
Original Paper

Semantic Classification of Consumer Health Content

Qing T. Zeng, PhD, Jonathan MS; Jonathan Crowell, MS

Harvard Medical School, Decision Systems Group, Brigham and Women's Hospital, Boston, MA, USA

Corresponding Author:

Qing T. Zeng, PhD, Jonathan MS

Decision Systems Group

Brigham and Women's Hospital

75 Francis Street

Boston, MA 02467

USA

Phone: +617 732 7694

Fax: +617 739 3672

Email: qzeng@dsg.harvard.edu

Abstract

Background: While the Semantic Web concept holds considerable promise, it requires that “machine-readable meaning” of Web content be explicitly marked up. Yet accurate and consistent manual annotation of large quantities of consumer health content is infeasible.

Objectives: We set out to develop computerized methods for the task.

Methods: First, we created a partial taxonomy of consumer health information retrieval needs, based on their question, reason, or purpose for doing health information retrieval. Text-based content materials were then processed to extract words, phrases and concepts, which served as features for classification algorithms.

Results: Through 10-fold cross validation, classifiers were successfully trained and evaluated on 3 levels of the taxonomy with accuracy of 92% to 95%.

Conclusions: Automated semantic classification of consumer health content is a promising approach to annotate the genre of content for Semantic Web.

KEYWORDS

Consumer health; text classification; semantic web; information retrieval; information needs

Introduction

A plethora of health information is available on the Web and tens of millions of consumers conduct self-initiated Web-based health information retrieval (HIR) each year. Individual HIR queries, nevertheless, are often unsuccessful. Two of the main obstacles contributing to the failed queries are that: (1) many people are not good at articulating their specific purpose for seeking information; (2) Web contents have complex and non-standardized representations (e.g., free-text) and hidden semantics, which hinder effective and efficient retrieval and utilization. For instance, there are many Web pages providing an overview of asthma which do not explicitly declare themselves as relating to that topic, while many other pages that contain the words “overview” and “asthma” do not address that topic.

Regarding the first obstacle, we have developed a health information query assistant (HIQuA) system to improve query formation [1]. To address the second obstacle, much research is being carried out to create a Semantic Web, which will greatly

facilitate information retrieval, synthesis and utilization by enabling deeper understanding of the content. Although the Semantic Web concept holds considerable promise, it does require the Web pages/sites to explicitly mark up the “machine-readable meaning” of their content. For consumer health content, the effort involved in manually annotating content accurately and consistently would be time-consuming and laborious and require significant knowledge of the domain ontology. Thus, we set out to develop computerized methods for the task.

To create methods for classifying consumer health contents, we created a partial taxonomy of consumer HIR needs, containing 3 levels and 8 classes (excluding the root class). Text-based contents were processed to extract words, phrases and concepts, which served as features for classification algorithms. Using 10-fold cross-validation, the 3 classifiers were trained and evaluated. They demonstrated good accuracy ranging from 92% to 95%.

Background

Consumer HIR

HIR is a widespread practice of health care consumers. A Pew study in 2003 reported “fully 80% of adult Internet users, or about 93 million Americans, have searched for at least one of 16 major health topics online.” [2]. Another Pew report stated that despite the existence of a digital divide, among adult members of low-income and minority populations in California, 45% to 60% have internet access and 70% to 80% of those who have internet access also conduct HIR [3].

Despite the overall success of HIR, there are barriers that make the experience between an average consumer and the available relevant information suboptimal. Most consumers have no training in medicine or information retrieval and thus often are not expert at formulating queries. We and other researchers have found that consumer terminology has lexical, semantic and mental model differences from clinician terminology [4]. The consumer queries have also often been observed to be short and to not characterize well the information needs [5].

Since the Internet is vast, consumer queries typically do not fail by not returning any results, but rather by returning a plethora of results that don't relate to the user's retrieval purpose. The challenge is to make the search process more effective and efficient. One obstacle to achieving effective and efficient search is the fact that the content being searched often has a complex and non-standardized representation with hidden semantics. For instance, by our observation, only a small fraction of the numerous pages dealing with menopause have titles or headings correctly indicating on which aspect(s) of menopause (e.g., diagnosis, treatment, or personal experience) they focus.

Consumer Information Needs

Understanding consumer health information needs is an essential step in devising computerized solutions for HIR. We have noticed from literature searches that as more attention is focused on consumer health, the number of papers on consumer information needs has rapidly increased. The Pew study is the largest study so far on consumer information needs that are associated with the use of the Internet [2]. The 2003 Pew report categorized consumer questions into a dozen topics and determined their associated frequency. Besides the Pew study, some research has been conducted on information needs of specific populations, particularly cancer patients [6-11]. Other research has examined information needs of patients in particular settings such as a doctor office [12]. Many studies utilized surveys or interviews, while Web/mail log data analysis was also used by a few [13].

Nevertheless, there is no standard taxonomy for representing health information needs of consumers or physicians and the different studies used classification schemes of varying granularity, scope and views. Common themes, however, do exist. For instance, disease/problem, procedure and medication are common topics sought. For medications, side effects and contraindications are often of interest to consumers.

Semantic Web

In 2001, Berners-Lee et al published the paper “The Semantic Web – a new form of Web content that is meaningful to computers will unleash a revolution of new possibilities” [14]. This paper eloquently pointed out that the Web in its current state is far more suited for publishing documents for human reading than for automatic processing of data and knowledge. The Semantic Web will require content to be more rigorously represented in eXtensible Markup Language (XML) and the Resource Description Framework (RDF). Standardized ontology will be used to explicitly annotate the meaning of a piece of content. Software agents will then be able to interpret and make inferences based on the content. For instance, an agent may be able to answer the question “what are the risk factors that contribute to female heart disease?” by not only collecting relevant pages for consumer review, but also interpreting their meaning and compiling them into a ranked list of unique items along with some explanations.

Many technical obstacles have to be addressed before we can provide such “wizard” agents for consumers. Adding correct and meaningful metadata to Web contents is the critical need of any Semantic Web. The decentralized nature of the Web, however, makes it very difficult to control the quality of the metadata, and the sheer size of the available content makes it daunting to mark up the content manually. Trust is also a significant concern for researchers: it is predictable that commercial sites will attempt to use metadata to manipulate software agents. In the context of e-commerce, such manipulation may be viewed as legitimate marketing. In the context of HIR, however, misinformation can be quite unethical and dangerous.

Automating the generation of metadata will facilitate the development of the Semantic Web. In the biomedical domain, trained professionals index published literature by hand for the MEDLINE database, though automation of the process has long been researched. Health-related contents on the Web are much more diverse and abundant than the peer-reviewed literature, and it is not feasible to have professional coders creating metadata for them. An automated metadata generator could function independently or assist content developers to mark up materials. It could also be used to validate metadata provided by unfamiliar sources.

Text Annotation

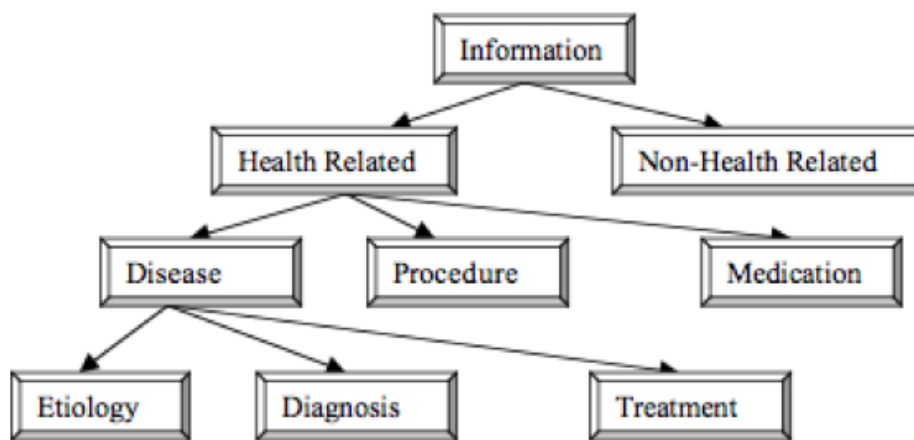
Although there are other types of information on the Web (audio, graphic, video, etc.), query and retrieval are still mostly performed on text contents. To annotate such contents requires free text analysis and understanding, which is a form of natural language processing (NLP). Much of the NLP of clinical reports has focused on key diagnoses and findings [15-17]. NLP has also been used to index medical literature by extracting keywords, particularly MeSH keywords [18].

Text document classification has become an active research area for information retrieval and text routing and filtering in the past decade [19,20]. Outside of the health domain, documents have often been classified by topic and author. In the health domain, classifiers have typically been applied to a

specific type of clinical report (e.g., pathology reports) to detect or infer the presence of specific medical conditions (e.g., pneumonia). To test the methods, large text corpora are established, often using news articles or articles from newsgroups. Example of categories include “graphics”, “hardware”, “religion”, “sports”, “grain” and “stock”. Depending on data sets and classification goals, the state of the art accuracy rate typically runs from low 80% to mid-90% [21-23].

Several different classification methods have been applied to text categorization: Naïve Bayes, Support Vector Machine (SVM), Neural network, KNN, Decision tree, etc. No single method has been shown to be consistently better than others.

Figure . Partial taxonomy of consumer information needs



Health and Non-health Classification

Sample Collection: For health and non-health classification, we semi-randomly gathered 501 web pages of each type manually. A list of health-related and non-health-related terms (e.g., “blood”, “heart”, “basket ball”) were submitted to the Google™ search engine to generate a set of Web sites as starting points, and pages linked to these sites were explored. We excluded any pages that could not be determined, by an informatics researcher without medical background, to be either health-related or non-health-related. Some pages contained mostly links to other pages; the number of such “hub” or “entry” pages was limited to 10 of each type by us because their low text content makes them poor candidates for text-based classification. The resultant pages originated from 426 distinct hosts.

Feature Selection: We focused on semantic type and high frequency word as features and carried out the following steps:

- We parsed the Web pages for free text – scripts, tags, images, etc., were filtered out.
- Free text was split into small chunks along separators. Punctuation marks, numerical strings, and stop words were considered separators.
- Free text chunks of more than one word in length were mapped to one or more UMLS concepts.
- Free text chunks of one word were checked to see if they belonged to the top 1000 frequent words in general English

Naïve Bayes, for instance, is a fairly simple algorithm. Yet it often does surprisingly well in text categorization even when the assumed feature independence does not hold.

Most applications use word (or n-gram) distributions as features to represent the documents. In the health domain, concepts are sometimes used as features instead of words or n-grams.

Methods

For this experiment, we created a partial taxonomy of consumer information needs (Figure 1) and manually collected and classified web contents according to the taxonomy. Classification was performed on 3 levels with accuracy of 92% to 95% percent.

texts. If not, they were mapped to UMLS concepts. (Though some general frequent words are health-related, they are not good features for classification because they lack the distinguishing power.)

- Semantic types of the mapped UMLS concepts were retrieved.
- The number of concepts of a set of health-specific types was counted.
- The number of high frequency health words was counted. (High frequency health words were identified by extracting the top 500 most frequently used words from the MEDLINEplus Web Site and exclude from those the top 1000 frequent general words and non-content words (e.g., “PDF”, “URL”).)
- Both the number of concepts per health-specific type and the number of high frequency health words of a page were normalized by the number of words in that page.

The resultant features are the normalized number of concepts of health-specific types and high frequency health words. The health-specific types we used can be found in Appendix 1.

Classifier: For training and testing of classifiers, we used the Waikato Environment for Knowledge Analysis (WEKA) [24]. After experimenting with different classification algorithms (e.g., Decision Table, C4.5 Decision Tree, SVM, KNN), the Repeated Incremental Pruning to Produce Error Reduction (RIPPER) algorithm with the following parameters was selected:

Number of folds for reduced error pruning: 5; Minimal weights of instances within a split: 2; Number of runs of optimizations: 2; Seed: 1; Use Pruning: True. The RIPPER classifier was trained and tested through 10 fold cross-validation – 10 fold cross-validation selects 90% of the sample for training and 10% for test; the process is repeated 10 times.

Disease, Procedure, and Medication Classification

Sample Collection: For disease, procedure, and medication classification, we semi-randomly gathered 501 pieces of Web content of each type through manual crawling. Because a page sometimes contains information of more than one type, we cut and pasted paragraphs of content (minimum length of 100 letters) from Web pages. Content of ambiguous class was discarded.

The resultant content originated from 1217 distinct URLs and 163 distinct hosts.

Feature Extraction: The feature extraction process was similar to that of health/non-health classification, except that the high-frequency health words count was replaced with the high-frequency disease, procedure and medication words count. The sets of high-frequency disease, procedure and medication word were extracted from contents of these three types on MEDLINEplus. Overlapping high-frequency words of two or three classes were removed from the sets. Concept frequency and word frequency were both normalized according to content length.

Classifier: Using WEKA, we selected an SVM classifier using sequential minimal optimization. The key parameters of the classifier were set to be: Complexity constant C: 1.0; Standardize training data; RBF kernel; Gamma for the RBF kernel: 0.01; Filter: True. The classifier was trained and tested through 10 fold cross-validation.

Etiology, Diagnosis and Treatment Classification

Sample Collection: To perform etiology, diagnosis, and treatment classification under the disease class, sample collection started with the previously collected 501 chunks of disease contents. Extra disease-related texts were collected to reach a larger sample size – 400, 405 and 422 pieces of texts were gathered and labeled as etiology, diagnosis, and treatment respectively. The resultant content originated from 631 distinct URLs and 60 distinct hosts.

Table . 1 Summary of performance statistics

| Total Number of Instances | Correctly Classified | Incorrectly Classified | Kappa statistic | Mean absolute error | Root mean squared error |
|---------------------------|----------------------|------------------------|-----------------|---------------------|-------------------------|
| 1002 | 952 (95.01 %) | 50 (4.99 %) | 0.90 | 0.08 | 0.22 |

Table . 2 Detailed performance by class

| Class | TP Rate | FP Rate | Precision | Recall | F-Measure |
|------------|---------|---------|-----------|--------|-----------|
| Health | 0.96 | 0.06 | 0.94 | 0.96 | 0.95 |
| Non Health | 0.94 | 0.04 | 0.96 | 0.94 | 0.95 |

Disease, Procedure, and Medication Classification

Feature Extraction: Besides the same semantic type features described before, we experimented with the use of n-grams. The high-frequency word feature used in health/non-health and disease/procedure/medication classification is a one-gram. For etiology/diagnosis/treatment classification, we extracted n-grams (n=1, 2, 3, 4) using an open source software program - the Ngram Statistic Package (NSP) [25].

We first collected an extra set of texts of each category, besides the samples for classification. After extracting n-grams (n=1, 2, 3, 4), the top 50 terms for each n value were selected to be signature n-grams of a category. Signature n-grams should only be extracted from training data, not testing data. Because of the use of 10-fold cross validation, all sample data were employed for testing at some point. An additional data set had to be collected to extract signature n-grams. Given that the text set was relatively small (≤ 200 KB per category), a threshold of n-gram occurrence of 3 and above was set for signature n-grams. Overlapping signature n-grams among the categories were removed.

When processing a piece of text in the sample set, extracted n-grams were checked against signature n-grams of each category and matched ones were counted with an assigned weight of n (e.g. n=2 for bigrams, n=3 for trigrams). Thus, for each piece of text, three features of class-specific signature n-gram matching count were extracted.

Classifier: After experimenting with several classification algorithms (RIPPER, Neural Network, SVM, KNN), we chose the SVM classifier with the following parameters: Complexity constant C: 1.0; RBF kernel: False; Exponent: 1; Filter: Normalize training data. The classifier was trained and tested through 10 fold cross-validation.

Results

Three classifiers were trained and evaluated using the method of 10-fold cross-validation. The accuracies were good, ranging from 92% to 95% (Table 1, 3, and 5). Taking a closer look at the classification for each of the 8 classes, the performance was also very consistent (Table 2, 4, and 6).

Health and Non-health Classification

Table . 3 Summary of performance statistics

| Total Number of Instances | Correctly Classified | Incorrectly Classified | Kappa statistic | Mean absolute error | Root mean squared error |
|---------------------------|----------------------|------------------------|-----------------|---------------------|-------------------------|
| 1503 | 1389 (92.42 %) | 114 (7.58 %) | 0.89 | 0.24 | 0.30 |

Table . 4 Detailed performance by class

| Class | TP Rate | FP Rate | Precision | Recall | F-Measure |
|------------|---------|---------|-----------|--------|-----------|
| Disease | 0.92 | 0.05 | 0.90 | 0.92 | 0.91 |
| Procedure | 0.92 | 0.04 | 0.92 | 0.92 | 0.93 |
| Medication | 0.93 | 0.02 | 0.96 | 0.93 | 0.95 |

Etiology, Diagnosis and Treatment Classification

Table . 5 Summary of performance statistics

| Total Number of Instances | Correctly Classified | Incorrectly Classified | Kappa statistic | Mean absolute error | Root mean squared error |
|---------------------------|----------------------|------------------------|-----------------|---------------------|-------------------------|
| 1227 | 1159 (94.46 %) | 68 (5.54 %) | 0.92 | 0.24 | 0.30 |

Table . 6 Detailed performance by class

| Class | TP Rate | FP Rate | Precision | Recall | F-Measure |
|-----------|---------|---------|-----------|--------|-----------|
| Diagnosis | 0.94 | 0.02 | 0.96 | 0.94 | 0.95 |
| Etiology | 0.92 | 0.03 | 0.94 | 0.92 | 0.93 |
| Treatment | 0.97 | 0.03 | 0.94 | 0.97 | 0.96 |

Discussion

This report describes an experiment in automated semantic classification of consumer health contents. As the target for classification, a proof of concept and partial taxonomy of consumer information needs was created. The taxonomy contains 3 levels and 8 classes (excluding the root level). A total of 3732 text samples were collected to train and test 3 classifiers, one for each level. Evaluations results obtained through 10-fold cross validation was very promising, with accuracies = 92-95%.

This experiment demonstrated that automated semantic classification is feasible and can be fairly accurate. As mentioned in the Background, manual annotation of Web content can be unreliable if carried out by individual content providers and not feasible if carried out by certain central authorities. Automated methods for annotating content are thus much to be preferred.

To support intelligent Semantic Web functions, content need to be annotated beyond the basic genres we employed in this paper. Identifying the genres, however, is not only helpful for locating documents for certain information needs, but could also guide further annotations. The type of metadata or keywords one would expect for an article on disease etiology are different from that for an article on medication side effects.

In the area of keyword identification, decades of research have been done [26-29]. Most notable is the work by researchers at the NLM. The Indexing Initiative System combines three different methods in suggesting MeSH main headings: noun phrase parsing, trigram phrase matching and extracting headings

from related citations [18]. Since most of the indexing was conducted on abstracts of literature, it is not clear how successful it would be in providing metadata for full-length Web pages. We postulate that combining genre classification with keyword extraction would provide a large portion of the necessary annotations for the Semantic Web.

We experimented with several types of classifiers and features in this study. No classifier was a clear winner, although SVM generally did well. Concept semantic type and word were effective features. In disease, procedure, and medication classification, multi-word n-grams were also used to enhance the performance.

In this experiment, all text samples were classified by the authors and each sample was annotated by one reviewer. The human annotations, thus, are not really “gold” standards. Partially due to the simplicity of the taxonomy, our decision to classify paragraphs rather than whole documents except for the health/non-health task, and overall good quality of the texts collected, it was relatively easy for us to assign semantic labels. It can be expected, however, that detailed classification of some complex and ambiguous text would be very difficult for even human reviewers.

Although there are numerous ways to classify Web content, the use of consumer information needs taxonomy seems to be a natural one. Given the consumer health focus of our study, it is also important to classify content’s the quality and reading level. Much study has been done on quality and reading level assessment, although there is still a lack of reliable automated methods for health materials. For future work, we are interested

in exploring the benefit of the annotations for HIR. Besides that could better capture the rich semantics of text. genre and keyword, we also hope to study other representations

Acknowledgments

This research is funded by NIH NLM R01 LM007222-05. The authors thank Dr. Robert Greenes for reviewing the manuscript. QTZ designed the study and drafted the manuscript. JC performed most of the sample collection and feature extraction.

References

1. Zeng QT, Crowell J, Plovnick RM, Kim E, Ngo L, Dibble E. Assisting consumer health information retrieval with query recommendations. *J Am Med Inform Assoc* 2006;13(1):80-90.
2. Fox S, Fallows D. Health searches and email have become more commonplace, but there is room for improvement in searches and overall Internet access: PEW INTERNET & AMERICAN LIFE PROJECT; 16 July 2003.
3. Fox S. How Californians compare to the rest of the nation: A case study sponsored by the California HealthCare Foundation: The Pew Internet & American Life Project; 2003 December 14, 2003.
4. Zeng Q, Kogan S, Ash N, Greenes RA, Boxwala AA. Characteristics of consumer terminology for health information retrieval. *Methods Inf Med* 2002;41(4):289-98.
5. Zeng QT, Kogan S, Plovnick RM, Crowell J, Lacroix EM, Greenes RA. Positive attitudes and failed queries: an exploration of the conundrums of consumer health information retrieval. *Int J Med Inform* 2004;73(1):45-55.
6. Nikoletti S, Kristjanson LJ, Tataryn D, McPhee I, Burt L. Information needs and coping styles of primary family caregivers of women following breast cancer surgery. *Oncol Nurs Forum* 2003;30(6):987-96.
7. Pinnock CB, Jones C. Meeting the information needs of Australian men with prostate cancer by way of the internet. *Urology* 2003;61(6):1198-203.
8. Tamburini M, Gangeri L, Brunelli C, Boeri P, Borreani C, Bosisio M, et al. Cancer patients' needs during hospitalisation: a quantitative and qualitative study. *BMC Cancer* 2003;3(1):12.
9. Fukui S. Information needs and the related characteristics of Japanese family caregivers of newly diagnosed patients with cancer. *Cancer Nurs* 2002;25(3):181-6.
10. Houts PS, Ruseas I, Simmonds MA, Hufford DL. Information needs of families of cancer patients: a literature review and recommendations. *J Cancer Educ* 1991;6(4):255-61.
11. Tang PC, Newcomb C, Gorden S, Kreider N. Meeting the information needs of patients: results from a patient focus group. *Proc AMIA Annu Fall Symp* 1997:672-6.
12. Kravitz RL, Bell RA, Franz CE. A taxonomy of requests by patients (TORP): a new system for understanding clinical negotiation in office practice. *J Fam Pract* 1999;48(11):872-8.
13. Shuyler KS, Knight KM. What are patients seeking when they turn to the internet? Qualitative content analysis of questions asked by visitors to an orthopaedics Web site. *J Med Internet Res* 2003;5(4):e24.
14. Berners-Lee T, Hendler J, Lassila O. The Semantic Web - new form of Web content that is meaningful to computers will unleash a revolution of new possibilities. *Scientific American* 2001;284(5):34-43.
15. Fiszman M, Chapman WW, Evans SR, Haug PJ. Automatic identification of pneumonia related concepts on chest x-ray reports. *Proc AMIA Symp* 1999:67-71.
16. Haug PJ, Christensen L, Gundersen M, Clemons B, Koehler S, Bauer K. A natural language parsing system for encoding admitting diagnoses. *Proc AMIA Annu Fall Symp* 1997:814-8.
17. Spyns P. Natural language processing in medicine: an overview. *Methods Inf Med* 1996;35(4-5):285-301.
18. Aronson AR, Bodenreider O, Chang HF, Humphrey SM, Mork JG, Nelson SJ, et al. The NLM Indexing Initiative. *Proc AMIA Symp* 2000:17-21.
19. Park SB, Zhang BT. Large scale unstructured document classification using unlabeled data and syntactic information. In: *Advances in Knowledge Discovery and Data Mining*; 2003. p. 88-99.
20. Vittaut JN, Amini MR, Gallinari P. Learning classification with both labeled and unlabeled data. In: *Machine Learning: Ecml 2002*; 2002. p. 468-479.
21. Sun A, Lim E-P, Ng W-K. Web Classification Using Support Vector Machine. *WIDM'02* 2002:96-99.
22. Yu H, Zhai C, Han J. Text Classification from Positive and Unlabeled Documents. *CIKM'03* 2003:232-239.
23. Peng F, Schuurmans D. Combining Naive Bayes and n-Gram Language Models for Text Classification. *The 25th European Conference on Information Retrieval Research (ECIR) 2003*.
24. Witten IH, Frank E. *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations*: Morgan Kaufmann; 1999.
25. Banerjee S, Pedersen T. The Design, Implementation and Use of the Ngram Statistics Package. In: *The Fourth International Conference on Intelligent Text Processing and Computational Linguistics*; 2003 February; Mexico City; 2003. p. 370-381.
26. Wagner MM. An automatic indexing method for medical documents. *Proc Annu Symp Comput Appl Med Care* 1991:1011-7.

27. Humphrey SM, Miller NE. Knowledge-based indexing of the medical literature: the Indexing Aid Project. *J Am Soc Inf Sci* 1987;38(3):184-96.
28. Hersh WR, Greenes RA. SAPHIRE—an information retrieval system featuring concept matching, automatic indexing, probabilistic retrieval, and hierarchical relationships. *Comput Biomed Res* 1990;23(5):410-25.
29. Fowler J, Maram S, Kouramajian V, Devadhar V. Automated MeSH indexing of the World-Wide Web. *Proc Annu Symp Comput Appl Med Care* 1995:893-7.