MEDINFO 2007
HISA Ltd Australia
© 2007 The authors. All rights reserved.

Automatic Categorization of Google Search Results for Medical Queries using JDI

Anantha K. Bangalore, Guy Divita, Susanne Humphrey, Allen Browne, Karen E. Thorn

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.

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. 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. 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.

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 has been proven effective in characterizing MEDLINE documents. 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. 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)

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 score. 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:

Chronic bloody nose, Carotid artery surgery, Dark circles, Ventilator, Cold war

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. These 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. 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 categorized.
2. Pick the top *n* JDs where the difference between the score of the first JD and the rest of JDs is < 0.1.
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)

Results

Table 1 below summarizes the results of the experiments for each of the documents for each of the 5 query terms.

Table 1 - 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.

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.

Evaluation

For evaluation purposes we randomly selected 3 specific documents for each of the 5 terms. Three reviewers independently reviewed these documents

Table 2 - Summary of reviews by Reviewer

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%**

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. (Ads etc.)
3. High quality medical documents (MedlinePlus, Wikipedia) 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.