Retrieval and Classification of Dental Research Articles

W.C. Bartling[1]*, T.K. Schleyer[2], S. Visweswaran[1]

[1]Center for Biomedical Informatics, University of Pittsburgh, Pittsburgh, PA; [2]Center for Dental Informatics, School of Dental Medicine, University of Pittsburgh, Pittsburgh, PA; *corresponding author, wcb@cbmi.upmc.edu

Adv Dent Res 17:115-120, December, 2003

Abstract — Successful retrieval of a corpus of literature on a broad topic can be difficult. This study demonstrates a method to retrieve the dental and craniofacial research literature. We explored MeSH manually for dental or craniofacial indexing terms. MEDLINE was searched using these terms, and a random sample of references was extracted from the resulting set. Sixteen dental research experts categorized these articles, reading only the title and abstract, as either: (1) dental research, (2) dental non-research, (3) non-dental, or (4) not sure. Identify Patient Sets (IPS), a probabilistic text classifier, created models, based on the presence or absence of words or UMLS phrases, that distinguished dental research articles from all others. These models were applied to a test set with different inputs for each article: (1) title and abstract only, (2) MeSH terms only, or (3) both. By title and abstract only, IPS correctly classified 64% of all dental research articles present in the test set. The percentage of correctly classified dental research articles in this retrieved set was 71%. MeSH term inclusion decreased performance. Computer programs that use text input to categorize articles may aid in retrieval of a broad corpus of literature better than indexing terms or key words alone.

Introduction

The information that researchers use in their scientific activities can come from many sources, such as the published literature, Web sites, and databases. As the amount of scientific knowledge grows, and as an increasing portion of it becomes available more easily, researchers are faced with the difficult task of finding research relevant to their work. Typically, researchers concentrate their search on a relatively narrow problem area and hone their retrieval skills (or those of their librarian) to a satisfactory level. While many literature retrieval systems have been optimized for relatively detailed queries, such as "Which genes are predominantly expressed in oral squamous cell carcinoma?", the situation is different for more general queries. For instance, a question such as "For what types of research questions are murine models being used in dental research?" is far more difficult to answer using standard literature retrieval systems such as PubMed (www.nlm.nih.gov) and OVID (OVID Technologies), since the user must possess relatively deep knowledge of the indexing schemes and retrieval processes of each database.

The impetus for developing the methodology reported in this paper was a very high-level question: "What are the characteristics of the dental and craniofacial research literature available through MEDLINE, such as the frequency and relationship between research topics? What research trends are evident from the literature?" Bibliometric and content analyses of the literature are common methods to answer such questions and, in doing so, describe a scientific field. Co-occurrence of authors (Marion, 2002), journals (Morris and McCain, 1998), or indexing terms (Marion, 2002) in a body of literature can be used to visualize the structure of a scientific field and determine its boundaries. This type of content analysis has been performed in many fields, including medical informatics, information retrieval, and software engineering (Morris and McCain, 1998; Ding and Foo, 1999; Marion, 2002). Free-text analyses are becoming more prevalent to characterize documents and to determine relationships among them. Recent studies using free text to classify documents identified patient subgroups in dictated chest radiography reports (Chapman *et al.*, 2001) and classified Web pages (Moore *et al.*, 1997; Asirvatham, 2003). Once the relationships among documents have been determined, various statistical methods, such as multidimensional scaling and hierarchical clustering, can be used to quantify those relationships.

In this paper, we describe a method for retrieving the dental and craniofacial research literature from MEDLINE. We also present an analysis of a subset of this literature to illustrate what a "bird's-eye" view of the dental research literature could look like. We hope that our method, once refined, will be useful in answering more general research questions using the MEDLINE database.

Background and Significance

In most literature databases, key words or indexing terms are assigned to individual articles to facilitate retrieval. In MEDLINE, those key words are known as the Medical Subject Headings (MeSH). The MeSH are a hierarchical system of words and phrases that are assigned to documents by human indexers to describe content and other related information such as study methodology. As of this writing, MEDLINE contained approximately 12 million citations dating back to 1966, and there are approximately 22,000 unique MeSH terms (http://www.nlm.nih.gov/pubs/factsheets/mesh.html). However, since many terms occur in multiple places in the hierarchy, there are approximately 40,000 MeSH term locations, each with a unique identifier. Each article in MEDLINE is tagged with about 12-20 MeSH terms, with generally one to four being the major topics. Dental literature encompasses many different facets, such as molecular studies of dental materials, clinical procedures, population studies, psychological studies, and many more. Therefore, we assumed that dental articles, both research and non-research, were indexed with many different and unrelated MeSH terms. This study was intended to develop an information retrieval method that successfully retrieves this breadth of the dental and craniofacial research literature so that a comprehensive content analysis of it can be accomplished.

Human indexers make relatively complex decisions when assigning MeSH terms to an article. After reading the complete article, they pick major headings to describe the paper's main

Key Words

Dental research, information retrieval, inter-rater agreement, MEDLINE, MeSH, data mining, text classification.

Publication supported by Software of Excellence (Auckland, NZ)

Presented at "Dental Informatics & Dental Research: Making the Connection", a conference held in, Bethesda, MD, USA, June 12-13, 2003, sponsored by the University of Pittsburgh Center for Dental Informatics and supported in part by award 1R13DE014611-01 from the National Institute of Dental and Craniofacial Research/National Library of Medicine.

focus. Other MeSH terms add more detail to the article's description. The indexers must not only be very familiar with the terms in the MeSH hierarchy, but also they must predict the searcher's behavior and assign terms that the average searcher would be expected to choose when attempting to find articles about the specific subject(s). A study of indexing consistency in MEDLINE has shown that indexers tend to be more consistent with general MeSH terms and less consistent when using more specific terms (Funk and Reid, 1983). A key problem in indexing is that not all general descriptors that could possibly apply to an article can actually be assigned to it. For example, if a researcher wants to search only papers within the domain of dental research, MeSH indexing is of little help. Searching MEDLINE for articles tagged with the MeSH term DENTAL RESEARCH yielded approximately 850 citations as of August, 2003. Most of these papers were *about* the topic of dental research and were not dental research papers. In contrast, previous analyses of the dental literature have focused on specific and relatively narrow topics, such as clinical evidence in pediatric dentistry (Yang *et al.*, 2001), orthodontics (Sun *et al.*, 2000), implantology (Russo *et al.*, 2000) and prosthodontics (Nishimura *et al.*, 2002), oro-facial pain (Macfarlane *et al.*, 2001), and randomized controlled trials in dentistry (Sjögren and Halling, 2000).

Published analyses of the literature in a scientific field have used several approaches for identifying the target literature and content analysis. In a recent study of biomedical informatics, the investigators began by identifying journals associated with this field (Morris and McCain, 1998). Then, they performed intercitation studies among productive journal titles and examined co-citation data for proposed core journals using multivariate analyses. The study indicated the presence of a core literature and identified several major research areas. An analysis of the dental research literature faces several challenges and, therefore, must draw on multiple methods. First, it is very difficult to "bound" the field of dental research by attempting to identify a set of core journals, as Morris and McCain did in the biomedical informatics study. Due to dentistry's multi-disciplinarity, its research tends to be published in many different journals, including those in medicine, biomedical engineering, basic sciences, and psychology (Bush, 1996; Macfarlane *et al.*, 2001). Focusing on specific journals would therefore limit one's view of dental research significantly. On the other hand, locating the collection of individual dental research papers from MEDLINE is made difficult by the retrieval challenges discussed above.

Computer-based methods for analyzing text can be helpful in overcoming this obstacle. With these methods, a computer program is trained to search a collection of documents based on the characteristics of a training set of documents similar to those that the searcher would like to retrieve. Various algorithms have been developed and evaluated for classifying or categorizing text (Cavnar and Trenkle, 1994; Lewis and Gale, 1994; Lewis and Linguette, 1994; Yang, 1999). One advantage of these methods is that they can be applied to any text, regardless of source, and they can process very large collections of documents in a short period of time. The success of these methods depends on how well they can discriminate relevant documents from irrelevant ones, and usually involves some trade-off between sensitivity and specificity.

The approach described in this paper is significant for two reasons. First, it provides a method for retrieving and analyzing the dental research literature. Such a comprehensive analysis has not been conducted, but may be useful for several purposes. For instance, it could provide a "bird's-eye" view of dental research as a field, which may be useful to policymakers, funding agencies, and researchers. Analyzing changes in dental research topics over time may provide an indication of past, current, and emerging trends. Second, our methodology may be used to facilitate searches within global topics that cannot be easily delineated using current search engines. For example, our methodology could serve as a filter for searches that target only articles within a specific area of dental research and, thus, may increase specificity for researchers interested only in papers from their field. We are not attempting to evaluate or compare our retrieval performance with text classification methods; our aim is to demonstrate to the dental research community how the science of informatics can be applied to their field.

Methods

Our retrieval strategy involved four phases. First, we searched MeSH for terms related to dentistry and dental science and used these terms to retrieve a complete corpus of dental articles referenced in MEDLINE. Second, we randomly sampled this set and had dental research experts categorize these articles based on titles and abstracts. Third, using this gold standard set of rated articles, we trained Identify Patient Sets (IPS) (Aronis *et al.*, 1999), a probabilistic text classification computer program developed at the University of Pittsburgh, to determine the characteristics of dental research articles. We then applied the IPS models to test sets to determine retrieval success. Fourth, we looked at various characteristics of the categorized sets to see what distinguishes dental research articles from other articles. A more detailed description of the phases of our methodology follows.

Searching MeSH for dental and craniofacial terms

We examined the MeSH manually and recorded all terms that pertained to a dental or craniofacial topic. Iteratively, one of the author (WCB) and a dental research librarian compared results until they were confident that they had all relevant dental terms. If a term existed as a child of a term already chosen, that term was not included. All terms chosen were exploded for the final search. For example, ORTHODONTICS occurs lower in the MeSH hierarchy than DENTISTRY. Because we included DENTISTRY in our search, all terms below it, including ORTHODONTICS, were part of the final search. Terms were chosen with the goal of higher recall, or sensitivity, since we did not want to exclude any relevant articles from our retrieved set. As a consequence, we purposely accepted the trade-off of more irrelevant citations in our resulting set.

We used the OVID interface to search the MEDLINE database. Once our search strategy was complete, we limited it to English-language articles that contained abstracts published between 1966 and 2002. We omitted the following publication types: comment, editorial, biography, historical article, letter, news, review, review of reported cases, review, tutorial, case report, and dictionary. We randomly sampled 1000 articles from the resulting set for manual review by expert raters. These references were divided into 5 mutually exclusive groups of 200 articles each.

Expert ratings

We sought out dental research experts from academia, industry, and government as volunteers to review articles. We used referrals from dental school and medical informatics faculty, along with Internet searches, to identify possible participants. We e-mailed 50 experts, and 16 agreed to participate. We assigned three experts to review each group of 200 abstracts, with the exception of one group that had four reviewers. Our final set contained 990 references, because we removed ten duplicates.

We developed a World Wide Web interface that allowed experts to review articles at their convenience and in as many sessions as needed. The interface was developed in the

programming language PHP (Apache Software Foundation) and was connected to a MySQL database (MySQL AB). Reviewers were allowed to change their ratings if necessary. Reviewers were given 4-6 weeks to rate the 200 abstracts. Abstracts were randomized by publication date and among the 5 groups. The rating interface displayed only the title and abstract of each reference, and the rater, using radio buttons, was prompted to classify the text as either:

(1) dental or craniofacial research;
(2) dental or craniofacial topic, but not research;
(3) not a dental or craniofacial topic; or
(4) not sure.

Criteria for inclusion in each category were displayed on the instructions page within the interface.

The ratings for each abstract were counted, and the reliability (Cronbach's α) was calculated for each rater group. Classifications 3 and 4 were combined, so that three classes were used in the reliability measures: 1, 2, and (either 3 or 4). Each reference was placed in a class based on a majority rating of experts. For example, in a group of abstracts with three raters, two raters must have rated a document as "Dental or craniofacial research" for an article to be placed in that category. Those articles without a majority rating were placed in the "3 or 4" (non-dental or not sure) group for classification purposes.

Probabilistic text classification

Identify Patient Sets (IPS) (Aronis *et al.*, 1999) is a general-purpose medical record retrieval program developed at the University of Pittsburgh. Given a set of documents, IPS compiles a dictionary of all words and UMLS (Unified Medical Language System, www.nlm.nih.gov) phrases within them. This dictionary is then used to create a vector for each document, representing the presence or absence of each word in the dictionary. A training set of labeled documents is used to create probabilistic models for the classes of documents in the set.

IPS is trained in a binary way, that is, a document that is of the desired class is a "HIT", and one that is not in that class is a "MISS". We chose to combine the three categories that were not "dental or craniofacial research" so that IPS would make a binary classification. By using this method, we had more documents in our training and test sets than we could have if we trained and tested on 3 or even 4 classes. For instance, if an article were rated by a majority of raters in that group as "dental or craniofacial research", that article would be labeled as a "HIT" for input to IPS. An article that was not categorized as "dental or craniofacial research" was labeled as a "MISS". The "HIT" and "MISS" groups of documents are analyzed to find those words or UMLS phrases that discriminate between the two groups.

The 990 articles were divided into a training set (n = 693) and a test set (n = 297) such that the proportion of dental research articles was the same in the two sets (60%), and each article had a 30% chance of being randomly assigned to the test set. We used the single-holdout method of cross-validation in this study. That is, 70% of the rated documents were used to train IPS, and 30% of the documents were used to test the IPS models generated. These sets were mutually exclusive. In other words, a document occurring in the test set was "held out", or did not occur in the training set, and *vice versa*.

Three experiments were conducted with the following inputs for each article to IPS: (1) title and abstract only, (2) MeSH terms only, and (3) title, abstract, and MeSH terms. The same training set was used in each experiment to create corresponding IPS models. Those models that attained both sensitivity and specificity of at least 0.60 on the training set were then applied to the test set to predict the category of each article in the set. For instance, a sensitivity of 0.60 of an IPS model indicated that the model correctly identified 60% of the articles rated by experts as "Dental or craniofacial research". Sensitivity, specificity, precision, and F-measure were calculated for each model applied to the test set.

Sensitivity, or recall, is the ratio of documents that IPS correctly categorized as dental research to all documents classified by experts as dental research. Specificity is the ratio of articles that IPS did not categorize as dental research to those articles judged by experts as not being dental research. Precision is analogous to positive predictive value (PPV), or the proportion of dental research documents out of all documents retrieved. F-measure is commonly used to find the best models in information retrieval and takes into account both sensitivity and precision. The F-measure was calculated according to the following equation:

$$F = 2*(\text{Sensitivity} * \text{Precision})/(\text{Sensitivity} + \text{Precision}).$$

The constant "2" is used to weight sensitivity and precision equally. This constant can be varied if unequal weighting of sensitivity and precision is desired.

Characteristics of gold standard set

We performed several analyses to characterize our gold standard set. As described above, these analyses can serve as a template for a larger set. We generated counts of unique journal titles in the rating categories: dental research, dental non-research, and non-dental. We then looked at the types of journals in which these were published (www.ncbi.nlm.nih.gov/entrez), *i.e.*, dental, medical, basic science, etc., and calculated the percentage published in dental journals. We also calculated total counts of all MeSH terms and major MeSH terms that occurred in each rating category. We used programs written in Python (www.python.org) to extract these data from the OVID output files (Reprint/Medlars format) of the references retrieved.

Results

MEDLINE search

After limiting our search to English-language articles published between 1966 and 2002, three MeSH terms, when exploded, contained the most dental articles. These were: DENTISTRY, STOMATOGNATHIC SYSTEM, and STOMATOGNATHIC DISEASES. We removed articles indexed with PHARYNX and PHARYNGEAL DISEASES from the last two terms, respectively. These three major categories and combinations contained the following numbers of documents:

(1) DENTISTRY: 228,629
(2) STOMATOGNATHIC SYSTEM (NOT PHARYNX): 188,826
(3) STOMATOGNATHIC DISEASES (NOT PHARYNGEAL DISEASES): 238,839

Together, these three categories contained 459,758 articles. Approximately 70 additional dental MeSH terms—such as GINGIVAL CREVICULAR FLUID, STREPTOCOCCUS MUTANS, and DENTAL RECORDS—were not children of the three large categories above and were included. Many terms occurred in isolated locations in the hierarchy. For example, GINGIVAL CREVICULAR FLUID and DENTINAL FLUID are present only under EXUDATES AND TRANSUDATES in the ANATOMY category. They do not occur elsewhere in the MeSH hierarchy. Many articles were indexed with these isolated terms, including: terms in BIOMEDICAL AND DENTAL MATERIALS, specific oral microbial species names such as PORPHYROMONAS GINGIVALIS, and many dental public health and education terms such as DENTAL WASTE and DENTAL EDUCATION. 62,255 articles were indexed with MeSH terms that were not indexed within the three major categories above. After unwanted publication types were filtered out and those containing abstracts were retained,

TABLE 1 — Characteristics of Gold Standard Set

Classification	n	Percentage	Unique Journal Titles	Unique Major MeSH Terms	Percentage in "Dental" Journals
Dental research	591	60%	250	1027	62%
Dental non-research	115	12%	78	224	68%
Non-dental	129	13%	128	411	5%
Not sure	3	< 1%			
No majority	152	15%			
TOTAL	990	100%			

137,816 articles remained.

Characteristics of gold standard set

Sixteen dental research experts used the Web interface to rate articles. It took approximately two months for all 16 raters to complete their ratings. All five groups had acceptable reliability (Cronbach's $\alpha > 0.70$). Table 1 contains a summary of characteristics of our rated set of 990 articles. Our gold standard set contained 72% dental articles (60% dental research and 12% dental non-research), and 13% non-dental articles. Fifteen percent of the articles did not have a majority rating. The 591 dental research articles were published in 250 different journals. Similar diversity in journal titles was seen in the other categories. Sixty-two percent of dental research articles and 68% of dental non-research articles were published in journals that are considered "dental" titles. A surprising finding was that 5% of non-dental articles were published in dental journals. The journals indexed in MEDLINE are assigned one or more of 127 general subject types (www.ncbi.nlm.nih.gov/entrez/query/fcgi?db=journals). We considered those titles labeled with "Dentistry" as dental titles.

One thousand twenty-seven different major MeSH terms were used to index the 591 dental research articles, 224 major MeSH terms for the 115 dental non-research articles, and 411 major MeSH terms for the 129 non-dental articles. Table 2 shows the most commonly used major MeSH terms in each category. DENTAL CARIES was the most commonly assigned major MeSH term in the dental research set, occurring in 4.5% of the articles. MOUTH NEOPLASMS was the second most commonly used, in 3.7%. The remaining terms in the set included many dealing with oral pathology, clinical dentistry, periodontology, and dental materials. DENTAL EDUCATION and DENTURE DESIGN were the most common major MeSH in the dental non-research set, each occurring in 5.9% of the articles. In contrast to their occurrence in dental research, DENTAL CARIES and MOUTH NEOPLASMS each occurred as major MeSH terms in less than 2% of the dental non-research articles. The dental non-research category also included many prosthodontic terms but fewer describing pathology and neoplasms. However, many major MeSH terms in this set focused on surgical or dental procedures. MeSH terms describing salivary glands were common in the non-dental set, with SUBMANDIBULAR GLAND being the most common major term assigned (3.1%).

Probabilistic text classification (IPS)

Each IPS model consists of a unique set of terms, either text words or UMLS phrases, that the program determined to discriminate between dental research and all other articles. The same training and test sets were used for each of the models generated. The acceptable models generated by the training set with different inputs are shown in Table 3. The numbers of models generated for each input above the threshold of 0.6 sensitivity and specificity are as follows: six models were generated by title and abstract only; three models by title, abstract, and MeSH terms; and two models by MeSH terms alone. The best models when applied to the test set resulted in the following F-measures: (1) title and abstract only, F = 0.74; (2) title, abstract, and MeSH, F = 0.70; and (3) MeSH only, F = 0.65. The highest sensitivity, 0.80, was achieved on a model with only title and abstract. The highest specificity, 0.71, was achieved on a model with title, abstract, and MeSH. This same model also resulted in the highest precision, 0.74. That is, including more words in the model decreased the number of articles incorrectly identified (false-positives) as dental research.

IPS models trained on title and abstract alone provided the best results. Some dental terms were discriminatory between dental research and all other articles in these models, *e.g.*,

TABLE 2 — Numbers of Major MeSH for Each Rating Category and Probability of MeSH Term in Each Category

Dental Research (n = 591)			Dental Non-research (n = 115)			Non-dental (n = 129)		
n	Major MeSH	p	n	Major MeSH	p	n	Major MeSH	p
27	Dental Caries	0.045	7	Education, Dental	0.059	4	Submandibular Gland	0.031
22	Mouth Neoplasms	0.037	7	Denture Design	0.059	4	Skin	0.031
19	Mouth Mucosa	0.032	6	Mandible	0.051	4	Epidermal Growth Factor	0.031
19	Gingiva	0.032	5	Maxilla	0.042	4	Adenylate Cyclase	0.031
18	Periodontal Diseases	0.030	5	Dentistry	0.042	3	Salivary Glands	0.023
17	Periodontitis	0.029	4	Surgical Flaps	0.034	3	Hip Prosthesis	0.023
17	Mandible	0.029	4	Osteotomy	0.034	3	Fluorides	0.023
17	Dental Implants	0.029	4	Denture, Complete	0.034	3	Chromosomes	0.023
15	Carcinoma, Squamous Cell	0.025	4	Dentists	0.034	3	Chlorides	0.023
14	Submandibular Gland	0.024	4	Dental Restoration, Permanent	0.034	3	Cementation	0.023
14	Dental Pulp	0.024	3	Tooth	0.025			
14	Composite Resins	0.024	3	Practice Management, Dental	0.025			
14	Cleft Palate	0.024	3	Malocclusion	0.025			
13	Tongue	0.022	3	Jaw Relation Record	0.025			
13	Dental Bonding	0.022	3	Jaw	0.025			
12	Fluorides	0.020	3	General Practice, Dental	0.025			
12	Dentin	0.020	3	Education, Dental, Continuing	0.025			
12	Dental Enamel	0.020	3	Denture, Partial, Removable	0.025			
11	Saliva	0.018	3	Denture, Overlay	0.025			
11	Parotid Gland	0.018	3	Denture Precision Attachment	0.025			
11	Dental Restoration, Permanent	0.018	3	Dental Implants	0.025			
11	Dental Plaque	0.018	3	Dental Care	0.025			
10	Gingivitis	0.017	3	Crowns	0.025			
			3	Anesthesia, Dental	0.025			

periodontal, tooth, amalgam, cervical, crest, plaque, and *pulp.* However, many non-dental terms were discriminators, including: *statistically, human, study, mean, layer, females, hard, exogenous, weeks, differentiation, evaluate,* and *subjects.*

Table 3 — Identify Patient Sets Results

IPS Models	Training Set (n = 690) Sensitivity	Specificity	Precision	F-measure	Test Set (n = 300) Sensitivity	Specificity	Precision	F-measure
Title & abstract	0.61	0.80	0.82	0.70	0.63	0.58	0.69	0.66
	0.63	0.79	0.82	0.71	0.64	0.62	0.71	0.67
	0.68	0.74	0.80	0.74	0.64	0.53	0.67	0.65
	0.70	0.68	0.76	0.73	0.71	0.59	0.72	0.71
	0.80	0.65	0.77	0.78	0.79	0.48	0.70	0.74
	0.80	0.64	0.77	0.78	0.80	0.45	0.69	0.74
MeSH only	0.61	0.70	0.75	0.67	0.61	0.57	0.68	0.64
	0.68	0.68	0.76	0.72	0.63	0.53	0.67	0.65
Title, abstract, & MeSH	0.81	0.64	0.77	0.79	0.74	0.47	0.67	0.70
	0.69	0.69	0.77	0.73	0.62	0.58	0.69	0.65
	0.61	0.78	0.81	0.70	0.56	0.71	0.74	0.64

Discussion

Many dental research articles are indexed with MeSH terms that occur in unusual places in the hierarchy, and our inclusion of these increased our recall. We assumed that many dental articles were indexed with less obvious MeSH terms, and our results support this. Since our goal was high recall and not necessarily high precision, the 13% occurrence of non-dental articles in our gold standard set was not unexpected. Since our gold standard set included 72% dental articles (60% research and 12% non-research), our MEDLINE search had 0.72 sensitivity for dental articles and 0.60 sensitivity for dental research articles, provided that all possible dental articles were in the retrieved set.

The variety of journal titles included and major MeSH terms assigned in the dental research and dental non-research sets show that the science is very diverse and that our retrieved set encompasses various areas of clinical dentistry, the basic sciences, and biomaterials. The diversity of major MeSH terms used to describe dental and craniofacial research articles shows that using MeSH alone to retrieve such articles may be very costly in terms of time needed and knowledge of MeSH required. Using MeSH in combination with text analysis may improve success. Also, since dental research articles occur almost 40% of the time in non-dental journals, we may consider including other databases in further studies, *e.g.*, BIOSIS or PsycINFO, since these articles may occur in journals that are not indexed in MEDLINE.

For this study, we considered a sensitivity of 0.70 as acceptable. Only 4 of our 11 IPS models met this goal. Since we have determined that discriminating research from non-research involves analyzing free text, and that many non-dental terms were discriminators, further work with text classification methods to increase performance is necessary. Filtering of dental articles from all others may increase performance. Since 60/72 or 84% of dental articles are dental research in our gold standard set, we may be able to retrieve the dental research literature more successfully if non-dental articles are filtered out. Considering that many words that IPS models used to discriminate dental research from non-research were non-dental terms, *e.g.*, *statistically, differentiation,* and *subjects,* IPS may perform better when non-dental articles are not included.

Another method that may increase performance is parsing the text semantically and linking it to UMLS concepts and semantic types. A semantic knowledge representation computer program, MetaMap (Aronson, 2001), developed at the National Library of Medicine, may help achieve this task of discriminating dental research from non-research by providing additional information to classify articles.

Limitations

One limitation to our study was that we grouped the dental non-research with non-dental articles for our text classification procedure. Because some dental terms were determined by IPS to be discriminatory between dental research and other articles, combining these two categories may have decreased performance. That is, a dental research article may have been excluded because it contained a dental term that did not occur in many research articles, but occurred in many non-research articles. The word may be associated with a narrow research area, and its exclusion may result in our missing a growing area of research.

A limitation to our text classification method was that the IPS models were constructed with the use of only one training set. A more comprehensive analysis of performance would include cross-validation with many training and test sets. With either ten-fold cross-validation, or the single-hold-out method, we could determine whether the performance between the different models was statistically significant. We plan this for future work.

The high prevalence (60%) of dental research articles in our gold standard set may have been another limitation. Because of this prevalence, with similar sensitivity, our precision was greater than it would be in a set with a lower percentage of dental research articles, such as all of MEDLINE. However, training IPS on such a set would have required a much larger set of documents. That is, to have a large enough number of dental research documents for a good IPS model to be constructed, more raters rating many more documents would be needed.

Conclusion

Analyzing the content of a corpus of literature in a scientific discipline may help to define the science and to provide information for researchers to base studies or to find relationships between subdisciplines. The dental and craniofacial research literature is very diverse and difficult to retrieve. Using a combination of indexing terms and free-text analysis may be required to retrieve such a broad scientific literature. Upon successful retrieval of this literature, visual representation and statistical analysis may be used to examine trends and topics in dental and craniofacial research.

Acknowledgments

The authors thank Amy Gregg, MLIS Falk Library for the Health Sciences, University of Pittsburgh. This work was supported in part by an NLM/NIDCR Training Grant.

References

Aronis JM, Cooper GF, Kayaalp M, Buchanan BG (1999). Identifying patient subgroups with simple Bayes'. *Proc AMIA Symp* 1999:658-662.

Aronson A (2001). Effective mapping of biomedical text to the UMLS metathesaurus: the MetaMap program. *Proc AMIA Symp* 2001:17-21.

Asirvatham AKR (2003). Web page categorization based on document structure (www.iiit.net/students/stnd_pdfs/kranthi.pdf), pp. 1-9. Last accessed 9/30/2003.

Bush R (1996). Biomaterials: an introduction for librarians. *Sci Technol Libr* 15(4):3-17.

Cavnar W, Trenkle J (1994). N-Gram-based text categorization. In: SDAIR-94, 3rd Annual Symposium on Document Analysis and Information Retrieval, April 11-13, 1994, Las Vegas, NV. pp. 161-169.

Chapman WW, Fiszman M, Frederick PR, Chapman BE, Haug PJ (2001). Quantifying the characteristics of unambiguous chest radiography reports in the context of pneumonia. *Acad Radiol* 8:57-66.

Ding YCG, Foo S (1999). Mapping the intellectual structure of information retrieval studies: an author co-citation analysis, 1987-1997. *J Inf Sci* 25(1):67-78.

Funk ME, Reid CA (1983). Indexing consistency in MEDLINE. *Bull Med Libr Assoc* 71:176-183.

Lewis D, Gale W (1994). A sequential algorithm for training text classifiers. In: 17th Annual ACM/SIGIR conference, July 3-6, 1994, Dublin, Ireland. New York, NY: Springer-Verlag, pp. 3-12.

Lewis D, Linguette M (1994). A comparison of two learning algorithms for text categorization. In: Third Annual Symposium on Document Analysis and Information Retrieval, April 11-13, 1994, Las Vegas, NV, pp. 81-93.

Macfarlane TV, Glenny AM, Worthington HV (2001). Systematic review of population-based epidemiological studies of orofacial pain. *J Dent* 29:451-467.

Marion LSMK (2002). Contrasting views of software engineering journals, author co-citation choices and indexer vocabulary assignments. *J Am Soc Inf Sci Tech* 52:297-308.

Moore JEH, Boley D, Gini M, Gross R, Hastings K, Karypis G, *et al.* (1997). Web page categorization and feature selection using association rule and principal component clustering. In: 7th Workshop on Information Technologies and Systems (WIT3 '97), December 13-14, 1997, Atlanta, GA, pp. 1-10.

Morris TA, McCain KW (1998). The structure of medical informatics journal literature. *J Am Med Inform Assoc* 5:448-466.

Nishimura K, Rasool F, Ferguson MB, Sobel M, Niederman R (2002). Benchmarking the clinical prosthetic dental literature on MEDLINE. *J Prosthet Dent* 88:533-541.

Russo SP, Fiorellini JP, Weber HP, Niederman R (2000). Benchmarking the dental implant evidence on MEDLINE. *Int J Oral Maxillofac Implants* 15:792-800.

Sjögren P, Halling A (2000). Trends in dental and medical research and relevance of randomized controlled trials to common activities in general dentistry. *Acta Odontol Scand* 58:260-264.

Sun RL, Conway S, Zawaideh S, Niederman DR (2000). Benchmarking the clinical orthodontic evidence on Medline. *Angle Orthod* 70:464-470.

Yang S, Needleman H, Niederman R (2001). A bibliometric analysis of the pediatric dental literature in MEDLINE. *Pediatr Dent* 23:415-418.

Yang Y (1999). An evaluation of statistical approaches to text categorization. *J Inf Retrieval* 1(1/2):67-88.