Filtering for Medical News Items

Carolyn Watters
Faculty of Computer Science, Dalhousie University, Halifax, Nova Scotia. Canada. B3H 3W5
Correspondent: watters@cs.dal.ca

Wanhong Zheng
Faculty of Computer Science, Dalhousie University, Halifax, Nova Scotia. Canada. B3H 3W5
Email: wzheng@cs.dal.ca

Evangelos Milios
Faculty of Computer Science, Dalhousie University, Halifax, Nova Scotia. Canada. B3H 3W5
Email: eem@cs.dal.ca

In this paper we describe recent work to provide a filtering service for readers interested in medically related news articles from online news sources. The first task is to filter out the nonmedical news articles. The remaining articles, the medically related ones, are then assigned MeSH headings for context and then categorized further by intended audience level (medical expert, medically knowledgeable, no particular medical background needed). The effectiveness goals include both accuracy and efficiency. That is, the process must be robust and efficient enough to scan significant data sets dynamically for the user at the same time as provide accurate results. Our primary effectiveness goal is to provide high accuracy at the medical/ nonmedical filtering step. The secondary concern is the effectiveness of the subsequent grouping of the medical articles into reader groups with MeSH contexts for each paper. While it is relatively easy for people to judge that an article is nonmedical or medical in content it is relatively difficult to judge that any given article is of interest to certain types of readers, based on the medical language used. Consequently the goal is not necessarily to remove articles of higher readership level but rather to provide more information for the reader.

Introduction

Both medical and lay people have an interest in current medical information now available largely in electronic form, including electronic newspapers. Information needs for medical information can be satisfied or triggered by news reports as well as research articles. The number of electronic newspapers has grown from about 40 in 1989 to over 15,000 by 2001 and continues to climb (Abyz, 2001). Many health organizations, including hospitals, universities, government departments, are now providing validated medical information for use by a variety of constituents. The American Medical Association, for example, generates an online medical newspaper (Amednews, 2001) targeted at physicians. Overall, there has been a growing mass of medical news and related data that users, of all levels of sophistication, can access.

In this paper we address two related goals for using online sources for medical news. First, we would like to be able to identify accurately and quickly articles that have health or medical content. Second, we would like to be able to categorize these articles by intended audience, expert to layperson. The categorization results can then be used to further filter or rank the results. We describe recent work to provide a filtering service for users that identifies news items that are medical in nature, and associates articles with intended audience level (medical expert, medically knowledgeable, no particular medical background needed), and assigns MeSH (Medical Subject Headings) that describe the subject matter content of the article. The effectiveness goals include both accuracy and efficiency. That is, the process must be robust and efficient enough to scan significant data sets dynamically for the user at the same time as provide accurate results. Our primary effectiveness goal is to provide high accuracy at the medical/nonmedical filtering step. The secondary concern is the effectiveness of the subsequent grouping of the medical articles. While it is relatively easy to judge that an article is nonmedical or medical in content it is relatively difficult to judge that any given article is only of interest to certain readers. This evaluation is one that falls on a continuum and largely into the "it depends" class and as such we use it as a guideline for the reader rather than for exclusion. Categorization of articles is not straightforward. It is not always obvious when a news article is medically relevant. For example, an article reporting on the sports injury or treatment of an athlete could be categorized as sports or health. For the purposes of this work, we treated these articles as, indeed, medically related.

Background

As in many other areas, the introduction and wide spread adoption of the Internet has provided opportunities for advances in communication of health related information. Many people are now using the Internet to access a wide variety of medical information. This increase in awareness benefits the medical system if the data has validation and comes from respected sources. For physicians sources may be medical journals or online medical news sources, such as amednews.com (amednews, 2001) published by the American Medical Association. For laypeople, sources may be recognizable web sites, such as patientcenters.com or health.yahoo.com, and more probably health related articles in reputable newspapers.

The indexing of medical literature has a rich history and is gaining importance with the rapid growth of online medical and health related information. The UMLS (Unified Medical Language System) Metathesaurus (NLM, 2001) developed by the National Library of Medicine contains concepts and concept names from over sixty different medical vocabularies and medical classification schemes, including the Medical Subject Headings (MeSH). MeSH is the National Library of Medicine's controlled vocabulary within the structure of a hierarchy of subject headings (MeSH, 2001). The incorporation of such a wide range of concepts creates a powerful resource that has been used extensively to improve retrieval results. UMLS and its MetaThesaurus have also been used for query term expansion (Aronson, 1997) and for concept retrieval from medical information data. Success in improved recall has largely been at the expense of decreased precision (Wright et al, 1999). Researchers have also used the UMLS Information Sources Map to facilitate retrieval from multiple sources, where individual sources may be classified using different schema (Abiwajy & Shepherd, 1994; Voorhees et al, 1995; Humphreys et al, 1998). The Metathesaurus has been used extensively for the automatic classification and automatic indexing of medical documents. MetaMap, for example, is a program (Aronson, 2001; Wright et al, 1999) that uses the Metathesaurus to automatically index medical documents by matching noun phrases in the text to the Metathesaurus concepts and then choosing one or more of these concepts to represent the document. Ribeiro-Neto et al (2001) use the International Code of Disease from the World Health Organization to automatically index medical records, such as hospital discharge records. Other approaches to the automatic indexing of medical documents have included machine learning, neural networks, and latent semantic indexing algorithms. Dasigi (SAC'98), for example, used latent semantic indexing with neural network learning to attain indexing effectiveness of 40% with a test medical collection.

The automatic indexing and filtering of news articles for the creation of personalized online news services has been studied for over ten years. Research has shown (Shepherd & Watters, 2001) that fine-grained filtering, based on past behaviour, is not effective for the task of reading news but may be effective for specific information retrieval tasks on this data. Previous research (Carrick & Watters, 1997; Watters & Hang, 2000) found that feature extraction from news articles, such as names, locations, and dates, provides coarse-grained filtering and ranking that users found helpful.

The goal in the current work is to provide effective and efficient coarse-grained filtering by identifying those news articles with medical content and categorizing these roughly by intended readership on the basis of the complexity of the medical concepts within the documents. Our approach is based on identifying a feature set that will provide this level of discriminatory filtering. In the first trial we use the keywords of the documents mapped onto the medical concepts of the MeSH using the terms of the UMLS Metathesaurus as the feature set. In a second trial we employ machine learning techniques to automatically formulate classification criteria on the basis of a training set, in which news articles have been classified by an expert.

Keyword Based Approach

First Iteration: Using the UMLS Metathesaurus

Identification of medical articles based on keyword extraction depends very much on access to a vocabulary that is very specifically medical. Most online medical dictionaries contain a large proportion of words that, although used in medical discussions, are not helpful in discriminating medical articles from sports or financial articles. For example, "management", "back", or "administration" could be used with equal ease in a variety of domains. After considerable exploration of online medical vocabularies, we decided to engage a well known specifically medical vocabulary resource, the UMLS Metathesaurus (NLM, 2001).

We looked first at measuring the effectiveness of using the UMLS Metathesaurus vocabulary to distinguish medical articles from non-medical ones in three pairs of documents, titles shown in Table 1. Two of the groups had a medical article and a non-medical article, while the third group had a chapter from a medical text and a chapter from a programming text. Table 2 shows the percentage of keywords extracted that were UMLS terms.

Table 1. Titles of documents in groups

Group #	**Medical**	**Non-medical**
1	AIDS Treatment Guidelines Revised	Raps win minus Vince
2	First Baby to Receive Heart Transplant	Dot-commers Feel Pain of Withdrawal
3	Kidney and Urinary Tract Tumours and Cancers	C programming Strings

Table 2. Proportion of terms found in UMLS Metathesaurus

	Medical article	**Non-Medical Article**
Group 1	16.22%	17.74%
Group 2	35.09%	24%
Group 3	29.18%	9.41%

In neither of the two groups of news articles was the ratio of medical terms found in medical articles much greater than in the non-medical article. Only the two textbook chapters showed a difference in the occurrence of medical terms that was large enough to be worth pursuing. From examining these results we decided that too many of the UMLS terms occur regularly in non-medical writing and that the frequency of occurrence of UMLS terms in documents does not necessarily, in itself, say much about the content of the document. When reading the articles, however, this result is initially a surprise. No human would miscategorize these articles. Manual examination of terms found in the Metathesaurus led us to the conclusion that the Metathesaurus, by its very nature, is too broad in its coverage. UMLS words are often too general to be useful in discriminating medical content from non-medical content. For example, the following UMLS words occurred in nonmedical news articles: *research, Chicago, employment, jobs, unemployment, insurance, family, work, time, hard, industry, pain,* and *health*. Consequently, for this approach to work we needed a way to identify a subset of medical terms that had better discrimination results.

Second Iteration: Refining the Vocabulary with MeSH

We then hypothesized that we could improve the discrimination value of the terms extracted from the articles by focusing only on those terms that occurred in both the Metathesaurus and the MeSH subject headings. Explorations with the use of the Metathesaurus led us to believe that alone it was not specific enough to provide the high precision we required for this filtering task.

Consequently, we removed terms from the Metathesaurus that were not also MeSH terms. We worked with two raw UMLS sources, MRCON and MRCXT. MRCON contains 1,598,176 terms with the relationship of each of these terms to one of the concept names used in the Metathesaurus, linking all unique variations to the same concept identifier. From this file we created a database of concept names and preferred names to facilitate fast lookups. MRCXT provides the hierarchical context for each of the UMLS concepts, with 11,690,136 entries. The contexts are derived from the hierarchies of each of the source vocabularies, including MeSH. From this file we created a data set of only MeSH concepts and ancestors. Using these derived files we could very quickly identify the MeSH context for each term extracted from the news articles.

We then processed those terms found in the articles that found a match in both the Metathesaurus and the MeSH subject headings. The results from the test articles are shown in Table 3.

Table 3. Proportion of UMLS terms extracted with MeSH context

	Medical	**Non-medical**
Group 1	24%	28%
Group 2	39.58%	26.67%
Group 3	37.84%	44.92%

To further distil the set of specifically medical terms we used two online medical dictionaries to filter out non-medical terms from the original term set. Using a dictionary from MedicineNet.com and another from FastHealth.com, we generated a set of 26,333 medical terms. We used this set to filter the UMLS terms and reran the three document sets. Table 4 shows the percentage of UMLS terms remaining after that filtering.

Table 4. % UMLS terms, after Filtering

	Medical	**Non-medical**
Group 1	6.36%	3.53%
Group 2	16.96%	5.5%
Group 3	20.93%	3.11%

Better, but still not riveting. We recognized that to the main factor in this apparent inability to really discriminate based on the Metathesaurus and MeSH vocabulary matching is that the MeSH headings and hence the Metathesaurus still includes large numbers of terms from areas which are not, actually, medically specific. For example, the term *bankruptcy* produces a MeSH context subtree that includes: Health – Health Care Economics – Economics – Financial. Keyword matching on any of these terms is not helpful.

Third Iteration: Customization of Vocabulary

Since our goal is to be able to filter document terms quickly, we need to have a concise vocabulary of medical terms that have two properties: good discrimination value and match at least one concept in the MeSH. Identifying terms that are most likely to be useful for the filtering task early on has significant advantage in on the fly algorithms. Proportionally more computing can be used in evaluation of a smaller set of more promising candidate concepts at the second stage of the process.

The first phase of the customization included pruning the vocabulary. One of the authors, a medical doctor, developed a customized vocabulary for use in this project by identifying major sections in the MeSH for pruning such as finance, administration, and employment, which have low discrimination value for our purposes. Of the fourteen general categories in the MeSH headings we removed seven; Physical sciences, Anthropology, education, sociology and social phenomenon, Technology and food and beverages, Humanities, Information science, Persons, and Geographic locations. The customized vocabulary was then drawn from the remaining 31,441 headings.

The second phase of the customization was a subjective weighting by the same doctor of the remaining terms. Each term was assigned a weight indicating its value for categorizing articles (1- nonmedical term, 2-lay medical term, 3-general medical term, and 4-specifically medical term). Three categories, shown in Table 5, were roughly targeted to be terms that would be understood by three groups of users: medical specialists, generally medically knowledgeable patients or health care workers, and laypersons.

Table 5. Document Categories

	Examples
Specific medical	Inguinal Canal Peritoneum Douglas Pouch
General medical	Anatomy Umbilicus Pelvis
Lay medical	Body regions Stomach Brain

Filtering Process

The process using the customized vocabulary is relatively straightforward. The keywords from each article are extracted and matched against the customized vocabulary. Terms for which a match is found are determined to be of interest and the corresponding MeSH context is retrieved. A tree structure is created for the article in which each node represents a MeSH category. A filtering algorithm is used first to determine if enough medical content is present to categorize the article as having medical content. This is a simple threshold algorithm using the weights of the terms in the context hierarchy. A classification algorithm is used to categorize the articles with medical content into intended readership levels. This algorithm also uses the assigned weights of the terms along with the relative position in the context tree.

Finally, the context tree for each article is used to determine the most appropriate MeSH categories for the article for additional information for the user. The context tree is traversed recursively to determine the relative weights of the upper nodes with a threshold imposed at the top level.

Test Results

Using a set of seventy electronic newspaper articles from the New York Times, Washington Post and Doctor's Guide, we manually classified each of the articles into four categories: non-medical, medical for general interest, medical for knowledgeable reader, and medical for experts. This process was very subjective with low inter-rater reliability. Each rater was simply asked to interpret the intended audience for each article. The classification task was not based on whether or not only experts could or would read articles categorized as medical for experts but rather that the vocabulary in those items indicated that the article had been written for a specialized audience.

First, we checked to see how reliable the process was in filtering out the non-medical articles. The results of the results for the humans and for the system are shown in Table 6.

Table 6. Classification Results

	Non-med	**Lay**	**Gen**	**Exp**
Human	8	18	4	40
System	7	12	2	33

The system correctly classified 87% of the non-medical articles, 66% of the lay articles, 50% of the general articles, 75% of the expert articles, and 23% of the articles were classified by the system at variance with the human classifier. We are not saying these are incorrect just different. Nonetheless, 77% of the articles were classified the same over the 70 articles by the human and the system.

Since one of our goals is to perform this filtering on the fly from large document sets we examined the relationship between the number of terms, starting at the beginning of the article, used to form the MeSH tree and correctness (i.e., same categorization as human classifier). For this test we chose twenty news articles, five for each category. We then ran the classification process based on a varying number of terms, counted after removing stop words, extracted from the articles. The results are shown in Table 7.

Table 7. Classification on Reduced Term Sets

	First 50 terms	**First 100 terms**	**First 200 terms**	**Full text**
Non-medical	5	5	5	5
Lay	5	5	3	3
General	4	5	4	4
Expert	3	5	5	5

We see that thresholds exist beyond which more terms do not make the result more accurate. In this sample, 100 terms are enough to perform the task accurately. More terms may deteriorate performance perhaps as the accumulative effect of non-medical terms may increase. Also, we see that individual category classification may have individual thresholds. The lack of expert vocabulary in the first 50 terms of an article may be a good indicator that this article either non-medical or lightly medical while the accumulation of expert vocabulary in the article may take longer to differentiate between general and expert levels.

Machine Learning Approach

One of the difficulties of the above approach is formulating classification criteria appropriate for the domain. In the experiments described in the previous sections, we used a human expert to come up with criteria to classify news articles on the basis of frequencies of different term categories. Clearly this is an ad hoc process that involves trying criteria and thresholds based on a number of news articles and adjusting them to improve classification accuracy. In this section, we describe the application of two supervised machine learning techniques, Decision Trees and Naïve Bayes, to automatically formulate classification criteria on the basis of a training set, in which news articles have been classified by an expert.

In supervised machine learning (Mitchell 1997; Witten & Frank 2000), a preclassified data set is available. Preclassification of the data items in the data set is objective, i.e. based on the judgment of a human expert, or from prior knowledge of the origin of the data items. Learning consists of the automatic formulation of classification criteria based on the preclassified data set (training set) for correctly classifying new data items. Data items are described by a set of attributes, while the classification criteria are tests applied to the values of the attributes to decide on the correct class, to which the data item belongs. Different supervised machine learning techniques assume different models about the data, resulting in different algorithms for formulating the classification criteria.

Decision trees follow the "divide-and-conquer" approach. A node of a decision tree typically involves comparing a particular attribute of the data item being classified with a constant. The data item is routed down the tree according to the result of the comparison. A leaf node gives a classification for any data item that reaches it. Learning a decision tree involves choosing the order in which attributes are tested, and the constants against which they are tested at nodes of the tree corresponding to the attributes, so as to maximize the "homogeneity" of the subsets of the training set that fall on the same side of the comparisons at the nodes of the tree.

Classification with a Naïve Bayes classifier involves the calculation of the probability $P(C_i \mid E)$ of the new data item belonging to each one of the possible classes C_i, i=1,2,..,N, given the evidence E provided by the values E_1, E_2, … E_K of its K attributes. The class assigned to the new data item is the one with the highest probability. Learning a Naïve Bayes classifier involves estimating from the training set the probabilities $P(E_j \mid C_i)$ of attribute j having value E_j given that the data item belongs to class C_i. In the classification stage, Bayes' theorem together with the "naïve" assumption that attributes are statistically independent from each other, is used to calculate the probability that the data item whose attributes provide evidence E belongs to class C_i

$$P(C_i \mid E) = P(E_1 \mid C_i) * P(E_2 \mid C_i) * \ldots * P(E_K \mid C_i) * P(C_i) \;/\; P(E)$$

where probability $P(C_i)$ is the prior probability (i.e. before considering any evidence). The probability of the evidence P (E) is not required, as it simply scales all probabilities $P(C_i \mid E)$ and therefore it does not change their ranking. In spite of the naïve assumption of independence, Naïve Bayes classifiers have proved remarkably reliable in text classification tasks.

In practice, it is not possible in general to have learned classifiers that are always correct. Therefore, a learned classifier needs to be evaluated based on the accuracy it achieves on test data, i.e. a preclassified data set that has not been used in learning (or training) the classifier. Since the amount of preclassified data is often limited, setting aside some of it as test data reduces the amount of data available for training. Ten-fold cross-validation is a standard method for addressing the evaluation of a learning method. It consists of breaking the preclassified data into 10 equal disjoint subsets, and using one subset as test data, and the rest as training data. This is repeated 10 times with a different subset as test data each time. The average classification error over the 10 trials is a good estimate of the overall classification error of the learning method.

For the purposes of exploration we simplified the problem to classifying articles into three groups: non-medical, medical intended for experts, and medical intended for other readers. For this experiment we used 302 articles; 100 articles at the expert level from The Doctor's Guide Website (Doctor's Guide, 2001), 102 medical news articles for general readers from the Health Section of Toronto Star (Toronto Star, 2001) and Washington Post (Washington Post, 2001) on line newspapers, plus 100 non-medical news articles from the same papers. The feature set used consists of six features. The first four features are the fraction of Level 1, Level 2, Level 3, Level 4 words in the article respectively. The fifth and sixth features are the fraction of Level 1, Level 2 and Level 3 words combined in the text of the article and in the title of the article respectively.

Results from the Decision tree algorithm demonstrated (on the training set) 80% accuracy of the decision tree overall with 92% accuracy in detecting the non-medical articles. Using this derived decision tree on a separate test sample of 30 online articles from The Doctor's Guide site and Boston Globe, Washington Post and New York Times, the classifier identified all of the medical for expert articles correctly and 60% of the medical for laypersons correctly. On the same test data the Naïve Bayes classifier correctly identified 80% of both expert and layperson articles.

For the non-medical articles both the Decision Tree and Naïve Bayes classifiers identified 9 out of 10 of the non-medical articles correctly and both identified the other non-medical article as being expert level.

Further experiments involved modifications of the feature set, still based on the number of Level 1-4 words, and classification of the training instances by human subjects, instead of the classification by source we used above. No substantial improvements in classification accuracy were observed from these experiments using the cross-validation approach for evaluating the resulting classifiers.

Prototype

A prototype system has been developed to test this approach to retrieving medically relevant articles based on keywords, level of complexity of vocabulary, and MeSH headings. Figure 1 is a sample screen with a result from a search for general medical news showing the MeSH categories assigned to this article, the MeSH sub-tree generated for the article, and confidence in the categorization as general medical. Any of the MeSH fields, keywords, or category of readership level can be used to refine or change the query.

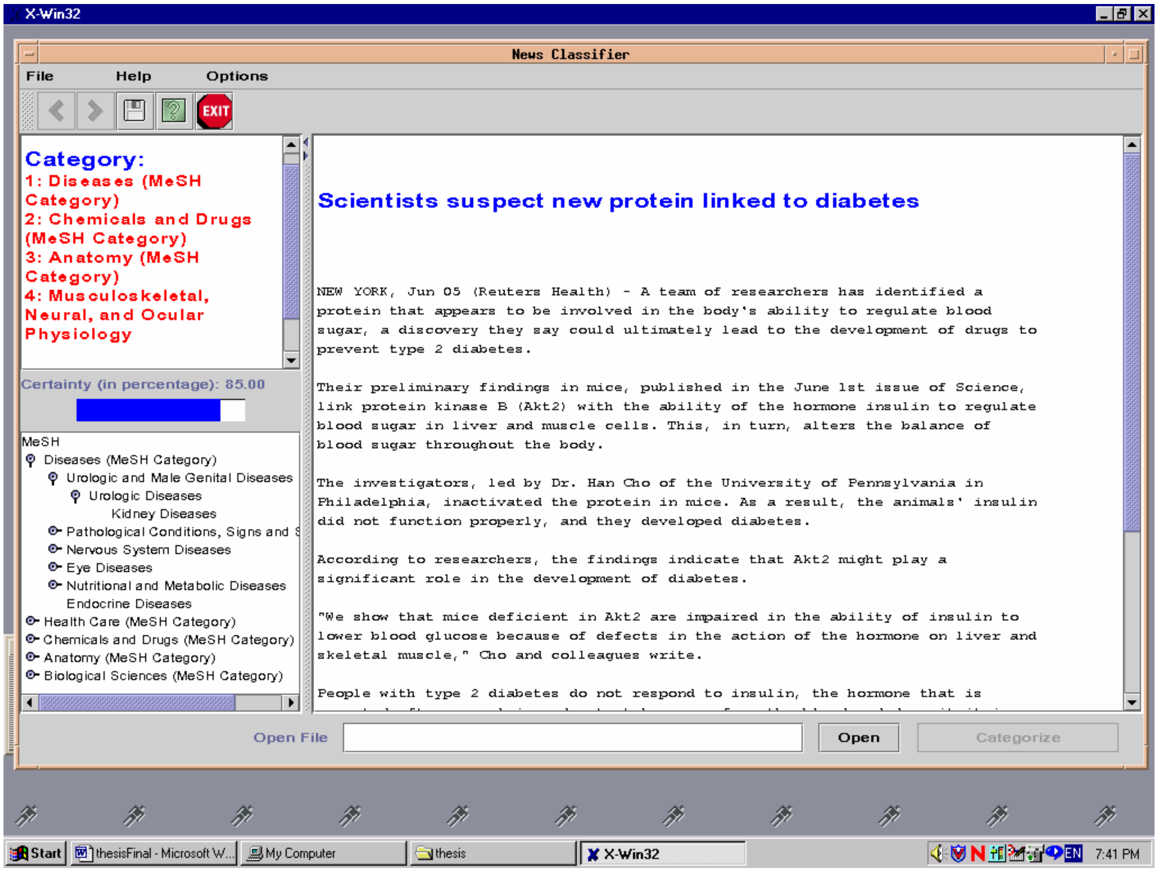

Figure 1. Sample Screen from Prototype

Results

Overall we are able to perform at about the 90% level in separating medical from non-medical articles. In all cases the errors were false positive, i.e., classifying a non-medical article as a medical one rather than missing medical articles.

Using the MeSH concept hierarchy and customized vocabulary provided good results in the determination of categories of medical depth in medical news articles based on simple keyword extraction methodology. There are, of course, several limitations to this approach. First, human input was required in the customization of the MeSH hierarchy and weighting of individual terms. Second, this approach is domain dependent. Third, although the distinction between medical and non-medical is relatively straight forward, the distinctions between intentions of authors are very subjective and fuzzy. Most people can read most, if not all, of articles intended for physicians, especially where definitions are provided. A continuum of complexity might be a better model than strict classification.

Preliminary results from the machine learning approach also provided good results, particularly with the binary classification for medical and non-medical identification.

Acknowledgment

Support for this work was provided by the National Science and Engineering Research Council of Canada.

References

Abyz Web Links. (2001). Online at: [www.abyznewslinks.com] Last Accessed: Dec.5, 2001.

Abiwajy,J. and M.Shepherd. (1994). Framework for the Design of Coupled Knowledge/Database Medical Information Systems. *Proc. Of Annual IEEE Symposium on Computer Based Medical Systems.* Winston-Salem, N.C. June. p 1-5.

Aronson, A.R. (1996). MetaMap: Mapping Text to the UMLS Metathesaurus. Online at: [skr.nlm.nih.gov/papers/references/metamap.fm4.html] Last Accessed: Jan 3, 2002.

Aronson, A.R. (2001). Effective Mapping of Biomedical text to the UMLS Metathesaurus: the MetaMap Program. *Proc of the AMIA Symposium.* Nov.3-7, Washington, DC. p.17-21.

Carrick, C. and C.R. Watters. (1997). Automatic Association of News Items. *Information Processing & Management*. Vol.33, No.5. p. 615-632.

Dasigi, V. (1998). An Experiment in Medical Information Retrieval. *Proc. Of the ACM Symposium on Applied Computing.* Atlanta, Georgia. Feb. 1998. p. 477-481.

Doctor's Guide. (2001). Doctor's Guide Home Page. Online at: [www.docguide.com] Last Access: Sept.9, 2001.

Fasthealth Affilitates. (2001). Fasthealth Affilitates Home Page. Online at: [www.fasthealth.com] Last Access: Sept. 9, 2001.

Humphreys, B., D.Lindberg, H.Schoolman, and G.Barnett. (1998). The Unified Medial Language System: An Informatics Research Collaboration. *Journal of American Medical Informatics Association.* Vol 1. p. 51-60.

Medicinenet. (2001). Medicinenet.com Home Page. Online at: [www.medicinenet.com] Last Access: Sept.9, 2001.

Mitchell, T. (1997) *Machine Learning,* WCB McGraw-Hill, Boston, MA.

NLM, National Library of Medicine. (2001). UMLS Metathesarus. OnLine at: [www.nlm.nih.giv/research/umls/META2.HTML] Last Access: January, 2002.

MeSH. Medical Subject Headings. (2001). OnLine at:[www.nlm.nih.gov/mesh] Available on: Dec 5, 2001.

New York Times. (2001). Online at: [www.nytimes.com] Last Access: Sept 9, 2001]

Ribeiro-Neto, B., A.H.F. Laender, and L.R.S. deLima. (2001). An Experimental Study in Automatically Categorizing Medical Documents. Journal of the American Society for Information Science & Technology. 52(5). P. 391-401.

Shepherd, M., C.Watters, and R.Kaushik. (2001) Lessons from Reading E-News for Browsing the Web: The Roles of Genre and Task. *Proc. of the Annual Conference of the American Society for Information Science and Technology.* November 2001, Washington. p. 256-67.

Toronto Star. (2001). Online at: [www.torontostar.com] Available on: Dec 5, 2001.

Voorhees, E. and R.Tong. (1997). Multiple Search Engines in Database Merging. *Proc. Of the 2nd ACM Conf. On Digital Libraries*. p.93-102.

Watters,C and H.Wang. (2000). Automatic Rating of News Documents for Similarity. *Journal of the American Society for Information Science*, 51(9) 793-804.

Witten, I., E. Frank. (2000). *Data Mining: Practical Machine Learning Tools and*

Techniques with Java Implementations, Morgan Kaufmann, San Francisco, CA.

Wright,L.W., H.K.G.Nardini, A.R.Aronson, and T.C.Rindflesch. (1999). Hierarchical Concept Indexing of Full-Text Documents in the Unified medical Language System Information Sources Map. *Journal of the American Society for Information Science*. **50**(6), p. 512-23.