DOCUMENT RESUME

ED 429 549 IR 019 457

AUTHOR Tsinakos, Avgoustos A.; Margaritis, Kostantinos G.
TITLE On the Use of Librarians Selection Routines in Web Search.
PUB DATE 1997-11-00
NOTE 8p.; In: WebNet 97 World Conference of the WWW, Internet & Intranet Proceedings (2nd, Toronto, Canada, November 1-5, 1997); see IR 019 434.
PUB TYPE Reports - Descriptive (141) -- Speeches/Meeting Papers (150)
EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Access to Information; *Computer System Design; Databases; Expert Systems; Foreign Countries; Information Retrieval; Librarians; *Online Searching; Optical Data Disks; Search Intermediaries; Search Strategies; User Needs (Information); *World Wide Web
IDENTIFIERS *Search Engines

ABSTRACT

Information retrieval on the World Wide Web has a major obstacle: although data is abundant, it is unlabeled and randomly indexed. This paper discusses the implementation of a consultative Web search engine that minimizes the expertise level that is required from a user to accomplish an advanced search session. The system takes advantage of the meta-knowledge (Selection Routine), used by expert librarian searchers and applies it to a heterogeneous search space such as CD-ROM databases and Web-based environments acting as an intermediary expert system. Topics discussed include: (1) the Selection Routine, including the three basic stages of a typical online search--definition of query structure, selection of search keys, and feedback review; (2) system description, including the Web-based interface, spell checker, consultative core, and retrieval component, as well as examples of the Selection Routine and Metaknowledge rule sets; and (3) implementation issues. Three figures present components of an online search, the decision tree of the Selection Routine, and system description. (AEF)

On the Use of Librarians Selection Routines in Web Search.

Avgoustos. A. Tsinakos

Kostantinos. G. Margaritis

Department of Informatics
University of Macedonia
54006 Thessaloniki
GREECE
Tel : +30-31- 891 891
E-mail: tsinakos/kmarg@kirki.it.uom.gr

Abstract: Information retrieval on the Web has a major obstacle: although data is abundant, it is unlabeled and randomly indexed. This paper discusses the implementation of a consultative Web search engine that minimizes the expertise level that is required from a user in order the latter to accomplish an advance search session. The system takes advantage of the meta-knowledge (Selection Routine), used by expert librarian searchers and apply it to a heterogeneous search space such as CD-ROM Data Bases and WWW based environments acting as an intermediary expert system.

Introduction

Due to the rapid growth of data, taming the information relaying on the Internet repositories has become a difficult and time consuming process. Beyond the Web based search engines such as Lycos and AltaVista that have appeared on the net, some other sophisticated mechanisms have been developed, to confront with the problem of information retrieval. ALIWEB (Archie-Like Indexing in the WEB) [Koster 1994], GENVL (an interactive hierarchical system for cataloguing Web resources in a sense of "Virtual Libraries") WWW Worm - (a resource location tool) [McBryan, 1994] are some representative examples.
Recently several new software products have emerged on the Web space. The common target of those products is to reduce the user effort during a information retrieval session on one hand, and on the other, to increase the productivity and accuracy of the retrieval process using AI and parallel searching techniques. Intelligent Agents used by the MORE LIKE THIS [MORE LIKE THIS] and AUTONOMY [AUTONOMY] products, are trained by the user and released in web space in order to locate an derive the requested term-concept. Additionally meta-search engines such as WEB COMPASS 2.O [WEB COMPASS], MetaCrawler [MetaCrawler] and ECHO SEARCH [ECHO SEARCH], are applying parallel searching on pre-selected Web based search engines and filter the retrieval set by eliminating the duplicates. Web miners are another category of information retrieval systems that relay on a combination of test queries and domain specific knowledge to automatically learn descriptions of Web services such as product catalogues or personal directories. Internet Learning Agent (learns to extract information from unfamiliar resources by queering them with familiar objects) [Perkowitz & Etzioni, 1995] and Shopbot (learns to extract product information from Web vendors) [Doorenbos et al, 1996] are such systems. Internet Softbot can automatically extract information or learned descriptions collected by such intelligent agents [Etzioni, 1994]. Some of the latest products in knowledge-based information retrieval technology for the WWW are: FAQFinder which is an automated question-answering system that uses the FAQ files which are associated with many USENET newsgroups, in combination with FindMe and RentMe systems (market search agents) [Bruke & Hammond et al 1997]. The increasing use of AI techniques in the information retrieval process, reveal a new tendency and need for more intelligent and flexible systems with a high degree of search expertise regarding the procedural and declarative knowledge, in order to perform a search task.
Similar requirements have been outlined in the area of database and on-line search, by the catalogers and reference librarians. There is a number of inherent characteristics of on-line catalogs that make them difficult

to use, especially when someone was seeking subject information. [Bates, 1972] [Borgman 1986] [Connel, 1991]. The identification and characterization of the knowledge used by experienced librarians during a subject searching process in on-line catalogs, is considered an important topic for investigation since an understanding of the specialized knowledge used by the librarians may facilitate the design of more usable systems [Connel 1995]. Tackling this problem, a number of systems had been implemented like Source Finder [Bailey 1992] and Reference Expert [Myers 1994].

In this paper an effort to combine the needs of both librarian reference search and Internet information retrieval, is attempted. The main idea of the system that will be discussed is its intelligence of taking advantage of the meta-knowledge called "Selection Routine", used by expert librarian searchers, in order to construct a search plan. Furthermore this plan will be applied on heterogeneous search spaces such as Data Bases and WWW based environments. An analysis of the rules which consist the Selection Routine and system's architecture for the co-operation of system's core and the retrieval mechanisms follows.

Selection Routine

The intellectual components of a typical on-line search can be analised in to three basic stages. I) The definition of query structure stage II) The selection of search keys stage and III) The feedback review stage [Fig 1].

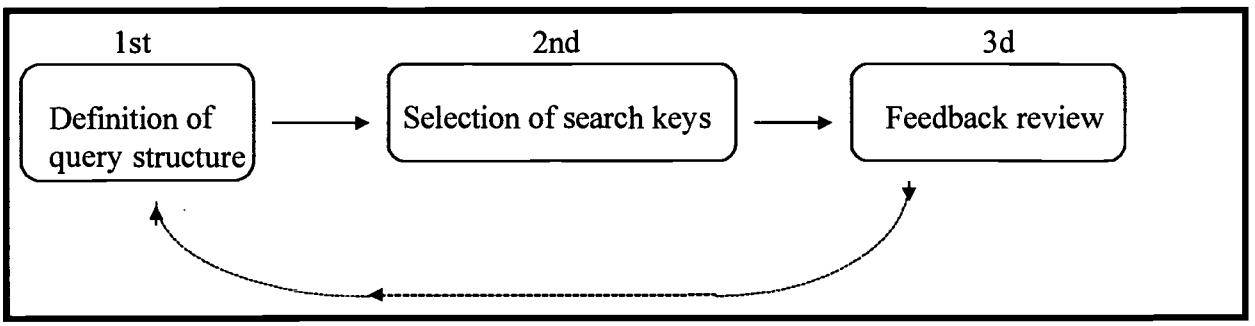
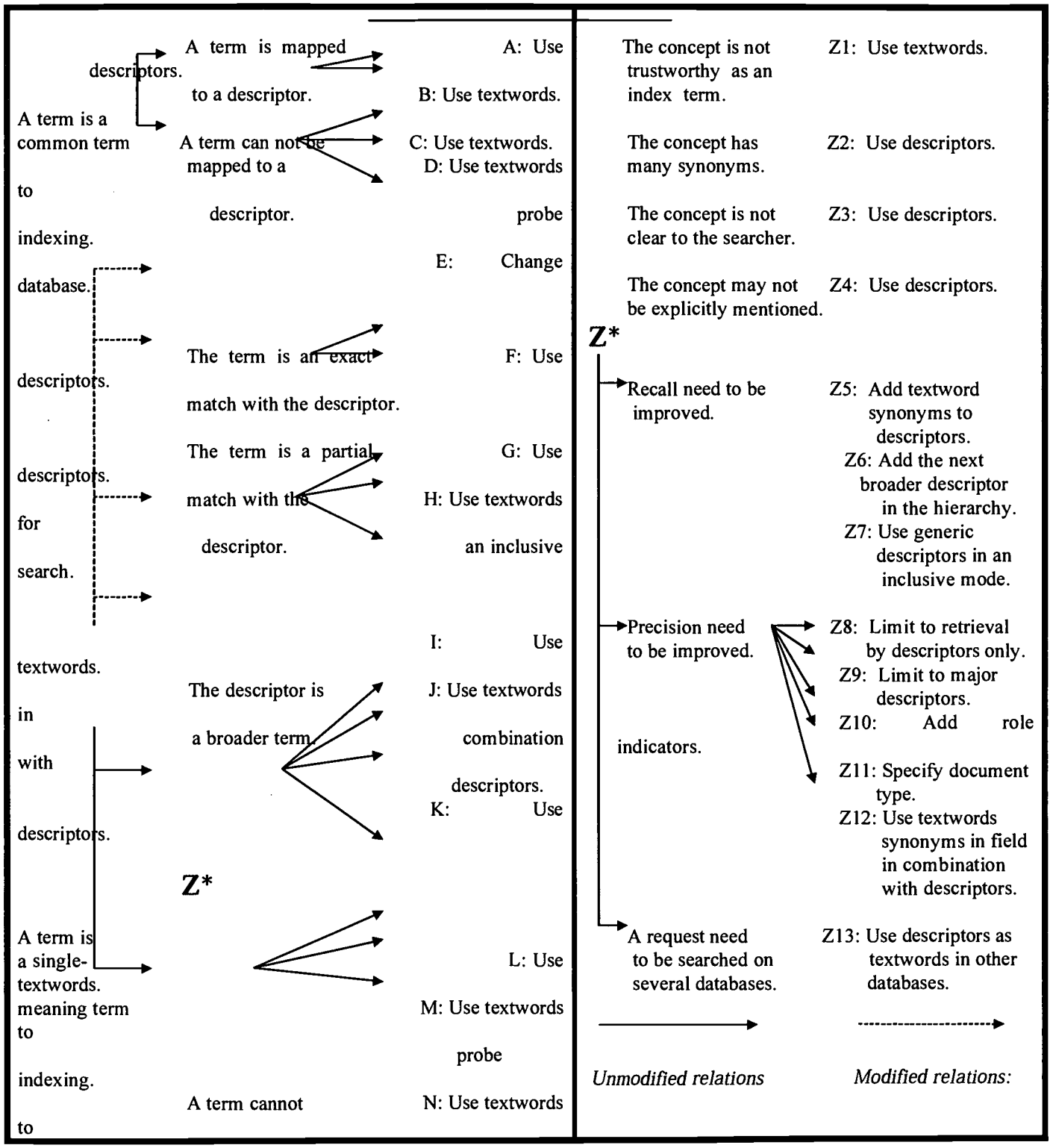


Figure 1
Components of on-line search

At the second stage, the search expert having clarified both the Semantic and the Pragmatic aspect of the request, proceeds to the construction of the search strategy. Four layers comprises the search strategy plan. At the first layer the expert selects the Data Base to be searched towards user's request. At the second layer expert considers regarding which terms- search keys will be used during the search process. At the third layer expert opine which search key will be entered first, and finally at the forth stage , which at the same time affects the third stage of Feedback review, experts mediate how to review unsatisfactory results.

The basic dilemma of the librarian search experts, is the appropriate use of free text search or the controlled vocabulary search according the type of the search key. If a search key is a single meaning term (uniquely defined and specific to the concept that represent), then using free text search seems to be the most promising choice to be followed. On the contrary in the case where the search key is a common term having a broad and vague meaning, then free text search destroys the relativeness and preciseness of the retrieval set, and thus controlled vocabulary search is preferred. The advantage of the controlled vocabulary search type, is the use of descriptors which are single meaning terms used for thesaurus construction in databases. Many concepts are accurately indexed under such descriptors. Therefore a crucial point for the performance of the search process, is the selection of the search key that will be used. On the research project *"Searchers' Selection of Search Keys* (part I II III)", of Raya Fidel, is expressed the idea that expert searchers usually follows some general rules in order to conclude which search key to select before the search session is initialised. This set of rules is defined as *Selection Routine* [Fidel 1991]. An overview of the decision tree of the Selection Routine and the corresponding paths from the initial assumption to the final decision is sown in Fig. 2.

As an explanatory example of the searchers selection routine, consider the case where a search term is a common term and is mapped to a descriptor. Then this descriptor will be used as a search key instead of the common term (Case A).This fact implies a controlled vocabulary search type. However when a common term can not be mapped to a descriptor then fact implies the use of free text search (Case C). In the above figures the continuous lines represent the relations of the initial assumption to the final decision as they are shown in the original paper, while the dotted lines correspond to the modified relations that are .



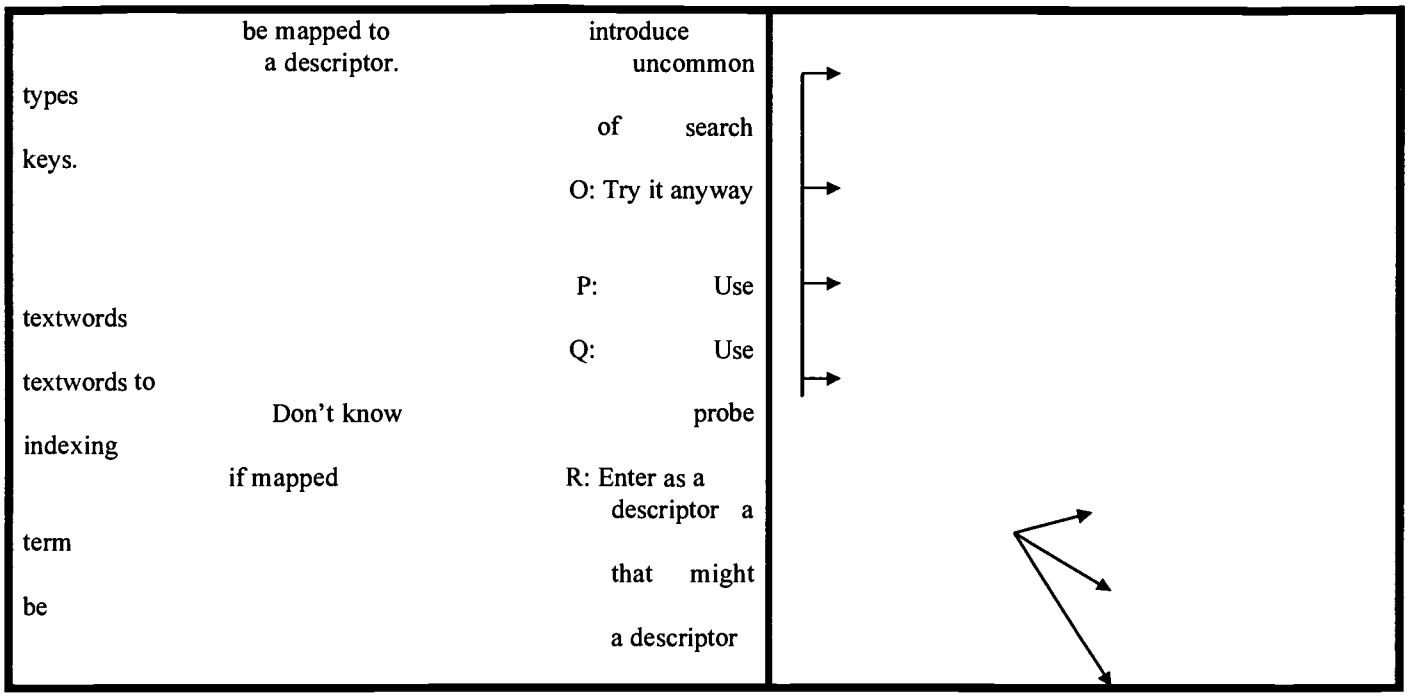


Figure 2
Decision tree of the Selection Routine

System description.

The system comprises of the following components [Fig 3].

1. Web based Interface: This component is the front end user interface where the user can interact with the system and define the desirable search term. Additionally user is interviewed by the system in order the semantic and the pragmatic aspect of the search to be clarified.
2. Spell checker: A spell checker is used in order to eliminate misspelled search terms. Speller fires optionally after user's suggestion.
3. Consultative core: This component includes the knowledge base of the system (Selection Routine, Metaknowledge rules), and interacts with the retrieval component.
4. Retrieval Component: Retrieval component combines a variety of retrieval tools which co-operate with the consultative core in order the retrieval set to be achieved.

A further description of the Consultative core component will follow.

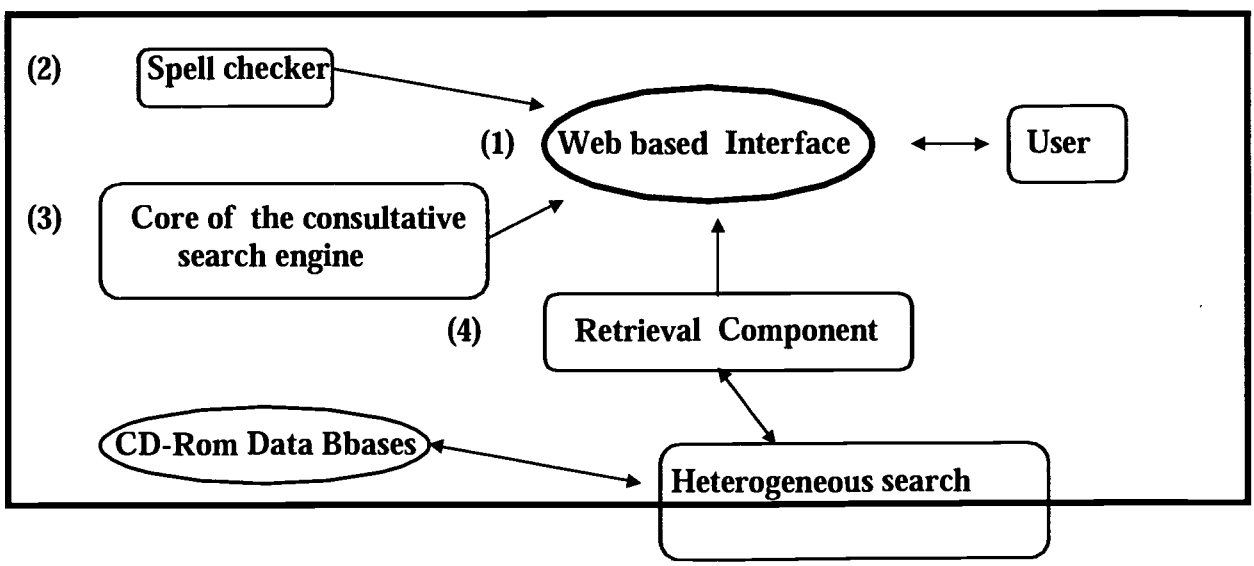


Figure 3
System description

In a typical search session user accesses the system via WWW. The system in order to clarify the Semantic and the Pragmatic aspect of user invoked search session, initialize an interview session with the user by displaying a number of form based questions. Filling these forms, user among others, is requested to define the search term, the repositories that he/she prefers to be searched (CD-Rom Data Bases or Internet or both), the corresponded topic to the search term (i.e. music or education) or if an expert's suggestion regarding the selection of an appropriate keyword is required (use of thesaurus). The defined search term can optionally be spell checked. Furthermore, consultative component using the selection routine, opines regarding the appropriated search key that will be used during the search session and the corresponding repositories that they will be reached. The results of the retrieval set are displayed to the user and in case that results are unsatisfactory, the whole search session can be refined.

Consultative core includes the knowledge base of the system in a form of *if... then...* rules and using is a forward chain inference engine constructs the search strategy. Two kinds of rule sets are included in this component. The first rule set represents the decision tree of the Selection Routine. This rule set affects the selection of the search key and the search type (free text or controlled vocabulary). The second rule set is the Metaknowledge rule set which have an effect on conflict resolution cases. A representative example of both rule sets is given.

Selection Routine Rule Set: In this example, cases A and B of Fig 2 are represented where the user defined term [term] is common term [CTR] and is mapped to a descriptor, so search expert can either use descriptors. [DSRC] to apply a control vocabulary search method (case A), or can use textwords [TXTWRD] to apply a free text search. So Rule_A corresponds to case A, and Rule_B corresponds to case B.

Rule_A: If
is_CTR <term>
& is_mapped_to_DSRC <term>
then
use_DSRC

Rule_B: If
is_CTR <term>
& is_mapped_to_DSRC <term>
then
use_ TXTWRD

Metaknowledge Rule Set: As it is earlier stated this rule set concerns conflict resolution cases. Conflict resolution in general, corresponds to the system "making up its mind" which rule to fire [Jackson]. During the contraction of the search strategy, it is very often the case where two or more rules are eligible to fire. In such cases meta-rules take effect in order to solve the conflict session by suggesting to the system which rule to fire first. Perceiving the above rules statements, Rule_A (RA) and Rule_B (RB), typical example of a conflict session can be noticed. Both left hand side premises of Rule_A (RA) and Rule_B (RB) are the same: *The defined term [term] is common term [CTR] and is mapped to a descriptor [DSRC],* while the right hand side premises are completely different. So it is obvious that in a forward chain session where these premises are true [T], both rules RA and RB will be loaded on the working memory [WM]of the inference engine and will be both eligible to fire causing a conflict to the system. At this point, meta-rules becomes activated. An example of the structure of meta-rules are the MR_ 1 and MR_11 rules

MR_1: If
is_not_nil <WM>
& RA and RB member_of <WM>
not_need_improve_recall <T>
then
use_RA
just
Searcher almost always prefer to enter descriptor as search key in

MR_11: If
is_not_nil <WM>
& RA and RB member_of <WM>
not_need_improve_recall <F>
then
use_RB
just
When the recall set using descriptors

order to restrict the retrieval set.	*is poor, searchers prefer to use text-word search alternatively to increase the number of recalls.*

In this example MR_1 MetaRule, on the left hand side examines *if the Working Memory of the system is not empty,* in other words system has started the chaining, *if RA and RB are loaded on the WM,* causing a conflict, and additionally examines *if the retrieval set do not need improvement,* in order to assure that the retrieval set had not been obtained yet-search session is still on progress. In that case meta-rule MR_1 loans priority to RA in order to fire first and additionally provides to the user, the justification for this selection (optionally). In case where the retrieval set had already been obtained and considered to be poor or irrelevant, then MR_1 loans priority to RB to fire in order recalls to be improved (refinement session). Again the justification for this selection is available to the user.

Implementation issues

System implementation issues have been also addressed in the discussion of Mentor system [Tsinakos & Margaritis 1996]. Consultative core component of the system is being implemented in Alegro Lisp, while the front end interface of the system is hosted on a CL-HTTP server CL-HTTP is a full-featured server for the Internet Hypertext Transfer Protocol, implemented in Common LISP in order to facilitate exploratory programming in the interactive hypermedia domain and to provide access to complex research programs, particularly artificial intelligence systems [Mallery, 1994].

The Retrieval Component, in co-operation with the Consultative core component, reaches the appropriate repositories in order the retrieval set to be achieved. In case where the search session regards information retrieval from a database- CD-ROM, retrieval component uses the SilverPlatter information retrieval system for the Internet environment called WebSPIRS [WebSPIRS]. WebSPIRS provides potential to the user to search a remote CD-ROM database using WWW interface. In case where the retrieval is applied on Internet repositories, retrieval component uses a number of intelligent meta-search engines such as Quarterdeck WebCompass 2.0. Such meta-search engines can "work" in conjunction with popular search engines such as AltaVista, Yahoo, WebCrawler, Excite and Lycos, as well as many others and are able to filter summarize and categorize the acquired information. Retrieval Component can apply a search in both environments (CD-ROM Data Base and Internet), using at the same time the WebSPIRS and meta-search engines retrieval tools.

The ability of remote search of a CD-ROM database using WebSPIRS software, has been accomplished by using SilverPlatter's Electronic Reference Library Technology. Electronic Reference Library is a multi-user application server implementation of SilverPlatter's CORE technology. ERL client/server model provides local and wide area networking access to all SilverPlatter databases and enables easy loading of pre-indexed and ready to search information from CD-ROM or tape.

References.

1. [AUTONOMY] http://www.agentware.com
2. [Bailey 1992] Charles W. Bailey "The Inelligent Reference Information System Project" Information Technology and Libraries 11, September 1992, 237-44.
3. [Bates, 1972] M. J Bates, *"Factors affecting subject catalog search success"*, Unpublished doctoral dissertation, University of California, Berkely 1972, USA.
4. [Borgman 1986] C. L Borgman, *"Why are online catalogs hard to use? Lessons learned from information retrieval studies"*, Journal of the American Society for Information Science, 1986, 37, 387-400.
5. [Bruke & Hammond et al 1997] Robin Bruke and Kristian Hammond and Benjamin Young and Julia Kozlovsky, *"Intelligent Web Search Engines"*, PC AI Jan/Feb. 1997, 39-42.
6. [Connel, 1991]. Tschera Harkness Connell, "*Subject searching in on line catalogs: An explanatory study of knowledge used"*, Unpublished doctoral dissertation, University of Illinois, Urbana-Champaign 1991, USA.
7. [Connel 1995] Tschera Harkness Connell, "*Subject searching in on line catalogs:Metaknowledge Used by Experienced Searchers"*, Journal of the American Society for Information Science, 1995, 46 (7), 506-18.
8. [Doorenbos et al,1996] R.B. Doorenbos , O.Etzioni and D.S. Weld. *"Ascalable comparsion -shopping agent for the world-wide web"*, Technical Report, Dept. of Computing Science and Engineering, Univ. of Washington, Jan 1996 USA.

9. [ECHO SEARCH] http://www.iconovex.com/ECHO/ECHOS.HTM
10. [Etzioni, 1994]. O.Etzioni and D.S. Weld. *"Asoftbot-based interface to the Internet"*, Comm of ACM 37 (7), Jul. 1994, 72-76. http://www.cs.washington.edu /research/softbots.
11. [Fidel 1991]. Raya Fidel, *"Searchers' Selection of Search Keys: I. The Selection Routine, II. Controlled Vocabulary or Free-Text Searching, III. Searching Styles"*, Journal of the American Society for Information Science, 1991, 42 (7), 490-527.
12. [Jackson].Peter Jackson, *"Introduction to Expert Systems"*, Second Edition, Ch 8 pg. 142-143.
13. [Koster 1994] Martijn Koster *"ALIWEB - Archie-Like Indexing in the WEB"*, First International WWW Conference, May 1994, CERN, Geneva Switzerland.http://www.nexor.co.uk/public/aliweb/aliweb.html
14. [Mallery, 1994] John C. Mallery *"A Common LISP Hypermedia Server"*, First International WWW Conference, May 1994, CERN, Geneva Switzerland.http://www.ai.mit.edu/projects/iiip/doc/cl-http/server.html
15. [McBryan, 1994] Oliver McBryan, *"GENVL and WWWW: Tools for Taming the Web"*, First International WWW Conference, May 1994, CERN, Geneva Switzerland. http://wwwmbb.cs.colorado.edu/~mcbryan/bb/summary.html
16. [MetaCrawler] http://www.metacrawler.com
17. [MORE LIKE THIS] http://www.morelikethis.com
18. [Myers 1994] Judy E. Mayers *"A newer version of Reference Expert, now beeing tested, has a keyword capability"*, Personal communication, Dec. 16, 1994.
19. [Perkowitz & Etzioni] M. Perkowitz and O. Etzioni. *"Category translation: Learningto understand information on the Internet"*, Fifteenth Internetional Joint Conference on AI, Aug 1995, Montreal CA, 930-936.
20. [Tsinakos & Margaritis 1996] Avgoustos A. Tsinakos and Konstantinos G. Margaritis, *"MENTOR Internet Search Advisor and Information Retrieval System"*, WebNet 96 - World Conference of the Web Society, Oct. 1996, San Francisco USA, 583-84. http://129.115.62.189:80/info/webnet96/html/114.htm
21. [WEB COMPASS 2.O] http://webcompass.qdeck.com
22. [WebSPIRS]. http://www.silverplatter.com/sampler/webspirs.html

Acknowledgments.

To Dr. Raya Fidel who is an Associate Professor at the Graduate School of Library and Information Science, University of Washington.

U.S. Department of Education
Office of Educational Research and Improvement (OERI)
National Library of Education (NLE)
Educational Resources Information Center (ERIC)

NOTICE

REPRODUCTION BASIS

☒ This document is covered by a signed "Reproduction Release (Blanket) form (on file within the ERIC system), encompassing all or classes of documents from its source organization and, therefore, does not require a "Specific Document" Release form.

☐ This document is Federally-funded, or carries its own permission to reproduce, or is otherwise in the public domain and, therefore, may be reproduced by ERIC without a signed Reproduction Release form (either "Specific Document" or "Blanket").