

INTELLIGENT MEDICAL INFORMATION FILTERING

Yuri Quintana

New Media Lab, Faculty of Information and Media Studies

Correspondence to:

The University of Western Ontario
London, Ontario, Canada
N6G 1H1
Tel: 519-661-2111 ext. 8500
Fax: 519-661-3506
yquint@julian.uwo.ca
<http://newmedia.slis.uwo.ca/>

MeSH Terms: Information Services*, Artificial Intelligence*, Algorithms*, Information Storage and Retrieval*, Abstracting and Indexing, Physicians, Human, Software Design.

Keywords: Information Filtering, Physicians, Consumers, Information Needs, Relevance.

SUMMARY

This paper describes an intelligent information filtering systems to assist users to be notified of updates to new and relevant medical information. Among the major problems users face is the large volume of medical information that is generated each day, and the need to filter and retrieve relevant information. The Internet has dramatically increased the amount of electronically accessible medical information and reduced the cost and time needed to publish. The opportunity of the Internet for the medical profession and consumers is to have more information to make decisions and this could potentially lead to better medical decisions and outcomes. However, without the assistance from professional medical librarians, retrieving new and relevant information from databases and the Internet remains a challenge. Many physicians do not have access to the services of a medical librarian. Most physicians indicate on surveys that they do not prefer to retrieve the literature themselves, or visit libraries because of the lack of recent materials, poor organization and indexing of materials, lack of appropriate and available material, and lack of time.

The information filtering system described in this paper records the online web browsing behaviour of each user and creates a user profile of the index terms found on the web pages visited by the user. A relevance-ranking algorithm then matches the user profiles to the index terms of new health care web pages that are added each day. The system creates customized summaries of new information for each user. A user can then connect to the web site to read the new information. Relevance feedback buttons on each page ask the user to rate the usefulness of the page to their immediate information needs. Errors in

relevance ranking are reduced in this system by having both the user profile and medical information represented in the same representation language using a controlled vocabulary. This system also updates the user profiles automatically relieving this burden from the user, but also allows the user to explicitly state preferences.

An initial evaluation of this system was done with health consumers using a web site on consumer health. It was found that users often modified their criteria for what they considered relevant not only between browsing sessions but also during a session. A user's criteria for what is relevant is constantly changing as they interact with the information. New revised metrics of recall and precision are needed to account for the partially relevant judgements and the dynamically changing criteria of users. Future research, development and evaluation of interactive information retrieval systems will need to take into account the user's dynamically changing criteria of relevance.

INTRODUCTION

This paper investigates the problem that doctors and consumers have with keeping up with new and relevant medical information. Among the major problems users face is the large volume of medical information that is generated each day, the need to filter and retrieve relevant information, and the physical limitations of print media (time needed to physically publish material, reproduction and distribution costs). The Internet has dramatically increased the amount of electronically accessible medical information and reduced the cost and time needed to publish. Examples of health related information on the Internet include bibliographies and full text articles from conferences and journals, clinical treatment guidelines, information from health organizations and, newspapers, newsgroups, and discussion forums. The Internet presents both opportunities and problems for the medical profession and health consumers.

The immediate opportunity of the Internet for the medical profession and consumers is to have more information to make decisions and this could potentially lead to better medical decisions and outcomes. In order for this to occur users would need to have reliable and secure access to the Internet, the information would need to be of 'high quality' and be available when it is need. The information would also need to be interpreted and applied correctly. The quality of online medical information remains a problem since there are no limits to who can publish information on the Internet. The lack of a peer review process for publishing information on the Internet also means that inaccurate information will now be

disseminated faster. High quality web sites with peer reviewed and evidence-based information might become more prevalent in response to this quality problem. However 'high quality' is a term that is also dependent on an individual's opinion and biases. Each user's own selection criteria and preferences would need to be taken into account by any software that provides a retrieval function for answering clinical questions or an alerting system for current awareness.

Several studies have shown that the most frequently used sources of information used by physicians for answering clinical questions is colleagues, reference books such as the Physicians Desk Reference (PDR), and retrieval of literature via librarians [1,2,3,4,5,6]. Studies have shown that percentage of materials retrieved by librarians and judged to be relevant to clinical questions by physicians is between forty and sixty percent [7,8] and the cost of this is very high [7]. Many physicians do not have access to the services of a medical librarian. Most physicians indicate on surveys that they do not prefer to retrieve the literature themselves, or visit libraries because of the lack of recent materials, poor organization and indexing of materials, lack of appropriate and available material, and lack of time. [5,6,9]. Studies also show that most physicians do not regularly visit libraries [9]. It has also been shown that the access to materials by physicians depends more on the availability, searchability, understandability, and clinical applicability rather than extensiveness or credibility [9].

The need for physicians to keep up with literature has been argued for [10] and there are many papers that have been written on how physicians should do this [11,12]. However, many of the same problems encountered by physicians in trying to access the medical

literature for answering clinical questions also apply for current awareness efforts. Studies have shown the problems users have in searching medical information online [13]. In the next section, some current methods for Internet-based current awareness are described. The subsequent sections describe an information filtering system that overcomes some of the current problems of the Internet as a current awareness source of information.

CURRENT APPROACHES

The simplest approach for current awareness is to visit web sites that have links to medical information. However, some of the most current web sites are also the largest. Most indexes to information on the WWW have a 'What is New' page listing new items. A large index may have as many as 500 or more new sites added each day. 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. As the amount of information that is available on the WWW grows, it will be increasingly difficult to find relevant information for health related decisions. The number of WWW sites and the volume of WWW data being transmitted over the Internet is growing exponentially and could exceed one billion pages by the year two thousand. 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.

Another 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 the web or 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, which leads to incomplete searches [14,15,16]. Also, the interests of users change over time but users often don't update their search profiles. Traditional keyword filtering methods also have significant problems in the precision and relevance of retrieved information, mainly due to ambiguities in language [17,18]. The databases that are being used for searching are very large and a great number of irrelevant items are often retrieved. Users themselves rarely have the time, desire, or ability to filter the desired from the undesired information.

The National Library of Medicine (NLM) maintains MEDLINE, the world's largest and most indexed database of medical information. The Medical Subject Headings (MeSH) developed by the NLM has over 100,000 concepts and has a hierarchy that goes 11 levels deep [19]. 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 [20], 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.

Another approach to this problem is to use information filtering systems [21] that deliver new information that is believed to be relevant to users based on their explicit

selections or implied by their browsing behaviour. The WebWatcher System [22] 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 similar if some third page points to them both. The WebWatcher system ignores that there are different semantic relations among liked pages and that two pages may be related to a third in different ways. Another system, the Physician Information Customizer [23], builds a customized web page based on a user's response to explicit questions posed by the system. A similar system is the FishWrap system [24] that also builds a customized web page of online news based on user responses to explicit questions. These systems are limited because the questions can only determine the preferences of a user to a limited extent given the quantity and quality of questions. The user profiles are also not dynamically updated over time.

INFORMATION FILTERING

The rationale behind the design of intelligent filtering systems is to assist users to be notified of updates to new and relevant medical information. An Internet information filtering system has been built that has a self-adapting model of each user (See Figure 1). The system can be used to index information on both internal and external WWW sites. Each web page is indexed with one or more topics selected from a controlled vocabulary of topics that have thesaurus relations (narrower than, broader than, related to). Each user has a personal profile maintained by the information filtering system. This profile contains the web pages visited by the user, the topics used to index each web page visited by the user,

and the relevance feedback provided by the user of web pages visited. On each web page, the user has a choice of three relevance judgements (not relevant, somewhat relevant and very relevant). Additional levels of relevancy were found by test subjects to be too confusing. In addition to this, the user can specify in their user profile explicitly one or more topics that they feel are very important and relevant to them. Thus, each user profile that takes into account both their explicitly stated interests and their evolving interests as recorded by the topics assigned web pages they have visited and their relevance feedback.

Insert
figure 1
here



The system creates for each user a customized summary of new health care web pages that are added daily to the web server. Knowledge-based methods are used to generate a customized web page by comparing the topics assigned to new web pages with the dynamically updated user profile. Web pages listed on the customized summary are arranged based on the level of interest that a user is believed to have in each topic. The level of interest (L) in a topic (t) is computed with a heuristic weighted function. This function that takes into account the number of times the topic has been selected $s(t)$, the amount of time 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 very relevant with that topic, $vr(t)$, or somewhat relevant, $sr(t)$. The negative exponent provides a decaying exponential to reduce the level of importance of topics that have not been on any recent web page visited.

This type of informing service is also known as Selective Dissemination of Information (SDI). Unlike previous SDI methods and systems, this information filtering system uses topics that have been carefully assigned to each web page by professional

indexers. Future versions will incorporate an automatic indexing component. Previous SDI filtering 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 reduced in this system by having both the user profile and medical information represented in the same representation language using a controlled vocabulary. This system also updates the user profiles automatically relieving this burden from the user, but also allows the user to explicitly state preferences.

DISCUSSION

An initial evaluation of this system was done with health consumers using a web site on consumer health. A consumer health thesaurus was constructed for this evaluation. A preliminary evaluation of precision was conducted by asking subjects to indicate the relevance of each page visited with the feedback buttons on each web page visited. 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 that 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. New revised metrics of recall and precision are needed to account for the partially relevant judgments and the dynamically changing criteria of users.

Traditional evaluation systems such as TREC [25] are based on the assumption that items are either relevant or not relevant based on an explicit query that specifies the topic of

interest. However, in an interactive browsing environment, users can change their own criteria for what they consider to be relevant. There has been an extensive criticism [26,27,28] of this narrow view of relevance used in traditional evaluations primarily because of disagreements on the principle of relevance [28,29,30,31,32,33]. Harter [26] argues that evaluations should be based on the theories of "psychological relevance" and that users want to find information that alters their perception of a topic and introduces new concepts that are pertinent to their situational needs. Park [31] conducted an empirical study of user relevance judgements that revealed 28 different criteria for relevance judgements provided by the test subjects. Hersh [34] notes that "recall and precision may have serious problems in their external validity, at least as they are usually measured. The controversy is not so much related to whether these concepts are important, as they obviously are, but rather to how they are used and interpreted". Thus, information filtering and delivery systems need to be developed which take into account more user-oriented aspects of relevance and the cognitive aspects of information retrieval process [35,36,37]. In the case of medical information, the authority of the source might be one aspect of relevance. Many other criteria can be discovered by research.

The framework for understanding relevance is based on concept of psychological relevance first proposed by Sperber and Wilson [33] that is a mental model approach for understanding discourse. Harter has applied psychological relevance to information retrieval systems and defined information need as "the current cognitive state of an information seeker [that], as such, is fluid, constantly changing". Thus, users' criteria for what is relevant will be constantly changing as they interact with the information and make

connections with the context in which the information need occurs. He further argues that "references on topic may be less important than relevant references not on topic - references that allow the making of new intellectual connections or cause other cognitive change". Thus, future research, development and evaluation of interactive information retrieval systems will need to take into account the user's dynamically changing criteria of relevance.

CONCLUSIONS

As a result of our experimental findings with the information filtering system described in this paper, research is now being conducted to extend the indexing and filtering beyond just topics and to account for additional aspects of relevance. Additional features that will be indexed are source, author, affiliation, and type of document (guidelines, structured abstracts, randomized control trials). Research is also underway on an indexing tool to assist the classification of documents. A fully automated indexing system will need to deal with the many problems of ambiguity in natural language.

It has been suggested that a system that adapts to a user's preferences and biases may be reinforcing current bad habits in information seeking. In particular, some users may not have tendencies or desires for searching and reading evidence-based information. The system could be altered to recognize these behaviours and suggest additional sources of information that the user may want to consider. However, the issue of either tailoring to the preferences of users or overriding their preferences to cause behaviour change also needs to be examined from an ethical point of view. In particular, users that do not wish to see peer-reviewed, evidence-based information and tend to read only alternative medicine information may not react well to a system that provides them with suggestions to change

their preferences and suggests new web sites of evidenced-based information. More research needs to be done between balancing the privacy of users and bringing about change in their information seeking behaviour.

The system described in this paper has contributions to medical informatics, engineering, and library and information science. The system could potentially have a very significant benefit to the health and medical profession but further work is needed to understand the criteria used for selection of information by physicians and consumers. This system could reduce the time that doctors, patients and health professionals spend looking for information in repositories of information that are exponentially growing. This system also provides a framework for the management, indexing and dissemination of information within organizations and networks of consumers and physicians.

ACKNOWLEDGEMENTS

Partial funding for this research was obtained from HealNet, the Health Evidence and Application Linkage Network, a federally funded Canadian Network of Centres of Excellence. The author also wishes to thank Dr. Jochen Moehr, Dr. Vimla Patel, Dr. Andrew Grant, and their graduate students for their comments, questions and opinions shared at HealNet research exchanges and meetings.

REFERENCES

- [1] Connelly DP, Rich EC, Curley SP, and Kelly JT: Knowledge resource preferences of family physicians, *Journal of Family Practice*, 1990 March, 30(3):353-9.
- [2] Ely JW, Burch RJ, and Vinson DC: The information needs of family physicians: case-

specific clinical. *Journal of Family Practice*, 1992 September, 35:265-9.

[3] Haug, JD: Physician preferences for Information Sources: A Meta-analytical study. *Bulletin of the Medical Library Association*, 1997 July, 85(3): 223-232.

[4] Verhoeven AA, Boerma EJ, and Meyboom-de Jong B: Use of information sources by family physicians: a literature survey. *Bulletin of the Medical Library Association*, 1995 January, 83(1): 85-90.

[5] Covell DG, Uman GC, and Manning PR: Information needs in office practice: are they being met? *Annals of Internal Medicine*, 1985 Oct , 103(4):596-9.

[6] Curley SP, Connelly DP, and Rich EC: Physicians' use of medical knowledge resources: preliminary theoretical framework and findings. *Medical Decision Making*, 1990 October-December , 10(4): 231-41.

[7] Chambliss ML, and Conley J: Answering clinical questions. *Journal of Family Practice*, 1996 August, 43(2):140-144.

[8] Gorman PN, Ash J, and Wykoff L: Can primary care physicians' questions be answered using the medical journal literature? *Bulletin of the Medical Library Association*, 1994 April, 82(2):140-6.

[9] Dee C, and Blazek R: Information needs of the rural physician: a descriptive study, *Bulletin of the Medical Library Association*, 1993 July, 81(3):259-64.

[10] Haynes RB, McKibbon KA, Fitzgerald D, Guyatt GH, Walker CJ, Sackett DL : How to keep up with the medical literature: I. Why try to keep up and how to get started. *Annals of Internal Medicine*, 1986 July, 105(1):149-53.

[11] Wagner JD, and Wagner SA: Keeping abreast of the medical/dental literature: a simplified approach. *Journal of Oral and Maxillofacial Surgery*, 1992 Feb, 50(2):163-8.

[12] Jadad AR, and McQuay HJ: Searching the literature. Be systematic in your searching. *BMJ*, 1993 Jul 3;307(6895):66.

[13] Efthimiadis EN and Afifi M: Population groups: indexing, coverage, and retrieval effectiveness of ethnically related health care issues in health sciences databases. 1996 Jul , *Bulletin of the Medical Library Association*, 84(3):386-96

[14] Harbout, A.M, E.J. Syed and L.C. Kingsland III: The Ranking Algorithm of the COACH Browser for the UMLS Metathesaurus. *Proceedings of the Seventeenth Annual Symposium on Computer Applications in Medical Care*, 1993, 720-724.

- [15] Lowe, H.J., and G.O. Barnett: Understanding and using the Medical Subject Headings (MeSH) Vocabulary to Perform Literature Searches. *Journal of the American Medical Association*, 1993, 14: 1103-11.
- [16] Lowe, H. and G. O. Barnett: Understanding and Using the Medical Subject Headings Vocabulary to Perform Literature Searches", *Journal of the American Medical Association*, 1994, 14:1103- 11.
- [17] Humphrey, S.M: MeIndEx System: Medical Indexing Expert System. *Information Processing and Management*, 1989, 25(1): 73-88.
- [18] Humphreys, B.L. and D.A.B. Lindberg: The UMLS project: Making the Conceptual Connection Between Users and the Information They Need. In *Bulletin of the Medical Library Association*, 1993, 81(2):170-177.
- [19] Lowe, H.J., and G.O. Barnett: Understanding and Using the Medical Subject Headings (MeSH) Vocabulary to Perform Literature Searches." *Journal of the American Medical Association*, 1993,14: 1103-11.
- [20] Harbout, A.M, E.J. Syed and L.C. Kingsland III: The Ranking Algorithm of the COACH Browser for the UMLS Metathesaurus", *Proceedings of the Seventeenth Annual Symposium on Computer Applications in Medical Care*, 1993, 720-724.
- [21] Belkin, N.J. and W.B. Croft: Information Filtering and Information Retrieval: Two Sides of the Same Coin? *Communications of the ACM*, 1992, 35(12): 29-38.
- [22] Armstrong et.al. 1995] Armstrong, R, D. Freitag, T. Joachims, and T. Mitchell: WebWatcher: A Learning Apprentice for the World Wide Web. *AAAI Spring Symposium on Information Gathering from Heterogeneous, Distributed Environments*, 1995, AAAI Press.
- [23] Pratt W; Sim I: Physician's information customizer (PIC): using a shareable user model to filter the medical literature. *Medinfo*, 1995, 8 Pt 2:1447-51
- [24] Chesnais, P.R., M.J. Mucklo, and J.A. Sheena: The Fishwrap Personalized News System," *Second International Workshop on Community Networking*, 1995, IEEE Press.
- [25] Harman, D: Overview of the First TREC Conference". *Proceedings of the Sixteenth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 1993, 36-47.
- [26] Harter, S.P: Psychological Relevance and Information Science. *Journal of the*

American Society for Information Science, 1992, 43(9): 602-615.

[27] Kagolovsky, Y. and J.R. Möhr: A Structured Model for Evaluation of Information Retrieval. Proceedings of MedInfo98, 1998. In Press.

[28] Cool, C., N.J. Belkin, and P.B. Kantor: Characteristics of Texts affecting Relevance Judgements. Proceedings of the National Online Conference, 1993, 77-84.

[29] Harter, S.P: Variations in Relevance Assessments and the Measurement of Retrieval Effectiveness". Journal of the American Society for Information Science, 1996, 47(1): 37-49.

[30] Mizzaro, S.: Relevance: The Whole History. Journal of the American Society for Information Science, 1997, 48(9): 810-832.

[31] Park, T.K: The Nature of Relevance in Information Retrieval: An Empirical Study. Library Quarterly, 1993, 63(3), 318-351.

[32] Saracevic, T: Relevance: A Review of the Literature and Framework for Thinking on the Notion In Information Science. Journal of the American Society for Information Science, 1976, 26: 321-343.

[33] Sperber, D. and D. Wilson: Relevance: Communication and Cognition. 1987, Cambridge, MA: Harvard University Press.

[34] Hersh, WR.: Information Retrieval: A Health Care Perspective. 1996, New York: Springer-Verlag New York, Inc.

[35] Arocha, J.F., V.L. Patel, and Y.C. Patel: Hypothesis Generation and Coordination of Theory and Evidence in Novice Diagnostic Reasoning. Medical Decision Making, 1994, 13: 198-211.

[36] Belkin N.J: Cognitive Models and Information Retrieval". Social Science Information Studies, 1984, 4:111-29.

[37] Patel, V. and J.F. Arrocha: Cognitive Models of Clinical Reasoning and Conceptual Representation. Methods in Information in Medicine, 1995, 34: 47-56.

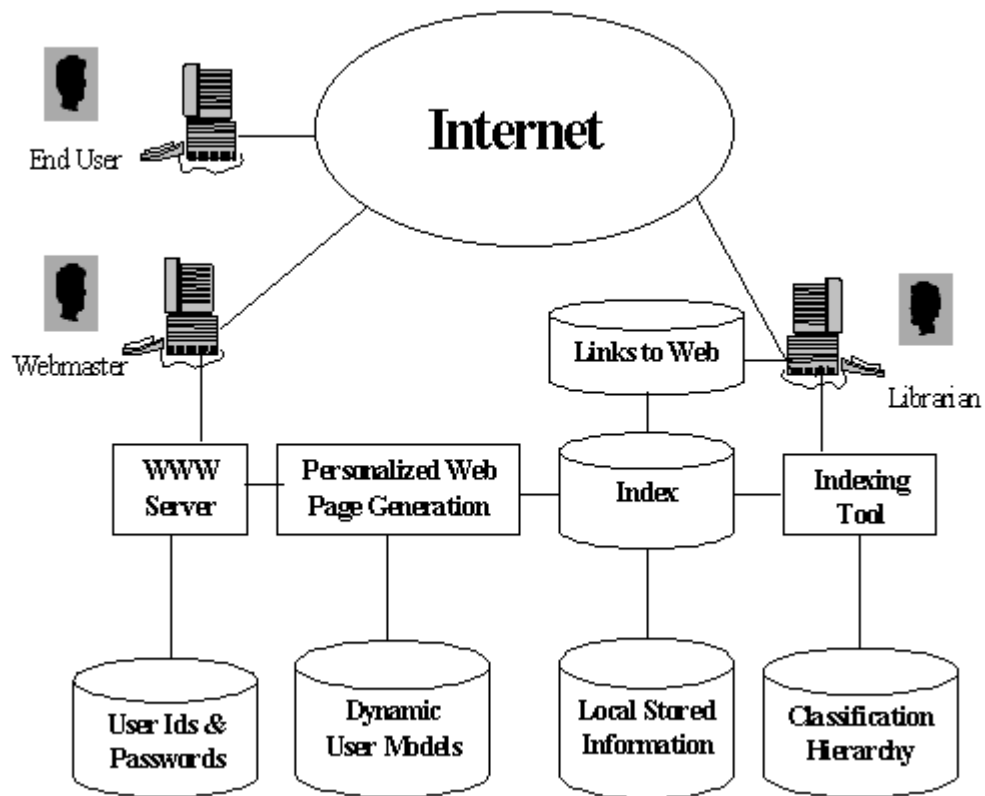


Figure 1: The Information Filtering System