# USER MODELLING AND INFORMATION FILTERING FOR CONSUMER HEALTH INFORMATION

**Yuri Quintana**
New Media Lab
University of Western Ontario
London, Ontario, Canada, N6G 1H1
yquint@julian.uwo.ca
http://newmedia.slis.uwo.ca/

## Abstract

This paper describes an information filtering system that creates personalized web pages with new and relevant information for users. The system maintains a model of each user based on web pages visited and relevance rankings made of those pages made by the user. The system has been used with a consumer health information web site. Focus interviews with test subjects were conducted. Among the most important issues raised by test subjects was a sense of loss of control and privacy. Human factors issues and future implications of intelligent agents in the consumer health care sector are discussed.

## 1 INTRODUCTION

The World Wide Web (WWW) has dramatically increased the amount of electronically accessible information. However, the size of the WWW making it increasingly difficult for users to find relevant information. The number of WWW sites and the volume of WWW data being transmitted over the Internet is growing exponentially. The number of users of the Internet is expected to double this year. A directory of World Wide Web sites may have hundreds of new sites added each day. As the amount information that is available on the WWW grows, it will be increasingly difficult to find relevant information.

This paper describes an information filtering system for consumer health information that can reduce the problem of information overload. Examples of health related information on the Internet include information from health organizations, government health departments, newspapers, newsgroups, discussion forums, bibliographies of the health and medical literature, abstracts and full text proceedings of conferences and journals. The quality of health decisions made by consumers and health professionals is to a large degree dependent on the amount, quality, relevance and timeliness of the information that is available to them. Any relevant information that is available on the Internet will be of no benefit to doctors, patients and health professionals if it can not be found in a time efficient way.

The information filtering system that creates personalized web pages with new and relevant information for each user. The system maintains a model of each user based on web pages visited and relevance rankings made of those pages made by the user. The user model is dynamic and automatically updated. The user model is represented using frames that represent topics or concepts of interest. Each web page is indexed using a frame that represents the meaning of the web page using concepts from a concept hierarchy. Information filtering is based on frame matching of the user model frame to frames representing the meaning of web pages using a weighted belief network that self-modifies over time based on the browse and search behaviour of each user.

The paper is organized as follows. The next section describes of current approaches to information filtering on the Internet. The architecture of the system is then presented followed by some observations on the use of the system with consumer health web pages. Some conclusions and future research are then described.

## 2 BACKGROUND

There are several indexing and information retrieval systems that have been developed for medical information, and some systems have been designed specifically for the WWW. One approach to meet the information needs of each user is to create a search profile based on keywords and boolean logic. This profile can then be used to automatically search databases either weekly or daily. However, most users find it difficult to create search profiles based on keywords and boolean algebra, and synonymous search terms are often not used resulting in incomplete searches. Also, the interests of users change over time but users often don't update their search profiles. Traditional keyword search and information filtering methods also have significant problems in the precision and relevance of retrieved information, mainly due to ambiguities in natural language. The databases that are being used for searching are very large and the number of irrelevant items that are sometimes retrieved can be several hundred. Users today don't have the time or desire to personally filter the desired from the undesired information.

MEDLINE is one of the world's largest and most indexed database of medical information and it is main-

0-7803-4053-1/97/$10.00 

tained by the National Library of Medicine (NLM). The Medical Subject Headings (MeSH) developed by the NLM has over 100,000 concepts and has a hierarchy that goes 11 levels deep [8]. Such a classification system is very overwhelming and difficult to use both for end users and indexers of medical information. To assist users in their searching of MEDLINE, the NLM has developed the Grateful Med system, a window-based user interface that allows users to compose queries off-line for MEDLINE, and COACH [6], an expert system for converting keywords and phrases from user queries into UMLS Metathesaurus concepts. However, these systems still require the user to learn the GratefulMed and COACH system. Most users find the learning curve and time requirements discouraging. These systems also don't remember what each user has previously seen, and each query will retrieve all information that matches the query terms. Ideally users would like to see only previously unseen citations.

Most indexes to information on the WWW have a 'What is New' page listing new items. This list can be very long and contains recently added items in all categories. However, users are only interested in seeing the new and previously unseen information on selected topics. The WebWatcher System [1] developed at Carnegie Mellon University allows monitoring of a user's behaviour on the WWW and it suggests related links to WWW pages similar to the one that the user is currently looking at. This system uses a similarity function that is based on keywords and uses a heuristic that is based on the assumption that two pages are of similar interest if some third page points to them both. The WebWatcher system ignores that there are different semantic relations among liked pages and the assumption they use for their similarity among web pages is not very sound. A research group at Stanford is currently developing an 'information brokerage' system to allow an intelligent agents to search the WWW for information [Fikes et.al. 1995]. The system will use a knowledge representation based on KIF [5] and ontologia [4]. This system does not track user's behaviour over multiple sessions, and it does not create personalized summaries of information.

## 3 ARCHITECTURE

An information filtering system has been designed for a World Wide Web server that can index and summarize new information relative to each online user's needs (See Figure 1). The system can be used index information from both internal and external WWW sites. Internal sites can be information that is locally stored on the WWW server, and could include locally added files (such as patient or practice guidelines), or incoming information from listservs, newsgroups, or information from online publishers sent via e-mail. External sites are be identified by a URL (WWW) address. The system manager can specify the Internet sources of information in order to control over the quality of information that is available on the system.

A frame-based knowledge representation is used to representation each document or information object. Frames were chosen because of their wide application to both text [7] and image data [9]. Frames also have object-oriented properties which allow more efficient and effective computation[3]. Each frame contains concepts that define the semantics of meaning of the object. The concepts used in the frames are be based on MeSH subject headings using frames, and the SNOMED (Systematised Nomenclature of Medicine) symbolic representation for patient records [2].

The system tracks the online browse and search behaviour of users and develops a profile of each user's information viewing preferences. The user profile is a frame that contain concepts that have been used to index the web pages viewed by the user. As new information is added to the WWW server, each user will have a personalized web page that ranks the new information based on their most frequently visited concepts. When users connect to the system, a *What's New* page is created dynamically for a particular user. This page shows the number of new items arranged by topic (concept) that have been added since the user's last visit or access to the system. The topics are arranged based on the level of interest that a user is believed to have in each topic, with topics that are of the highest interest displayed at the top. The level of interest (L) in a topic (t) is computed with a heuristic weighted function that takes into account the number of times the topic has been selected s(t), the amount of that has elapsed since a web page with that the topic was last visited, v(t), the number of times the user has indicated a page was relevant with that topic r(t), the number of times the user has indicated a page was not relevant with that topic n(t) .

$$L(t) = w1 \times \times s(t) + w2 \times \times * e^{-v(t)} + w3 \times r(t) - w4 \times n(t)$$

Between browsing sessions, the weights of concepts in any user profile is adjusted based on the number of times each concept appears in the set of web pages viewed by a user in the previous session. In other words, the concepts that appeared in each web page viewed are increased in weight. Information filtering is treated as a frame matching process between a user's profile and the frames representing the medical information using a weighted semantic similarity function.

A knowledge base management systems (KBMS) [10] is used to create and maintain medical knowledge bases. The components of the KBMS are shown in Figure 2. A knowledge base editor allows users to define schemas (implemented as frames). These schemas define the meaning of concepts that are in the concept hierarchy. Frames can refer to other frames and inherit prototypical properties. The knowledge base management system handles the retrieval of concepts and frames, and the inheritance of prototype frames, and the semantic similarity functions needed to do automatic indexing of medical document pages. The automatic indexer parses each web page and determines a set of concepts that can be used to index that web page. These concepts are then compared using a semantic similarity function to frames that describe the meaning of topic categories.

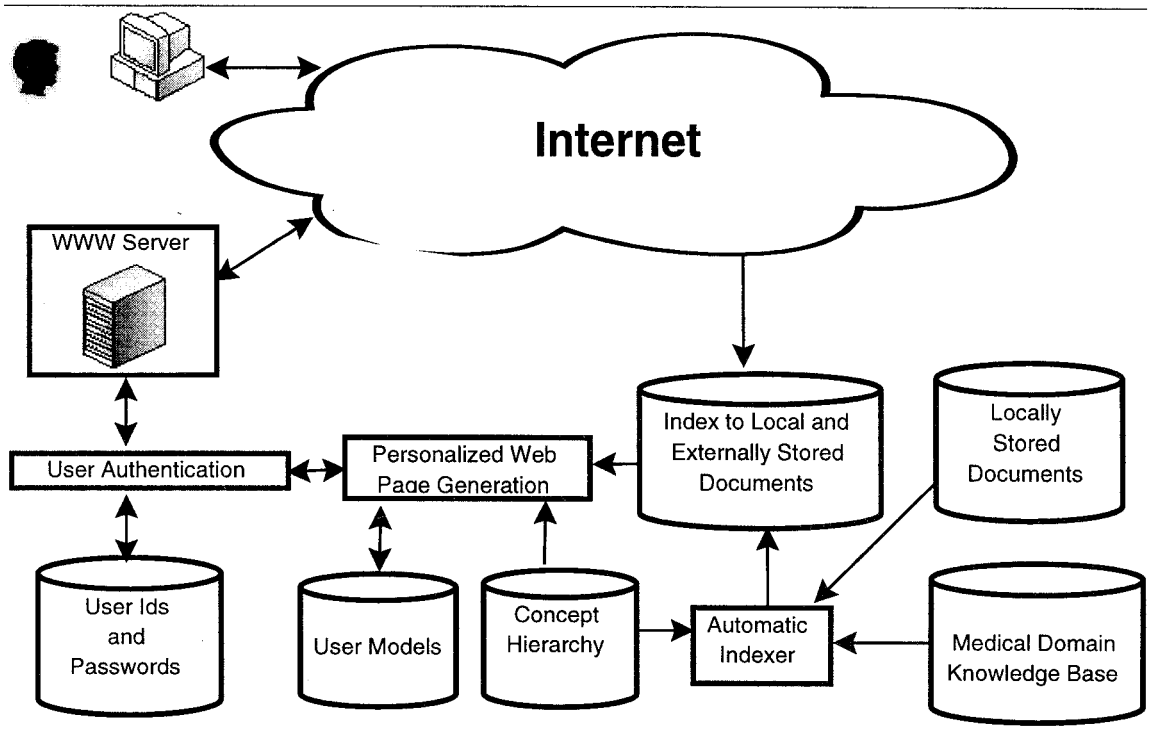


Figure 1: System Architecture

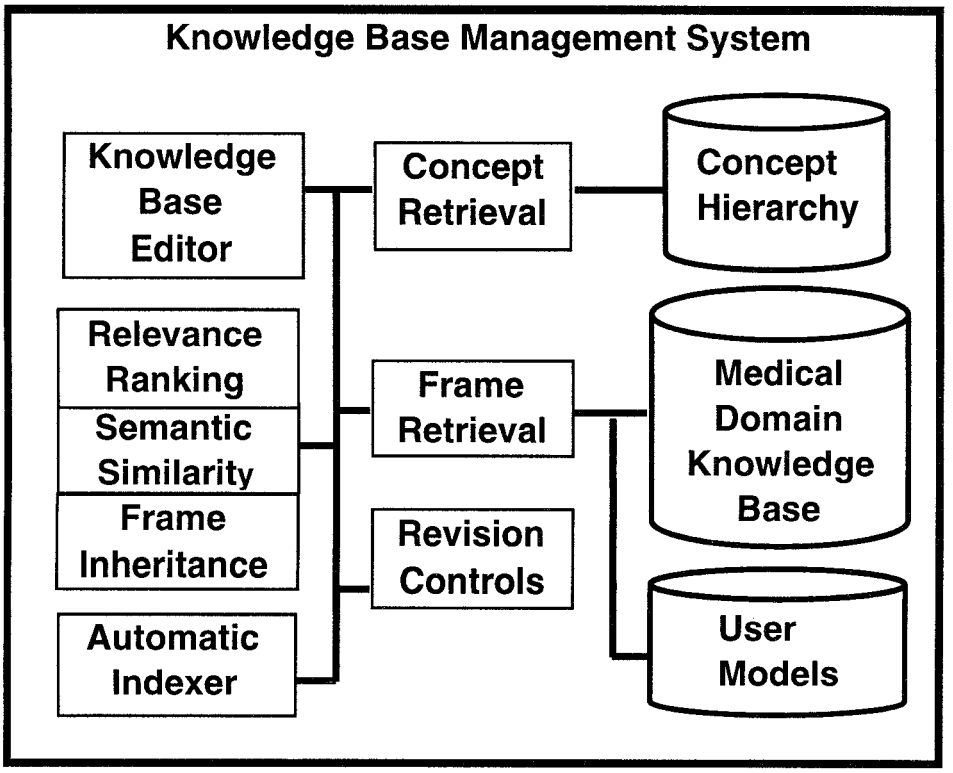


Figure 2: Knowledge Base Management System

## 4 DISCUSSION

The system has been implemented on Windows NT and tested with several hundred web pages on consumer health information. This information filtering system can be viewed as a Selective Dissemination of Information (SDI) service. Unlike previous SDI methods and systems, the proposed approach uses a knowledge-based approach to determine relevant information. Previous SDI methods have been based on keywords which have many problems associated such as ambiguity of natural language that cause errors in accuracy and precision of automatically filtered information. Errors in relevance ranking are be reduced by having both the user profile and medical information represented in the same knowledge representation language.

Among the most important feedback provided by users during the focus interviews is a need to control access and use of the personal profile. Changes have been made to the user interface to allow users to indicate the topics of concepts that are most important to them. These concepts are then given a higher value or score.Users can also explicitly request that the system ignore or lower the weight of other defined concepts which occurred in previously viewed pages. Security of consumer browsing behaviour is especially important for health information and personal privacy. One approach currently being explored is to store the personal profiles on the user of client's computer rather than on the server. The dynamic generation of the *What's New* web page would need to be done on the client computer, and this would require more processing demands on the user's computer.

A preliminary evaluation was conducted in terms of recall, precision. It was found that users often modified their criteria for what they considered relevant not only between browsing sessions but also during a session. This reflects a problem with the standard metrics of recall and precision which are based on the notion that a document is either relevant or irrelevant to the user. Our observations and interviews with users show that a document can be judged to be partially relevant by a user. Hence, new revised metrics of recall and precision are currently being devised to account for the partially relevant judgements. The user interface has also been modified to allow users to indicate how relevant a document is on a scale of one to five, as opposed a binary relevance judgement (i.e. relevant or not relevant). Furthermore, the weights on each concept are modified during browsing sessions, as opposed to after a browsing session is complete. These changes can provide a better indication of the user's relevancy criteria, and allow the system to do relevance rankings during the browsing stage resulting in a more adaptive and responsive system.

## 5 CONCLUSIONS

The system described has contributions to medical informatics, engineering, and library and information science. The system can have a very significant benefit to the health and medical profession. It will reduce the time that doctors, patients and health professionals spend looking for information in repositories of information that are exponentially growing. The knowledge-based management system will facilitate the development of knowledge-based medical information systems and have applications to other knowledge domains as well. Future work will focus on the development of medical knowledge bases using KIF [4; 5] as a standard representation language to share knowledge bases with other research groups in medical informatics.

## References


[1] Armstrong, R, D. Freitag, T. Joachims, and T. Mitchell. (1995) "WebWatcher: A Learning Apprentice for the World Wide Web", AAAI Spring Symposium on Information Gathering from Heterogeneous, Distributed Environments, AAAI Press.

[2] Cote, R.A., and Rothwell, D.J. (1994) "The Systematized Nomenclature of Human and Veterinary Medicine: SNOMED International, Canadian Medical Informatics, Nov./Dec.

[3] Fikes, R., and T. Kehler. (1985) "The Role of Frame Based Representation in Representation and Reasoning", Communications of the ACM, 28 (9), pp. 904-920.

[4] Gruber, T.R. (1992) Ontolingua: "A Mechanism to Support Portable Ontologies. Knowledge Systems Laboratory, KSL-91-66.

[5] Genesereth, M.R., and R. E. Fikes, Editors (1992). "Knowledge Interchange Format - Version 3.0 Reference Manual", Tech. Report Logic-92-1, Computer Science Dept., Stanford University.

[6] Harbout, A.M., E.J. Syed and L.C. Kingsland III. (1993) "The ranking algorithm of the COACH browser for the UMLS Metathesaurus", Seventeenth Annual Symposium on Computer Applications in Medical Care, pp. 720-724.

[7] Hayes, Patrick J. (1979) "The Logic for Frames", In: Frame Conceptions and Text Understanding. D. Metzing, (Ed). Berlin, Germany: Walter de Gruyter and Co., pp. 46-61.

[8] Lowe, H. and G. O. Barnett. (1994) "Understanding and Using the Medical Subject Headings Vocabulary to Perform Literature Searches." JAMA, Vol. 14, pp. 1103- 11

[9] Minsky, M. (1975). "A framework for representing knowledge", In: The Psychology of Computer Vision, P. Winston (Editor), New York: McGraw-Hill, pp. 211-277.

[10] Schmidt, Joachim W. and Costantino Thanos. (1989). Foundations of Knowledge Base Management. Berlin, Germany: Springer-Verlag.