeared in the item review. The aggregation process attempts to search for an optimal item representation (also represented by a vector, with the same dimension as a word vector) so that the item vector becomes more similar to these words that are in reviews and less similar those words that are not in reviews. More details of our approach can be referred to (Dai, Olah, & Le, 2015).

![Figure 1: An illustration of our proposed approach, in which both words and items are represented by vectors. The word vector is pre-trained based on Word2Vec and the item vector is obtained by maximizing the item generation probability \( p(item | R) = \sum_{w \in R} p(item | w) \).](image)

3 Experiment

3.1 Dataset

Our experiment utilizes all of the restaurants and their review information from a large-scale Yelp Challenge Dataset\(^1\). In total, it consists of 24,974 restaurants located in 10 cities and across four countries, and in total

\(^1\)https://www.yelp.com/dataset_challenge
1.3 millions of reviews. Based on these reviews, our model then generates the corresponding representation for each restaurant.

3.2 Understanding Item Representation

To better understand the learned representation for each restaurant, we conduct a simple qualitative analysis as shown in Table 1, in which we try to find the most similar restaurants for each query restaurant. Specifically, at first, we aggregate the item representations of the query restaurants. After that, we compute the cosine similarities of the aggregated representation with each of restaurants in our dataset. The top two or three restaurants are provided. According to Table 1, we find that through our representation, we could easily locate the most similar restaurants with shared attributes.

Querying a Chinese restaurant in Pittsburgh (e.g., china-palace-pittsburgh) enables us to locate other restaurants that also serves Chinese foods (e.g., long-jin-chinese-cuisine-las-vegas-2, yummy-yummy-chinese-restaurant-scottsdale) but in different locations. More interestingly, if we aggregate the representations of a Chinese restaurant in Scottsdale, Arizona (yummy-yummy-chinese-restaurant-scottsdale) with the representation of a Casino in Las Vegas (the-mirage-las-vegas-3), then we could locate a restaurant in Las Vegas and also serves Chinese food (long-jin-chinese-cuisine-las-vegas-2). We believe the above semantic operations would have many potential applications. For instance, a person who travels often can easily locate his desired restaurant in one city by providing his favorite restaurant in a different city.

Table 1: An qualitative analysis of item representation through locating the most similar restaurants

<table>
<thead>
<tr>
<th>Query restaurant</th>
<th>Most similar restaurants</th>
<th>Shared attributes</th>
</tr>
</thead>
</table>
| the-mirage-las-vegas-3 | bellagio-hotel-las-vegas  
new-york-new-york-hotel-casino-las-vegas  
monte-carlo-hotel-and-casino-las-vegas | Located in Las Vegas, Casinos and Hotels |
| valle-luna-phoenix | carlos-o-briens-phoenix-phoenix  
valle-luna-phoenix-2  
las-fonda-del-sol-scottsdale-2 | Mexican food                          |
| which-wich-middleton | which-wich-charlotte  
which-wich-chandler-3 | Sub-branch of which-wich            |
| pkwy-tavern-las-vegas-2 | carolina-ale-house-charlotte  
th-house-of-brews-gilbert  
duckworths-grill-and-taphouse-charlotte-2 | Having sports bar                     |
| china-palace-pittsburgh | long-jin-chinese-cuisine-las-vegas-2  
yummy-yummy-chinese-restaurant-scottsdale | Serving Chinese food                  |
located-inLasVegas and serving Chinese food |

3.3 Review-based Recommendation

To further understand whether this representation works in real-world scenarios, we follow the standard recommendation experiment protocol (Koren et al., 2009) and conduct item-based recommendations to predict users' ratings. At first, the dataset is split into 80/20 (80% for training and 20% for testing) based on the review posting time. Then, based on different approaches for modeling items, we locate the most similar items (top k items, where k is a parameter and we tried 3, 5 and 7) for each of these items. Finally, we predict user rating on each of the similar items as their averaged rating from other users. The prediction performance is evaluated based on the square error of the true rating and the predicted rating, i.e., root-mean-square error (RMSE).

<table>
<thead>
<tr>
<th>RMSE</th>
<th>Bag-of-Words</th>
<th>LDA</th>
<th>Our Model</th>
</tr>
</thead>
<tbody>
<tr>
<td>k = 3</td>
<td>2.841</td>
<td>2.731</td>
<td>1.507</td>
</tr>
<tr>
<td>k = 5</td>
<td>3.122</td>
<td>3.107</td>
<td>1.479</td>
</tr>
<tr>
<td>k = 7</td>
<td>3.308</td>
<td>3.276</td>
<td>1.472</td>
</tr>
</tbody>
</table>

Table 2: Results of the recommendation experiments. A small RMSE indicates a better performance.
Table 2 shows the result of our experiment, where two baselines (Bag-of-Words and LDA) and our approach are applied. They are utilized to find the top k similar items. The bag-of-word baseline computes similarity based on whether two items share exactly the same word, whereas LDA and our model tend to match top k documents based on semantic relations. For LDA, we also use 200 topics to align with our model. Meanwhile, this is also the common setting in many LDA applications (Blei, Ng, & Jordan, 2003).

As shown in Table 2, both LDA and our model outperform the bag-of-word approach, indicating the effectiveness of the semantic modeling of items. Our approach achieves significantly the best performance compared to LDA, denoting the necessity of modeling the semantic information based on contexts. In addition, our model tends to be insensitive to different configurations of k, which is a strong positive message to recommendation community since it is the most difficult parameter in a recommender system.

4 Conclusion

Most of the existing recommendation approaches remain rely on the simple user rating information, whereas such uni-dimension information cannot reveal many aspects of user preferences such as how and why a user prefers one item. Observing that item reviews often provide such rich information, this paper proposed an item representation based on review data, and the start-of-art word text modeling approach based on word contexts (Mikolov et al., 2013). This approach supports semantic operation, which could potentially empowers more recommendation scenarios. Our experiments further demonstrated that this approach gained much better performance than classical item representation methods based on words and semantic topics. We do think that the applications of our approach are not limited to the recommender systems. Similar ideas can be easily applied in any text-based system. We would like to explore more of these applications in the future.

References

SIE: Characterize iSchool Research Territory via Scholarly Data

Xiaozhong Liu¹, Wei Lu², Ying Ding¹, Sam Wunells¹
¹School of Informatics and Computing, Indiana University Bloomington
²School of Information Management, Wuhan University

Abstract
Comparing with other academic units, iSchool research can be more interdisciplinary and dynamic. In this SIE, we will organize a collaborative study with the goal of characterizing iSchool research territory. Meanwhile, Thomas Reuters will support this competition by providing us high quality scholarly data focusing on iSchol researchers. Unlike most prior SIEs, the proposed study will encourage iSchool researchers’ participation. We will organize the presentations and discussions at the conference session as well provide awards to the selected study winners, which will make this event appeal to the audience both with respect to content and format.

Keywords: iSchool; Research; Scholarly Data; Competition


Copyright: Copyright is held by the authors.
Contact: liu237@indiana.edu

1 Introduction

"While each individual iSchool has its own strengths and specializations, together they share a fundamental interest in the relationships between information, people, and technology." This is the classical definition of iSchool from http://ischools.org. It is clear that, comparing with other disciplines and academic units, iSchool research can be more interdisciplinary and dynamic, and we can hardly characterize iSchool research in an easy way.

In this SIE, participants are expected to characterize iSchool research territory in a larger scientific environment by leveraging novel scholarly data. For instance, by using information retrieval, data analysis, Bibliometrics, information visualization, data mining, etc. methods, participants can investigate and propose a number of interesting and novel questions. The proposed ideas are not necessarily restricted to the following exemplar topics:

1. What are the most important research topics of iSchool?
2. What are each individual iSchool’s strengths and specializations?
3. How iSchool researchers collaborate with other scholars? Do iSchool researchers prefer to collaborate with the scholars from other domains?
4. In the past few years, which topics are getting increasingly popular in iSchool and which are decaying?
5. Which iSchool topics are more likely contributed by external research communities?
6. How can one effectively visualize iSchool territory in a larger context?

For all the SIE participants, we will release a novel scholarly dataset from Thomas Reuters Digital Library, e.g., author, publication, citation, and venue, from the most important iSchool venues, i.e., the key conference proceedings and journals in information science. Participants do not necessarily use all the data. However, we expect each participant will use this dataset and the reasonable methodology to address the proposed research question(s).

The proposed SIE is designed to appeal the iSchool researchers (especially PhD students and junior scholars) who have experience and interest in scholarly data analysis. Meanwhile, the expected outcomes of this SIE will generate a more board impact on scholarly data mining and Bibliometrics communities. The organizers will design website and send out call for participation through different channels, e.g., email lists and special invitations.
2 Important Date

For this SIE, organizers will encourage iSchool researchers (especially PhD students and junior scholars) to submit their proposed problems and solutions via the EasyChair system. Meanwhile, we will enable collaborations between the participants. For instance, each participant can publish and advertise his/her proposed idea on the SIE website, and other researchers could join his/her team for collaboration. The selected ideas will be presented in the SIE session during iConference, and the organizers and presenters will lead the discussions in the session. For this SIE, we will follow the following timeline:

- Online registration: TBD
- Dataset release: November 15, 2016
- Paper submission: January 20, 2016
- Announcement of results: February 20, 2017
- Conference presentation: March 24, 2017

3 Awards

For this SIE, in order to encourage participation, we will offer awards to the selected winners. 1st Prize, 2nd Prize and 3rd Prize winners will be selected by the judge committee which consists of 3-5 domain experts from iSchools. Awards for winners will include a certain amount of monetized reward. Moreover, after the SIE, participants may be invited to submit the proposed study to Journal of Data and Information Science journal (one SIE organizer Dr. Ying Ding is one of the journal Co-Editors-in-Chief).

4 Organizing Committee

**Xiaozhong Liu, Indiana University Bloomington:** Dr. Xiaozhong Liu is an Assistant Professor at Department of Information and Library Science, School of Informatics and Computing, Indiana University Bloomington. His research interests include information retrieval, natural language processing, text/graph mining, digital library, metadata, and human computing. His dissertation at Syracuse University explored an innovative ranking method that weighted the retrieved results by leveraging dynamic community interests. In contrast to most existing studies in scientific resource recommendation, his research developed an enhanced understanding of the scholarly network from a topical content perspective and investigated the use of full-text citation data to improve the overall recommendation ranking performance (Jensen, Liu, Yu, & Milojevic, 2016). Meanwhile, the proposed algorithms/system help students and scholars understand the challenging scientific publications and formulas (Liu, Jiang, & Gao, 2015).

**Wei Lu, Wuhan University:** Dr. Wei Lu is the professor and the vice dean of School of Information Management, Wuhan University. His research interests include information retrieval, data mining and academic document understanding. He was also recently elected as Youth Yangtze River Scholar by the Ministry of Education of the People’s Republic of China. He did his postdoc at City University London, and worked as a visiting researcher at Royal School of Library Science, Denmark. He has published dozens of papers in journals, conferences and workshops.

**Ying Ding, Indiana University Bloomington:** Dr. Ying Ding is the Associate Director of Data Science Online Program and will serve as the primary advisor for the Data Science Online and Certificates Program, as well as an Associate Professor at School of Informatics and Computing, Indiana University. She was recently elected as Yangtze River Scholar by the Ministry of Education of the People’s Republic of China. The Yangtze River Scholar award is the highest academic honor given by the People’s Republic of China. Previously she worked as a senior researcher at the University of Innsbruck, Austria and as a researcher at the Free University of Amsterdam, the Netherlands. She has been involved in various NIH and European-Union funded Semantic Web projects. She has published 170+ papers in journals, conferences and workshops.
5  Acknowledge

This SIE is supported by Thomas Reuters Digital Library, and Thomas Reuters will provide scholarly data to all the SIE participants.

References


Second Interaction and Engagement on Information Research and Learning with Lifelogging Devices (IRLLD 2017)\(^1\)

Hideo Joho\(^1\), Cathal Gurrin\(^2\), and Frank Hopfgartner\(^3\)
\(^1\)University of Tsukuba, Japan
\(^2\)Dublin City University, Ireland
\(^3\)University of Glasgow, UK

\begin{abstract}

iSchools have their roots in the collection, storage, analysis, and dissemination of archived materials of human activities. We foresee that sensing data via lifelogging devices (or Internet of Things at large) will eventually shape its significant part in the coming years. Information Research and Learning with Lifelogging Devices (IRLLD) aims to offer a unique opportunity to experience various lifelogging devices such as wearable video recorders, wearable cameras, GPS sensors, and audio recorders. Following the successful 1st IRLLD at iConference 2016 SIE (Joho, Gurrin, & Hopfgartner, 2016), the 2nd edition of IRLLD at iConference 2017 SIE offers extended lifelogging devices such as biometric sensors. IRLLD 2017 also demonstrates how to access a large lifelog dataset called NTCIR-12 Lifelog Test Collection, created by the organisers of IRLLD 2017. The intended audience includes information behavioural researchers (both qualitative and quantitative), multimedia and/or UI developers, students who want to improve their work/life experience, and educators who explores the ways to develop reflective learning programs using lifelogging data.

\textbf{Keywords:} Lifelogging
\end{abstract}


\textbf{Copyright:} Copyright is held by the authors.

\textbf{Contact:} hideo@slis.tsukuba.ac.jp.

\section{Introduction}

\subsection{Purpose and Intended Audience}

Wearable devices and mobile devices have become commodity, and collecting a massive amount of digitised personal lifelog data has become much easier than before. This creates many opportunities for iSchool researchers, developers, learners, and educators. Some of the recent topics studied in this domain includes (but are not limited to):

\begin{itemize}
  \item How to conduct a mixed method study using conventional methods (e.g., questionnaires and interviews) and lifelogging devices
  \item How to develop a learning program that incorporates lifelogging as the core component
  \item How to synthesise data from lifelogging devices to enhance learning analytics (e.g., MOOCs)
  \item How to retrieve relevant objects from diverse personal multimedia collections
  \item How to visualise lifelog data to support recollecting, reminiscing, retrieving, reflecting, and remembering (Sellen & Whittaker, 2010)
  \item How to aggregate lifelog data for institutional lifelogging
  \item How to overcome privacy issues for various lifelogging applications
\end{itemize}

\(^1\)\url{http://irlld2017.computing.dcu.ie/}
• How to optimise lifelogging technologies to special social groups (e.g., senior citizens, children, etc.)

As can be seen, many of these challenges require to bridge among researchers, developers, learners, and educators, to fully leverage the power of lifelogging devices and their data. This session for interaction and engagement (SIE) at iConference 2017 is to create a living lab environment where participants can experience various lifelogging devices such as wearable video recorders, wearable cameras, GPS sensors, audio recorders, or biometric sensors. Participants can play with the wearable devices to see what kind of data can be collected, analysed, and visualized. The participants also have an opportunity to learn how to access a large lifelogging dataset created by the organiser of the SIE.

The intended audience includes information behavioural researchers (both qualitative and quantitative), multimedia and/or UI developers, students who want to improve their work/life experience, and educators who explores the ways to develop reflective learning programs using lifelogging data.

2 Organisers

IRLLD 2017 is internationally organized by the following team.

• Hideo Joho, Research Center for Knowledge Communities, Faculty of Library, Information and Media Science, University of Tsukuba, Japan.

• Cathal Gurrin, School of Computing, Dublin City University, Ireland.

• Frank Hopfgartner, Humanities Advanced Technology and Information Institute, University of Glasgow, UK.

Joho and Hopfgartner are from iSchool member institutions. Joho, Gurrin, and Hopfgartner have recently organised a panel session at JCDL 2015 (Gurrin & Hopfgartner, 2015), and a methodological panel at ASIS&T 2016 (Joho, Gurrin, Heinström, & Matsubayashi, 2016). Both panels motivated us to propose a technical SIE as a hands-on session with lifelogging devices at iConference 2017. The organising team is also co-organisers of NTCIR Lifelog Task (Gurrin, Joho, Hopfgartner, Zhou, & Albatal, 2016) which provides a large-scale reusable dataset for researchers and developers to develop and evaluate innovative lifelog systems (See http://ntcir-lifelog.computing.dcu.ie/ for detail).

3 Key participants

Following a success of IRLLD 2016², IRLLD 2017 was to set the various wearable devices as a central piece of the event, and to facilitate interactions different sectors of people such as researchers, developers, learners,

²http://irlld2016.computing.dcu.ie/
and educators. Due to the nature of our purpose, the organisers acted as the main contributors of the event. In addition, we invited contributions to present works on research, development, teaching, learning, and practising of lifelogging devices at the SIE.

4 Agenda

The program consisted of two sessions as shown in Table 1. The first session was designed to first set the scene for all participants by providing an introductory presentation on lifelogging. Then, a hands-on session provided an opportunity for participants to play with various lifelogging devices. The organisers provided additional information about the capability and limitation of the devices. Audiences were encouraged to share their experience on lifelogging devices too. We created a closed space where participants can safely enjoy the lifelogging experience with our SIE participants, while minimizing recording of other conference participants. Such a living lab experience was our central piece of instalment.

In the second session, we had interactive presentations where invited participants showed a demo system, early research outcomes, or crazy ideas about how to use lifelog data. The presentation was more informal than usual presentations, where other participants were encouraged to express their ideas and opinions. Finally, we had a round-table session to identify some of the core research directions regarding the development and use of lifelog devices in Information Research and Learning.

5 Relevance to the Conference/Significance to the Field

iSchools have its roots in the collection, storage, analysis, and dissemination of the recorded material of human activities. It used to be books and libraries for a couple of thousand years. Web contents on The Internet took over many parts of the place in the last two decades. Sensing data via lifelogging devices (or Internet of Things at large) will take over the significant part of human archiving in the near future. IRLLD 2017 looked into such core issues of Information Research. It also offered an opportunity to familiarise various lifelogging devices and obtained data to the iSchool community.

References