Automatic Categorization of Google Search Results for Medical Queries using JDI

Anantha K. Bangalore[a], Guy Divita[b], Susanne Humphrey[c], Allen Browne[d], Karen E. Thorn[e]

[a] *Lockheed Martin MSD, National Library of Medicine (NLM), United States*
[b] *Lockheed Martin MSD, National Library of Medicine (NLM), United States*
[c] *National Library of Medicine (NLM), United States*
[d] *National Library of Medicine (NLM), United States*
[e] *Lockheed Martin MSD, National Library of Medicine (NLM), United States*

Abstract

The web has become the primary source of medical information for consumers and health professionals. It is quite common for people to "Google" for information related to a medical topic. But the problem remains that as the number of documents increases on the web, the difficulty in quickly locating the best documents increases. Classifying results into meaningful categories, helps guide users to the most relevant set of results. Journal Descriptor Indexing (JDI) is a novel approach to fully automatic indexing. In this paper we explore the feasibility of using JDI to organize Google search results for medical queries into meaningful categories. For our experiments, we used JDI in combination with a set of heuristics to automatically categorize the search results for 5 query terms. Three independent reviewers reviewed and evaluated the automatic categorization for 3 documents for each query term. The results clearly suggest that this method offers promise. Additional work for improving the categorization as well as to determining whether a term is medical or not is also discussed.

Keywords:

Medical Subject Headings, Information Storage and Retrieval, Internet, Classification, MEDLINE, Indexing and Abstracting

Introduction

It is very common for consumers and health professionals these days to use search engines like Google to look up information related to various medical topics. Research shows that as many as 80 percent of Internet users have searched for personal health information online.[4] The number of documents for any given topic on the web is increasing exponentially. Even though search engines such as Google are worthwhile for searching the web, the results typically tend to be of low precision and high recall. More often than not, the results of interest are not among the top two pages. Users quickly lose interest, if the relevant results are not in the top 2 or 3 pages. To focus the search, users have to develop a good set of search terms, an often time consuming and challenging process. One approach to help users would be to quickly classify the results into meaningful categories that would allow users to drill down into a category of interest.

Google has implemented tools such as a spell checker and a query refinement tool to assist users in finding their documents of interest. Recently they have begun to categorize search results for some health related topics. The goal of this paper is to propose a complementary method to automatically categorize the web search results for medical queries. It is not to compare our method of categorization with Google's. Google does not provide categorization for any of the 5 terms chosen for our experiment.

In this paper we explore the feasibility of using the Journal Descriptor Indexing (JDI) methodology developed at NLM for automatic categorization of Google search results for medical queries. JDI is a novel approach to fully automatic indexing developed at the National Library of Medicine (NLM). In this approach, the JDI links a subject index to journal titles using journal descriptors (JDs) that correspond to biomedical specialties. [3]. JDI has been proven effective in characterizing MEDLINE documents. Typically documents from Google are more general in nature as opposed to the more specific nature of documents from MEDLINE. They are also not as well structured as MEDLINE documents and tend to include noise (external links, images, advertisements etc). Our experiments explore the feasibility of using JDI on Google documents for categorization.

For this paper we extracted 20 documents for 5 terms from Google using Google's SOAP-based application programmer interface (API) [6]. Of the 5 terms, 3 were medical, one was ambiguous with both medical and non-medical meaning, and the last was inherently non-medical. The JDI system created a JD profile for each of the 20 documents. We used the top 15 JDs for each document. We then developed a set of rules to determine the confidence level by which we could classify a given document into a particular category based on one or a

few JDs representing that document. Three reviewers independently evaluated the quality of the categorization for 3 of the same documents for each of the 5 query terms. Based on the results, this is a promising method for categorizing medical search results from Google.

Materials and Methods

Journal Descriptor Indexing (JDI)

JDI is based on NLM's practice of maintaining, in its serials file, a subject index to journal titles using a set of MeSH terms, known as JDs (journal descriptors) corresponding to biomedical specialties. For example, the Journal of Pediatric Surgery is indexed by the JDs Pediatrics and Surgery.

The JDI methodology associates JDs with words in titles/abstracts in a training set of about 435,000 MEDLINE records. Each record "inherits" JDs from the serial record matching the journal title. For example, a MEDLINE record with journal title Journal of Pediatric Surgery inherits the JDs Pediatrics and Surgery from the matching serial record. Each word in the training set can then be described by a JD profile, which is a list of JDs ranked according to the number of co-occurrences between the word and the JDs in the training set.

For example, the first three JDs, with scores in decreasing order, for the word "appendectomy" would be:

1	0.8631	Surgery
2	0.6787	Gastroenterology
3	0.5415	Diagnostic Imaging

Once JD profiles have been computed for each word in the training set, they can be used as the basis for indexing documents. We index a document by averaging the scores for each JD across the words in the document; these averages become the JD scores for the document. We then rank the JDs in decreasing order of these scores. That is, we treat the JD profile of a word as a JD vector, and the JD indexing of a document is computed as the centroid of these JD vectors.

For example, the first three JDs, with scores, returned by a MEDLINE title "Appendectomy in children" would be:

1	0.4623	Surgery
2	0.4230	Pediatrics
3	0.3675	Gastroenterology

The Surgery score is the average of the Surgery score for the word "appendectomy" (0.8631) and "children" (0.0614) in the training set. The score for Pediatrics is the average of the Pediatrics score for "appendectomy" (0.1645) and for "children" (0.6815); the score for Gastroenterology is the average of the Gastroenterology score for "appendectomy" (0.6787) and for "children" (0.0563).

Methodology

We selected 5 terms for this experiment. These are:

1) *Chronic bloody nose*
2) *Carotid artery surgery*
3) *Dark circles*
4) *Ventilator*
5) *Cold war*

Of these the first three are medical in nature. “Ventilator” is an ambiguous term that occurs in both medical and non-medical contexts. We picked this term to see if the JDI method would offer clues to distinguish medical from non-medical documents. “Cold war” is clearly non-medical and we wanted to investigate how the method would handle the results for non-medical terms in contrast to the results for the medical terms.

We used Google’s SOAP API to make queries to Google’s Search Web service. We retrieved the top 20 URLs for each of the 5 documents for a total of 100 documents. We then wrote a Java program to extract the actual HTML documents corresponding to each of these URLs. We extracted only the first level documents leaving the links embedded within each of these documents unexplored for this experiment. Some of these documents had no content wherein they either had only images on the first page or contained a redirection to another web document. In order to substitute for these documents we ran the queries corresponding to these documents on the Google browser search interface. We randomly picked documents to substitute for the ones with no content. We then manually edited each of these 100 HTML documents using Microsoft Word to remove as much of the header, footer, advertisements and any other extraneous information that may have been present in each of the documents.

We used a freeware tool called Emsa Html Tag Remover [7] to remove all HTML tags from each of the documents and create plain text files corresponding to each of the HTML files. These text files were then used as input to the JDI tool. The JDI tool created a JD profile consisting of the top 15 JDs for each of the documents.

In order to choose the list of the JDs which best categorizes a given document we developed a set of heuristics. A vast majority of documents retrieved from Google, unless they are from sources such as NLM’s MedlinePlus, tend to be non-specific in nature. It is very common for these documents to contain information about multiple topics. For example, documents retrieved for carotid artery surgery contained additional information related to “Vascular Diseases”. While some of these documents discussed the actual surgical procedure, others discussed side affects such as stroke. In order to classify these topics accurately, multiple journal descriptors may be required. Thus, we developed a means of selecting that set as the top *n* JDs based on the score. We also assigned confidence levels to each of the categorizations within that set. The following describes the heuristics used in determining the category and the confidence level:

1) If the first JD for a document has a score of less than 0.2, categorize that document as “Undefined”. The low score indicates that it is difficult to determine whether the corresponding document contains enough information to allow it to be meaningfully ca-

tegorized. This situation may occur for a variety of reasons.

- The document is completely non-medical in nature.
- The document may contain noise such as ads, images, links etc.
- The document is the first page for a given site that contains only introductory information and would require exploration of the links to determine the nature of the site.

2) Pick the top *n* JDs where the difference between the score of the first JD and the rest of JDs is < 0.1. If the difference between the top JD (primary) and the subsequent JDs is < 0.1, (secondary) then it is reasonable to assume that the document has some content in it relating to the secondary JDs as well. In this second case, it's feasible to assume that a combination of JDs indicate possible categorizations for that document. However, if the difference between the primary JD and the secondary JDs > 0.1, then it's feasible that the document from the secondary JD classification contains very little information about the topic the secondary JD represents. In these instances, the secondary JD selected for categorization can be discarded.
3) Assign a confidence level to the categorization based on the JD score.

0.1 – 0.2:	Undefined	(U)
0.2 – 0.3:	Weak	(W)
0.3 – 0.4:	Barely	(B)
0.4 – 0.5:	Strong	(S)
> 0.5:	Very Strong	(VS)

Table 1 shows categorization of a single document for each query and its associated confidence level. The actual document used for the categorization is listed in the reference section. Category is the JD value that represents the category for that query term.

Table 1: Examples of Categorization

Query Term	Category	Confidence Level
Chronic Bloody Nose [8]	Otolaryngology	S
Dark Circles [9]	Dermatology/ Ophthalmology	B
Ventilator [10]	Anesthesiology	S
Carotid artery surgery [11]	Vascular Diseases/ Neurosurgery/ Brain	W
Cold war [12]	Undefined	U

Results

Table 2 below summarizes the results of the experiments for each of the documents for each of the 5 query terms.

Table 2: Summary of categorization results by confidence levels

Query term	VS	S	B	W	U
Chronic Bloody Nose	5%	10%	30%	40%	15%
Carotid Artery Surgery	0	5%	30%	35%	15%
Dark Circles	35%	15%	20%	20%	10%
Ventilator	5%	30%	35%	25%	5%
Cold war	0	0	15%	25%	60%

The results of the experiment show that for those documents that contained high quality medical content, the system categorized them with a very strong (VS) or a strong (S) confidence level as expected. On review, the majority of the documents categorized with a VS or S confidence level appear to have been correctly categorized. The high quality documents in our set were from NLM's MedlinePlus or Wikipedia. Though some documents from a few commercial web sites were of high value, their resulting scores were low because the sites were peppered with advertisements and other extraneous information that proved difficult to remove. The resulting noise in these documents prevented JDI from generating JDs of high scores for these documents.

In the case of the term "Cold war", though most documents were high quality, they were non-medical in nature. Our system categorized these documents as Undefined as expected. This result can perhaps be used in the future to determine if a document is medical in nature or not.

Evaluation

For evaluation purposes we randomly selected 3 specific documents for each of the 5 terms. Three reviewers independently reviewed these documents to determine whether they agreed or disagreed with the automatic categorization of the system. Table 3 below summarizes the results of the reviewers for each of the 5 queries. An average for each reviewer is also computed.

Table 3: Examples of Categorization

Query Term	Reviewer 1	Reviewer 2	Reviewer 3
Chronic Bloody Nose	100%	100%	100%
Carotid Artery Surgery	67%	33%	100%
Dark Circles	33%	100%	100%
Ventilator	67%	100%	67%
Cold war	100%	100%	100%
Average	**73%**	**87%**	**93%**

The average for Reviewer 1 indicates agreement 73% of the time with the automatic categorization. Similarly, Reviewer 2 agrees 87% of the time, and Reviewer 3 agrees 93% of the time.

The evaluation implies that this is a useful method for automatic categorization of Google search results. As part of our future work, we plan to perform a more robust evaluation of the system with additional documents.

Discussion

In this experiment, several frequently occurring terms were chosen from the logs of a consumer health site (MedlinePlus), that, though determined to be medical, were not found in any of the medical vocabularies from the UMLS or medical dictionaries consulted. We suspect that Google relies on similar resources to determine if a query is medical or not. In addition those building medical vocabularies in particular, consumer health vocabularies are faced with just such an issue. Currently, extensive manual effort is employed to first determine if a given sequence of words is a term or not, and second, whether the term is medical. While there are other techniques [5] that can determine whether a word or sequence of words is a term, retrieving the tessitura, or overall topic range of a document, of each of the top ranked retrieved documents from Google for a term, can aid in the automatic determination of medical termhood. Although not the main thrust of this paper's effort, we are interested in exploring an elegant medical termhood methodology.

Conclusion

In conclusion we found:

1) Based on the results and the evaluations, the use of the JDI methodology may prove a valuable technique for categorizing medical search results from Google.
2) The quality of the categorization may be vastly improved if there were a consistent approach to removing noise from Google documents.
3) High quality medical documents were categorized with a very strong or strong confidence level with high scores. Non-medical documents tended to be categorized as undefined with very low scores.

Future Work

We plan to perform the following as part of our future work:

1) We plan to perform additional experiments expanding the number of input terms.
2) We plan to explore the feasibility of using the JDI-based Semantic Type indexing for categorization.
3) We plan to evaluate whether it is possible to extend our technique for determining the medical or non-medical focus of a document.
4) In our experiments we extracted only the first level documents (depth 0) for a given URL. In the future we want to experiment with extracting the second and third level documents for a given site.
5) We plan to use the origin of the document (NLM, Wikipedia) in determining the confidence level.

Acknowledgments

We would like to acknowledge the following people:
Olivier Bodenreider, Thomas Rindfleisch.

References

[1] Glover, E.J., Tsioutsiouliklis, K., Lawrence, S., Pennock, D.M. and Flake, G.W. Using web structure for classifying and describing web pages *Proceedings of the 11th International Conference on World Wide Web*, ACM Press, Honolulu, Hawaii, USA, 2002.

[2] Kules, B., Kustanowitz, J., Shneiderman, B., (2006 Categorizing Web Search Results into Meaningful and Stable Categories using Fast-Feature Techniques , *Proceedings of the 6th ACM/IEEE-CS Joint Conference on Digital Libraries* (Chapel Hill, NC, USA, June 11 - 15, 2006). JCDL '06. ACM Press, New York, NY. 210-219.

[3] Humphrey, S M, Rogers W J, Kilicoglu H, Demner-Fushman D, Rindflesch T, Word sense disambiguation by selecting the best semantic type based on Journal Descriptor Indexing: preliminary experiment. *J Am Soc Inf Sci Technol 2006* Jan 1;57(1):96-113. Erratum in: *J Am Soc Inf Sci Technol 2006* Mar;57(5):726.

[4] Stolz C, A 10 year Checkup: A Decade Into the E-Health Era, Online Medical Resources Pass a Real –Life Test, *Washington post*, August 1 2006, Page No. HE01.

[5] Wermter J, Hahn U., Effective grading of termhood in biomedical literature. *AMIA Annu Symp Proc*. 2005;:809-13.

[6] http://code.google.com/apis/soapsearch/api_terms.html

[7] http://www.winsite.com/bin/Info?21000000037106

[8] http://www.sleepnet.com/apnea106/messages/162.html

[9] http://www.skinenergizer.com/dark-circles-eyes.htm

[10] http://en.wikipedia.org/wiki/Ventilator

[11] http://health.allrefer.com/health/carotid-artery-surgery-info.html

[12] http://www.mtholyoke.edu/acad/intrel/coldwar.htm

Address for correspondence

Anantha K. Bangalore, bangal@nlm.nih.gov.