Facilitating Networks of Information

Erik M. van Mulligen[1], Mario Diwersy[2], Martin Schmidt[2], Henk Buurman[3], Barend Mons[4]
[1]Dept. of Medical Informatics, Erasmus University Rotterdam, The Netherlands
[2]SyynX WebSolutions GmbH, Frankfurt, Germany
[3]Collexis, The Netherlands
[4]Netherlands Organization for Scientific Research, The Hague, The Netherlands

In this paper we describe an approach to respond to a request for information with the identification and location of the appropriate person as a source of information for answering the question. The expertise of a person is characterized using a weighted profile that has been derived from a series of documents describing the expert's activities. Having these profiles, requests for information can be matched with these profiles. The best matches correspond with the people that are experts for providing information on the request.

Background
The information society is flooding us with information. Not only the web is drowning us with over a billion pages of information [1], the number of biomedical journals also doubles on the average every 19 years to a total number of about 10.000 in 1980 [2]. Since (web) search engines are yielding tons of web pages for non-specific queries, users are required to become more specific and, as a consequence, loose sensitivity (i.e., missing pieces of information). In addition, the major search engines are only covering at best 34% of all pages, confined as they are in bandwidth and processing power maintaining those large word indexes [3]. There is a clear need for effective information retrieval tools.

Clinical Information Seeking Strategies
With the need for ubiquitous information access, various studies have been performed to analyze how clinicians will find their information. Comparing the reported use of information sources with the observed use showed a discrepancy. Physicians reported to use print resources in 61% and human resources in 32% of the cases. However, the observation showed a different picture: in 53% of the questions human resources were consulted and only in 27% print resources [4]. Unavailability of print resources and the confined time during patient visits to seek for information are the most important reasons why human resources are playing an important role [5,6].

Digital Access to Information
Many projects have been conducted to facilitate and optimize the routine searching and retrieval of information. These projects can be roughly categorized in building information directories, improving the information retrieval facilities, ending up with knowledge management tools.

Information Directories
The most common step in improving access to information is maintaining directories classifying the information by subject. A user browsing the web can add a reference to an interesting web page to a local directory of references. A clear disadvantage of this approach is the fact that only after discovery of the information a bookmark can be set. Dynamic pages (i.e., generated from information in a database) can not be marked and static pages can be moved or deleted, causing broken links.

To support groups of users, the bookmark directory can also be maintained at a server side: the internet portals [7,8]. In this approach, references to information are collected for a common group of users and organized by the provider of the internet portal. Commercial interests may influence the organization and the contents of the portal and the question is whether the aim of the portal is consistent with the user's interests.

User-driven Information retrieval
Another approach is the use of search engines to find information that is interesting. Various search engines have been developed [9,10]. The underlying assumption in search engines is that phrasing a query using a number of words will yield the page(s) containing relevant information. Since most web search engines are generic search engines, (domain specific) synonyms and homonyms will not be dealt with in the search. For routine use, the information retrieved by the generic search engines is not good enough. Typically, a search will yield thousands of hits. However, adding search criteria in order to reduce the number of hits will in general diminish the sensitivity of the query. As a consequence, several projects focus on the development of a preprocessor that will extend the generic search engines with domain knowledge [11,12,13] and thus improve the sensitivity and specificity of the query.

Another approach to improve information retrieval is the active channeling approach. In this approach, one can define subjects of interest and the software will search in the background for information concerning the subjects. The advantage of this approach is that the querying is on the background and is not wasting the user's (web surfing) time.

System-driven Information Management
The approaches discussed above all have a transaction-oriented approach: a query is formulated and the results are fed back. There is hardly any interaction during the search that requires the user to refine and make choices during the information retrieval phase. Agent technology focuses on an approach that at certain moments interacts with the user to refine and optimize its search for information [14,15]. This approach is still a research topic: agent programming languages [16] and agent - user interaction are still research subjects.

Shortcomings
Despite these efforts and developments in information management, these approaches are still not used for daily practice. Typically, entering a non-specific request will yield a large amount of web pages with information that can not be managed quickly, whereas a more specific request will lead to missing the relevant information.

Approach in AWARE
The best source for information remains the human expert. All studies on the information seeking behavior of clinicians show that the best and first source for finding information is the clinical expert or colleague [4,5,6]. The effort to seek information using an expert is only a fraction of searching digital or print information sources. Not only clinicians, but also people in health research, publishing, companies, etc. are seeking for experts that can be consulted to provide answers on a particular request for information. Each of these is confronted with the fact that effective – effective meaning only a small set of relevant pages - information retrieval is still not available.

In the AWARE (Automated Web-based Archive Retrieval and Exchange) system, the answer to a request for information is the name of the source where the information can be found. In the first version of the system, the electronic business cards of experts are shown to the user as a result of a query. In the future, this can be extended to also include online information sources and online databases.

Structure of AWARE
The AWARE system consists of a web-based database with three major sections: institutions, people, and activities. The institutions section gives information on various institutes active in the application field (e.g., for health research in developing countries the institutions conduct research projects in those countries). The institution section links to the people section to indicate what people work in each institute. The people section contains the electronic business cards of people. The people section is linked with the activity section, specifying what activities the user is carrying out. These activities can contain project descriptions, curriculum vitae, scientific publications. In fact, anything that is related to the person's expertise can be entered here. Apart from matching a request for information with the most relevant electronic business cards, the user can also browse the database and follow the links between the various sections (i.e., traverse from organization to the people to the activities and vice versa) .

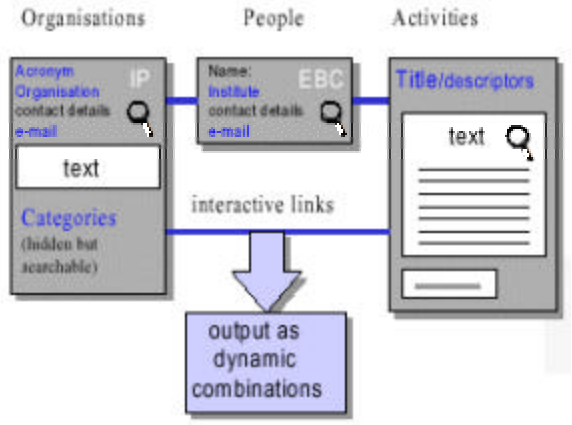


Figure 1. Overview of the different sections of the database. The Organizations, People, and Activities sections are linked and can be browsed, yielding dynamic combinations of these three sections.

Profile information
In order to be able to match a request for information with the information in the database, a categorization or indexing tool has been used. Attached to each person's business card is a list of activities. The text of all these activities is indexed. In this indexing step, the concepts from the Unified Medical Language System (UMLS) [17] that best fit the text are attached to the text including a weight factor. This index can be interactively adjusted by the user. All profiles from the texts attached to a person's business card are combined in a single profile for the user. All profiles of all people working for an institution are combined in a same way to generate the profile of an institution.

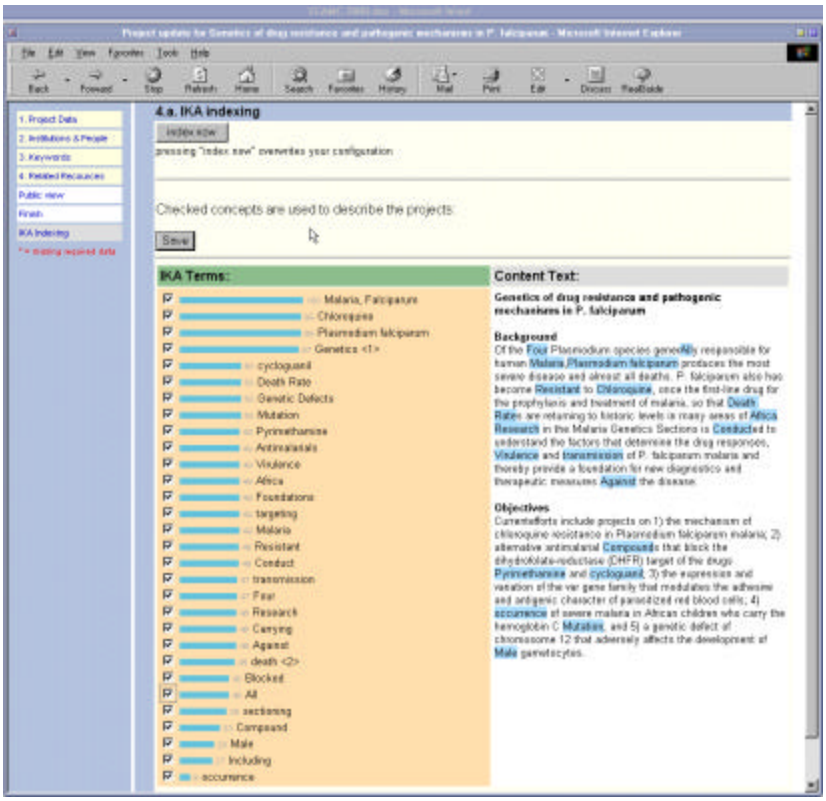


Figure 2. Overview of the AWARE system showing a project text and the weighted index as computed by the indexing software. The terms can

In the running version of AWARE, a request for information can only include a single concept from UMLS. In a next version of the software, the request for information can be entered as a description. The descriptive request will also be indexed and the resulting request profile will be matched against the profiles in the database.

Indexing software

The indexing software combines a statistical indexing algorithm with a rich thesaurus. This thesaurus has been derived from the contents of the UMLS and consists of three parts: concepts, terms (denotations for the concept from the various coding systems), and normalized or stemmed words derived from the words in the terms. The reference from normalized words to concepts is used in the indexing algorithm to find a list of candidate UMLS concepts. The words from the text are clustered: words within a certain distance that refer to the same concept are combined. After the clustering, statistical features are computed for each of the concepts. Finally, on basis of these features the most probable concepts are selected and shown to the user.

The following statistical features are computed:

Per concept:

- *Specificity.* The specificity is indicated by the number of concepts that are referred by a normalized word in the text. The specificity of a cluster is defined as the minimal specificity for all individual words contained in the cluster. A concept is more probable if it contains a word that is only rarely used in the vocabulary and the text contains that particular word.
- *Similarity.* The textual similarity feature indicates what fraction of the concept name is contained in the text. A concept is more probable if a large part of its term is covered by words from the text.
- *Co-occurrence.* The co-occurrence measure indicates how often a concept co-occurs (based on MedLine and other knowledge sources) with other concepts found in the document. If two candidate concepts are often used in combination, their probabilities will rise.

Per cluster:

- *Dispersion.* The dispersion is calculated per cluster. It is the mean distance in words between words referring to one concept

(since a cluster contains concepts with almost identical words, this measure is calculated per cluster). If there is only one word referring to the concept, the distance is set to the number of words in the document (this negatively discriminates concepts that are only referred by one word). The rationale behind dispersion is that a concept is more likely to be present if the words referring to it are close to each other.

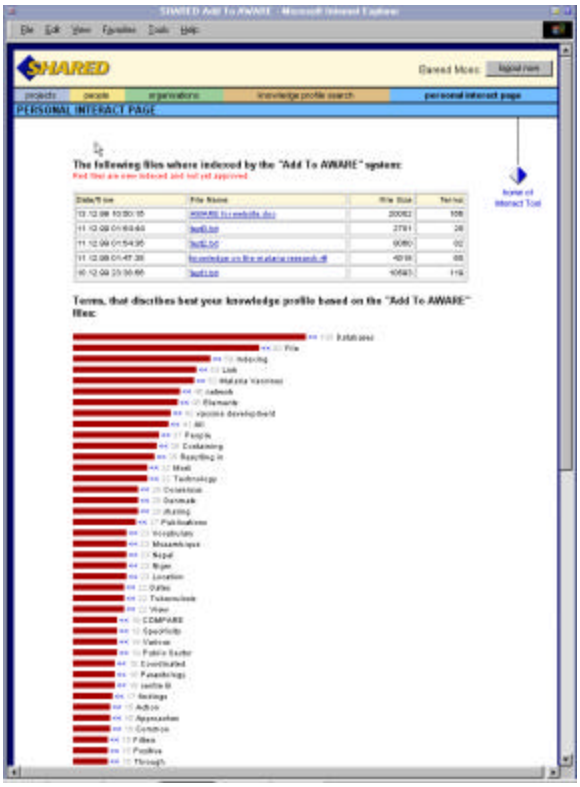
- *Cluster Size.* The cluster size is the number of concepts/alternatives in a cluster. A concept is less probable if there are many alternative concepts in the cluster.

Per document:

- *Frequency.* The frequency is the number of words or clusters referring to one concept. A concept is more probable if the text contains multiple references to that concept.

Add-to-AWARE button

In order to facilitate the uploading of documents or texts to the AWARE system, a button has been defined that can be included on the menu-bar of Microsoft Internet Explorer. With this button, the text contained in the Internet browser will be sent to the AWARE web-server, indexed, and the text profile will be added to the user's personal profile. Since this button is available in the user's desktop environment, the effort to maintain one's profile is minimal. Figure 3 shows the web-page that lists the documents that have been uploaded to the AWARE system. The terms that have been assigned by the indexing software can be manually reviewed and, if necessary, adjusted. If the terms are adjusted, the profile is automatically updated.



gure 3. This window shows the documents that ve been uploaded with the "add-to-aware" tton and the resulting personal profile.

Implementation

This Aware system has been implemented in an application for health researchers in developing countries. With this system, an information network of experts in health research in developing countries has been created that can be contacted for specific information. The profiles in this Aware-based system SHARED (Scientists for Health And REsearch Development, www.shared.de) are based on health research project descriptions in developing countries. The system has been developed as a light-weight system, giving people in developing countries the possibility to quickly access the web-based database and locating the most appropriate source of information.

Evaluation

The indexing software has been evaluated using a collection of about 1000 MedLine abstracts with their manually assigned MeSH headings. The indexing software is able to correctly identify 85% of the MeSH headings. Baring in mind the fact that the manual assignment of MeSH headings is based in the full article text rather than on the abstract and the fact that the indexing software has been limited to only MeSH headings rather than the whole UMLS, the results are quite well. A paper describing these results has been submitted for publication. A formal evaluation of the complete system has not yet been done, but will be started soon.

Discussion

Information delivery is becoming more and more important. However, the current electronic facilities are still not effective enough to be used for daily practice. Therefore, we have focused on information mediation, i.e., on identifying the most promising source for finding that information. Our first approach is aiming at developing a network of experts that can be consulted for requests of information. This concept has been implemented for health research and development in developing countries. However, these networks of information sources can also be beneficial for health care in

developed countries, for scientists who want to find other experts in the field, etc.

The availability of the add-to-aware button on the user's desktop minimizes the burden of maintaining the personal profile in the database. We are currently exploring ways in health science to create a starting personal profile from on-line available scientific publications.

For the near future, we will further improve the indexing software and improve the web-based database access. We also extend the information network with the location of print (web-based) information sources.

References

1 Keslik J, Kleine S. Web Surpasses One Billion Documents. http://www.inktomi.com/new/press/billion.html

2 Wyatt J. Use and sources of medical knowledge. The Lancet 1991;338:1368-73.

3 Lawrence S, Giles CL. Searching the World Wide Web. Science 1998;280(5360):98-100.

4 Covell DG, Gwen C, Uman RN, Manning PhR. Information Needs in Office Practice: Are They Being Met? Annals of Internal Medicine 1985;103:596-9.

5 Smith R. What Clinical information do doctors need? BMJ 1996;313:1062-8

6 Hall EF. Physical Therapists in Private Practice: Information Sources and Information Needs. Bull Med Libr Assoc 1995;83(2):196-201.

7 Garofalakis MN, Ramaswamy S, Rastogi R, Shim K. Of Crawlers, Portals, Mice, and Men: Is there more to Mining the Web? Proceedings of ACM SIGMOD'99, Philadelphia, Pennsylvania, 1999:504.

8 Hersh WR, Brown KE, Donohoe LC, et al. CliniWeb: Managing Clinical Information on the World Wide Web. J Am Med Inform Assoc. 1996;3:273-80.

9 Giles CL, Lawrence S. Accessibility of information on the web. Nature 1999;400:107-109.

10 Lawrence S, Giles CL. Searching the World Wide Web. Science 1998;280:98-100.

11 Joubert M, Fieschi M, Robert JJ, Volot F, Fieschi D. UMLS-based Conceptual Queries to Biomedical Information Databases: An Overview of the Project ARIANE. J Am Med Inform Assoc 1998;5(1):52-61.

12 Cousins SB, Silverstein JC, Fris se ME. Query Networks for Medical Information Retrieval – Assigning Probabilistic Relationships. In: Miller RA, ed. Proceedings of the Fourteenth Annual Symposium on Computer Applications in Medical Care. Los Alamitos CA: IEEE Computer Society Press. 1990:800-4.

13 Detmer WM, Barnett GO, Hersh WR. MedWeaver: integrating decision support, literature searching, and Web exploration using the UMLS MetaThesaurus. In: Proceedings of the AMIA Annual Fall Symposium. 1997:490-4.

14 Genesereth MR, Ketchpel SP. Soft ware Agents. Comm ACM 1994;37(7):48-53.

15 Bollacker KD, Lawrence S, Giles CL. CiteSeer: An Autonomous Web Agent for Automatic Retrieval and Identification of Interesting Publications. Proceedings of the 2nd International Conference on Autonomous Agents. ACM Press, New York, 1998:116-23.

16 Mendelzon AO, Mihaila GA, Milo T. Querying the World Wide Web. Int J Dig Libr 1997;1(1): 54-67.

17 Lindberg DAB, Humphreys BL, McCray AT. The unified medical language system. Meth Inf Med, 32:281-91, 1993.

Address for correspondence:
Erik M. van Mulligen
Dept. of Medical Informatics,
Erasmus University Rotterdam,
P.O. Box 1738, 3000 DR, Rotterdam
The Netherlands
vanmulligen@mi.fgg.eur.nl