Association for Information Systems
AIS Electronic Library (AISeL)

AMCIS 2005 Proceedings

Americas Conference on Information Systems (AMCIS)

2005

Medical Information Filtering Using Rule-based and Content-based Approaches

Surendra Sarnikar
University of Arizona, sarnikar@eller.arizona.edu

J. Leon Zhao
University of Arizona, lzhao@eller.arizona.edu

Amar Gupta
University of Arizona, gupta@eller.arizona.edu

Medical Information Filtering Using Rule-based and Content-based Approaches

Surendra Sarnikar
Department of MIS
University of Arizona
sarnikar@eller.arizona.edu

J. Leon Zhao
Department of MIS
University of Arizona
lzhao@eller.arizona.edu

Amar Gupta
Eller College of Management
University of Arizona
gupta@eller.arizona.edu

ABSTRACT

Healthcare professionals need to keep themselves updated with the latest medical developments by finding and reading relevant articles in order to provide the best possible care to their patients. The most popular technique for retrieving relevant articles from a digital library is keyword matching, which is known to retrieve a large amount of irrelevant articles without taking into account the knowledge requirements the user. Currently, the research community is making progress, but is still far from resolving this problem. In this paper, we propose a new method for generating rule-based stereotypical profiles to capture the knowledge requirements based on user roles, and an information filtering technique that combine content-based and rule-based filtering to deliver relevant articles to a user.

Keywords

Medical Information Filtering, Stereotypic profiling, Rule-based profiling

INTRODUCTION

It has been found that access to medical literature has a positive effect on patient care and outcomes (McKibbon and Walker-Dilks, 1995). The most widely used medical literature retrieval system is PubMed. However, due to the wide scope of PubMed, it is difficult to tailor the search results to particular users. Among the thousands of new research papers published every year, one needs to find information that is both related to the user's interest area, and is of relevance to the user's role. Doctors usually find it difficult and time consuming to search for medical literature and frequently rely on medical librarians to search for relevant literature. An alternative approach to information dissemination is information filtering or selective dissemination of information (SDI). Selective dissemination of information has been shown to result in increased productivity in an R&D environment (Mondschein, 1990). In order to deliver relevant and useful information to users it is necessary to accurately model the information requirements of the user. Healthcare professionals have different knowledge requirements based on their roles. For example, a General Physician would be interested in knowledge related to a wide variety of diseases and symptoms that would be helpful for early identification, while a specialist or a sub-specialist would have a narrower focus and would be more interested in deep understanding of a particular phenomena or disease. In this paper, we propose a new user profiling technique and an information filtering process, which together help in automatically identifying the articles that are of high relevance to a user. We use a combination of content-based and rule-based profiles to generate a user model that reflects both the user's specific interests and the role-specific information requirements of the user.

PREVIOUS WORK

A number of user profiling techniques have been proposed to deliver relevant information to users. Although several information filtering techniques have been developed, there has been a shortage of approaches that adapt such techniques to the special requirements and infrastructure available to the medical community. Quintana (1998) proposes a medical information-filtering algorithm that learns a user profile based on pages visited by a user. However, such an approach

requires initial training from the user. An alternative approach is to classify users based on certain characteristics and build stereotypical profiles to model user behavior (Rich, 1998). Research has shown that stereotypical profiles perform better than content-based profiles for novice users (Kulfik et al., 2003). Rule-based stereotypical profiles can be combined with user specific content-based profiles to personalize the results to individual users (Shapira et al., 1997). Another approach to distribute relevant information to users is to combine a network of concepts called a similarity network, and a user interest matrix in to an organizational concept space to aid in personalized distribution of information (Zhao et al., 2000). Initial results have shown that the organizational concept space approach is effective in distributing relevant information to users (Sarnikar et al., 2004). A network of terms with information on relationship between different terms already exists for the medical community in the form of Unified Medical Language System (UMLS). The UMLS consists of a meta-thesaurus and a semantic network. The meta-thesaurus is a large database of life sciences related terms. It consists of different concepts related to life sciences, the terms used to describe those concepts and relationships(narrower, broader) between the concepts. The concepts in the meta-thesaurus are categorized by assigning different semantic types to the concepts. The semantic types and relationships between different semantic types are described in the semantic network. The semantic network consists of 135 different semantic types and 54 types of relationships (UMLS Knowledge Sources, 2004). In this paper we propose a method for generating rule-based stereotypical profiles for medical filtering that exploits the capabilities of the UMLS semantic network and the meta-thesaurus system. The rule-based profile is then used along with a content-based profile to identify articles of high relevance to a user.

MEDICAL INFORMATION FILTERING TECHNIQUE

We propose a medical information filtering technique that uses stereotypical rule-based and content-based user profiles to identify articles of high relevance to the user. An incoming stream of documents is independently analyzed against both the content-based and rule-based user profiles and the relevance judgments from both the processes are combined to determine a relevance score for the documents.

User Profiles

A user profile models the user's interests and information requirements. We use two different kinds of user profiles, content-based and stereotypical rule-based profiles, to model user interest's and information requirements. The user interests are captured in a content-based profile and consists of a vector of terms that are of interest to the user. The user's information requirements are captured in a stereotypical rule-based profile that is based on the user's profession or role in healthcare. The stereotypical rule-based profile is defined in terms of rules describing information preferences for a particular group of users. In order to model the knowledge requirements of different users, a hierarchical set of stereotypical profiles representing different medical professionals is proposed, where child sets inherit characteristics from parent sets.

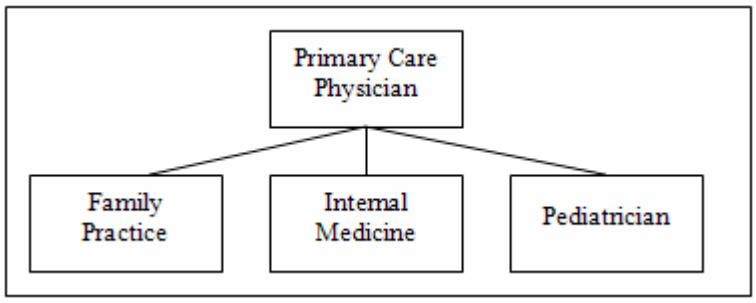

Figure 1. Hierarchical Set of Stereotypical Profiles

A sample hierarchy of user profiles is shown in figure 1. The rule-based user profile specifies the semantic type of the information of interest to the user. For example, primary care physicians are interested in a broad body of knowledge helpful in early diagnosis and prevention of diseases and general health care, while cancer biologists are interested in deep understanding of genetic and molecular processes related to cancer. Also, articles of interest to particular users are usually published in specialized journals such as journal of primary care medicine, journal of clinical nursing etc. Such preferential information is also captured in the stereotypical profile. A rule-based stereotypical profile can be constructed by domain experts and medical librarians who are familiar with information requirements of various healthcare professionals. As the rule-based stereotypical profiles can be reused over a large class of users, they can be constructed economically. A sample rule-based stereotypical user profile for primary care physicians is shown in figure 2.

If STY = <Sign or Symptom> THEN relevance = High
If STY = <Diagnostic procedure> THEN relevance = High
If STY = <Therapeutic or Preventive Procedure> THEN relevance = High
If STY = <Disease or Syndrome> THEN relevance = High
If Journal = <Primary Care> THEN relevance = High
If Journal = <Family Practice> THEN relevance = High

Figure 2. Sample Rule based Profile

Information Filtering Process

The filtering processes involves forming an initial set of documents to be distributed to interested users through a series of tasks: 1) extracting meta-thesaurus concepts from each of the documents; 2) processing these documents against rule-based and content-based profiles; and 3) ranking and routing the documents to the users.

Identifying the Initial Set: The initial set of documents to be distributed can be formed by retrieving recent literature either by query expansion using UMLS and content-based profiles (Aronson, 1997), or by retrieving all the recent publications in a predetermined set of journals.

Filtering Processes: After forming an initial set of documents, the concepts from the articles can be automatically extracted using the MetaMap (MMTx) program (Aronson, 2001). The semantics of the extracted concepts are matched against the rule-based profiles while the meta-thesaurus terms are matched against the content-based profile. A relevance value is then calculated for each of the documents.

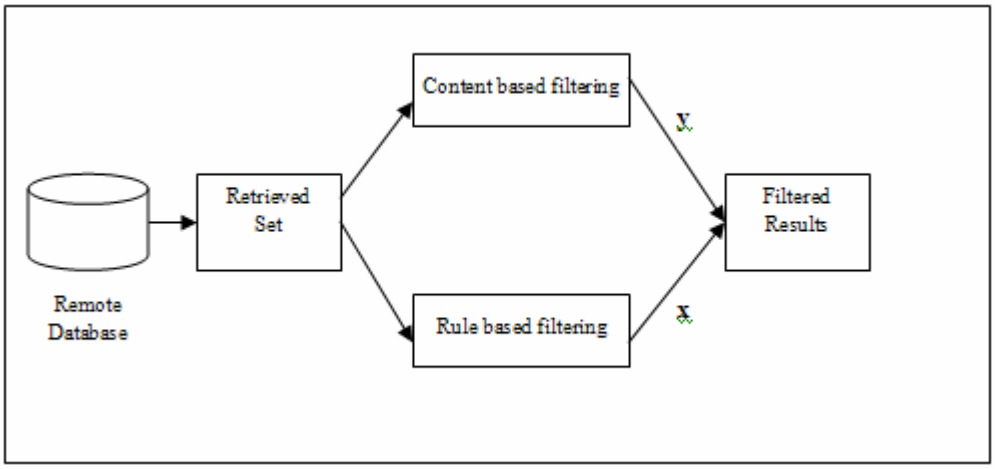

Figure 3. Information Filtering Process

Relevance Calculation: The relevance values from the two matching processes are then combined in the proposed approach, and the articles meeting a minimum threshold are routed to the user. The proposed filtering process is based on Shapira et al.'s (1997) parallel filtering model where rule-based and content-based filtering processes are conducted independently in parallel. However, instead of only selecting documents marked as relevant by both the processes, the final relevance of the document is calculated using a weighted combination of the relevance scores assigned by each processes. This weighting mechanism enables identifying articles that are missed either by the rule-based or the content-based profile. Also, documents that match both the rule-based profiles and content-based profiles would be ranked as highly relevant to the user, while documents that obtain a high relevance score only through one of the filtering processes would be ranked lower but not be eliminated. Documents that have low relevance scores on both the profiles will be filtered out.

PRELIMINARY EVALUATION OF RULE-BASED PROFILES

We intend to conduct a large scale user study to validate the above mentioned information filtering technique. However, although content-based filtering techniques have been used in medical domain, rule-based stereotypical filtering techniques have not been validated. In order to evaluate the effectiveness of the rule-based stereotypical profiles before conducting a full

scale user study, we conducted a simple experiment to test the effectiveness of our approach for constructing rule-based profiles using UMLS semantic types.

We constructed a simple rule-based profile for primary care that focused on knowledge related to identification of symptoms, diagnostic procedures for early detection of diseases and therapeutic or preventive procedures. A multiplying factor of one was specified for *sign or symptom*, *diagnostic procedure*, *therapeutic or preventive procedure* and *disease or syndrome* semantic types. The multiplying factor for all other semantic types was set to zero. Five recent articles on "common cold" were selected from PubMed by querying the MEDLINE database using "common cold" and sorting the results by publish date. The ranking of the articles in the un-sorted response was noted. The UMLS concepts and semantic types in the title and abstract portions of the article were extracted using the MMTx software. As a stop word list was not used during processing the article abstracts, some of the most frequently occurring semantic types like spatial, quantitative semantic types that defined frequently occurring terms in medical literature were not considered in the relevance calculation. The relevance weights of the semantic types were calculated using the tf*idf formula, $r = \frac{nterm}{\max n} Log_2 \frac{N}{n}$, where r is the relevance score, nterm is the number of occurrences of the semantic type, maxn is the number of occurrences of the highest occurring semantic type, N is the total number of documents, n is the number of documents in which the semantic type occurs. A relevance score was calculated for each semantic type by multiplying the tf*idf score of the semantic type by the multiplying factor specified for that semantic type in the rule based profile. The documents were ranked based on their relevance scores.

The results of our preliminary evaluation are summarized below. Table 1 shows the relevance weights of the semantic types in each of the articles. The relevance score for each article is calculated by summing up the relevance weight of the specified semantic types.

Semantic Type	A1	A2	A3	A4	A5
Disease or Syndrome	0.3219	0.0268	0.0000	0.2012	0.0402
Therapeutic or Preventive Procedure	0.2012	0.0134	0.0000	0.0402	0.2414
Sign or Symptom	0.0000	0.7370	0.7370	0.1842	0.0000
Diagnostic Procedure	0.0000	0.0000	0.0000	1.4512	0.0000
Relevance Score	0.5231	0.7772	0.7370	1.8769	0.2817

Table 1. Relevance Scores of Semantic Types in each document

The ranking of the documents based on the rule-based relevance scores and as retrieved by PubMed are shown in Table 2. A brief description of the documents is also given.

Document	M-Rank	R-Rank	Document Description
A1	4	4	"Nitric oxide and the common cold", Describes studies indicating faster recovery from common cold when using Nitric oxide based medications
A2	1	2	Studies the impact of pain states on children and adolescents in the context of common cold and other diseases.
A3	5	3	Study on nasal discharge and its relation to anti-biotic prescription rate for common cold patients
A4	2	1	This articles talks about certain types of cancer being misdiagnosed as common cold, and the onset of symptoms for those cancers.
A5	3	5	Studies the impact of a pesticide in controlling blue mold in cold storage conditions
M-Rank: Order of the document as returned by PubMed R-Rank: Ranking of the document based on rule-based filtering			

Table 2. Document Rankings and Descriptions

Qualitative Analysis

The rule-based filtering approach captures semantic type information from a document. For example, the document A1 which studies the effect of certain medications in treating common cold scores a higher relevance under t*herapeutic or preventive procedure* and zero under *diagnostic procedure* as there is no mention of a diagnostic procedure in the article text. Article A2 has a higher relevance score for *sign or symptom* semantic type as the article discusses various symptoms (pain states), but it has a lower relevance score under *disease or syndrome* as the primary focus of the study is a symptom, in the context of common cold (disease or syndrome). Similarly, article A5 has been assigned a low relevance score as none of the semantic types specified in the rule-based profile are predominant in the article. The article had scored higher relevance under the semantic types of *plant* and *laboratory procedure*. Article A4 scores consistently higher across all the semantic types specified in the profile and has a high relevance score for this profile. Based on the preliminary results, we observe that the rule-based filtering using UMLS semantic types shows considerable promise in extracting semantic information from a document, which can then be used to identify article of high relevance to particular group of users.

CONCLUSION

Our preliminary studies have shown that the rule-based profile can enhance content-based techniques by extracting semantic type information from a document that can be mapped to different kinds of knowledge requirements. We are currently implementing a full-scale system to conduct a user study for a detailed analysis of our approach. In future work, we intend to investigate a number of issues including automating rule-discovery for building rule-based profiles, integration of relevance feedback in to content-based and rule-based profiles, comparison of various ranking and weight adjustment mechanisms, and navigation of related documents using an adaptive navigation system.

REFERENCES

1. Aronson, A. R. (2001) Effective Mapping of Biomedical Text to the UMLS Metathesaurus: The MetaMap Program, *Proceedings of the AMIA Annual Fall Symposium,* Philadelphia, PA.
2. Aronson, A. R. and Rindflesch, T. C. (1997) Query Expansion Using the UMLS Metathesaurus, *Proceedings of the AMIA Symposium,* Philadelphia, PA.
3. Kuflik, T., Shapira, B and Shoval P. (2003) Stereotype-Based versus Personal-Based Filtering Rules in Information Filtering Systems, *Journal of the American Society for Information Science and Technology*, 54 (3), pp243-250
4. McKibbon KA, Walker-Dilks CJ. (1995) The quality and impact of MEDLINE searches performed by end users. *Health Library Review* 12(3):191-200.
5. Mondschein, L. G. (1990), "Selective dissemination of information (SDI): relationship to productivity in the corporate R&D environment", Journal of Documentation, Vol 46 No 2 pp.137-145.
6. Quintana, Y. (1998) Intelligent medical information filtering, *International Journal of Medical Informatics* 51,(2-3), 197-204.
7. Rich, E. (1998) "User modeling via stereotypes" in Readings in intelligent user interfaces, Morgan Kaufmann Publishers Inc., San Francisco, CA.
8. Sarnikar, S., Zhao, L. and Kumar, A. (2004) "Organizational Knowledge Distribution: An Experimental Evaluation", *Proceedings of the Americas Conference on Information Systems (AMCIS)* , New York, NY.
9. Shapira, B., Hanani, U., Raveh, A and Shoval, P. (1997) Information Filtering: A New Two-Phase Model Using Stereotypic User Profiling, *Journal of Intelligent Information Systems*, 8, pp155-165.
10. Zhao, L., Kumar, A., and Stohr, E. (2001) A Dynamic Grouping Technique for Distributing Codified-Knowledge in Large Organizations, *Proceedings of the 10th Workshop on Information Technology and Systems*, Brisbane, Australia.
11. NLM. *UMLS Knowledge Sources*, 2004