Automatic Categorization of Google Search Results for Medical Queries using JDI

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

U.S. NATIONAL LIBRARY OF MEDICINE

Introduction

- Consumers and Health care professionals routinely use Google to lookup medical information.
- No. of documents for any given topic is increasing exponentially on the web.
- Users quickly lose interest, if the relevant results are not in the top 2 or 3 pages.
- Quickly classifying the results into meaningful categories would allow users to drill down into a category of interest.
- Explore the feasibility of using Journal Descriptor Indexing (JDI) methodology developed at NLM for automatic categorization of Google search results for medical queries

Journal Descriptor Indexing (JDI)

- 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
- JDI has proven effective in characterizing MEDLINE documents
- JDI 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

Journal Descriptor Indexing (JDI)

- 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
- A document is indexed by averaging the scores for each JD across the words in the document; these averages become the JD scores for the document
- The JDs are ranked in decreasing order of these scores. That is, the JD profile of a word is treated as a JD vector, and the JD indexing of a document is computed as the centroid of these JD vectors.

Journal Descriptor Indexing (JDI) (example)

- The first 3 JDs for MEDLINE title “Appendectomy in children”
- 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).

1	0.4623	Surgery
2	0.4230	Pediatrics
3	0.3675	Gastroenterology

Methodology

- Extracted 20 documents for 5 terms from Google using Google's SOAP-based application programmer interface (API).
- 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 tool created a JD profile consisting of the top 15 JDs for each of the 100 documents
- Developed heuristics to choose the list of JD's which best categorizes a given document
- Multiple JDs required to classify documents containing information about multiple topics.
- Assigned confidence levels to each of the categorizations within that set

Methodology

- 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.
- Pick the top *n* JDs where the difference between the score of the first JD and the rest of JDs is < 0.1.
- 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)
- Three reviewers independently evaluated the quality of the categorization for 3 of the same documents for each of the 5 query terms.

Results

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%

Summary of categorization results by confidence levels

Evaluation

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%

Summary of reviews by Reviewer

Conclusion

- The results 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.
- Majority of google documents tend to generic in nature, containing information about multiple topics.
- On review, the majority of the documents categorized with a VS or S confidence level appear to have been correctly categorized.
- The quality of the categorization may be vastly improved if there were a consistent approach to removing noise from Google documents. (Ads etc.)
- Non-medical documents tended to be categorized as undefined with very low scores.

Future Work

- Plan to perform additional experiments expanding the number of input terms.
- Plan to explore the feasibility of using the JDI-based Semantic Type indexing for categorization.
- Plan to evaluate whether it is possible to extend our technique for determining the medical or non-medical focus of a document.
- 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.
- We plan to use the origin of the document (NLM, Wikipedia) in determining the confidence level.

References

- 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.
- 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.
- 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.
- 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.

- **Address for correspondence:** Anantha K. Bangalore, bangal@nlm.nih.gov.