IDEA GROUP PUBLISHING

701 E. Chocolate Avenue, Suite 200, Hershey PA 17033-1240, USA
Tel: 717/533-8845; Fax 717/533-8661; URL-http://www.idea-group.com

ITP5011

Fuzzy Intelligent Agents for Data Mining and Information Retrieval from World Wide Web Outline Databases

M. Mohammadian
University of Canberra, School of Computing, Canberra, ACT 2616, Australia, masoud@ise.canberra.edu.au, Tel: +61 262012917,
Fax: +61 262012917

R. Jentzsch
University of Canberra, School of Computing, Canberra, ACT 2616, Australia, Ric.Jentzsch@canberra.edu.au, Tel: +61 262012917,
Fax: +61 262012917

ABSTRACT

Each month, each year the amount of information available via the World Wide Web continues to increase. This continued increase makes the job of locating useful and relevant information more complex. This complexity drives the need for improved search and retrieval techniques, in particular using better search engines. Current search engines are primarily passive tools locked into a particular vendor's way of doing things. Intelligent agents can assist in improving the performance of search and retrieval engines. The use of existing search and retrieval engines with the aid of an intelligent agent will allow a more comprehensive search. This research study provides the building blocks for integrating intelligent agents with current search engines. It shows how an intelligent system can be constructed to assist in better information retrieval and filtering. This research study is unique in the way the agents are directed and how fuzzy logic and intelligent agents are combined to improve the information filtering and retrieval of information via the World Wide Web sources.

INTRODUCTION

The amount of information potentially available from the World Wide Web (WWW) that can be found in and between Web pages, documents, and databases continues to grow. The amount of potential "information overload" is staggering. Traditional approaches to information overload are now more commonly referred to as information filtering and are related to the established research fields of information retrieval and user modelling. Users are always looking for better and more accurate ways of filtering information to satisfy their particular needs. Existing search and retrieval engines provide more capabilities today then ever before. Yet, search engines still do not always provide sufficient assistance to users in locating and filtering the relevant information [Jensen 2002]. Search engines are increasingly unable to turn up useful results to search queries posed by users [Biology Internet Gateway 2003].

As of 2003, the Web accessible population was estimated at 682 million users world wide [Internet World Stats 2003]. As of 2001, it was estimated that there was more than 8.7 million unique Web sites [OCLC 2003]. Although the rate of growth has slowed, the Web and information accessible via the Web continues to grow rapidly. Researchers in 1999 estimated that none of the studied search engines indexes more than 16% of total index able WWW [Lawrence and Giles Lee 1999]. Search engines only partially meet the increased need for an effective means of finding information on the Internet [Lucas and Nissenbaum 2000].

There continues to be a need for more effective and intelligent search and retrieval tools for improving access to data and information that is stored and can be accessed via the WWW. Search engines such as AltaVista, Excite, Google, HotBot, Infoseek, Northernlight, Yahoo, [Search Engine List 2003] and numerous other search engines such as MSN, American On Line, Ask Jeeves, Netscape, and Dogpile are the sites that offer a large range of Web searching services. Nielsen Net Ratings "one of the leading Internet audience information and analysis services" and SearchEngine Watch publish information about search engines, search engine news, audience reach, average minutes spend on a search per visitor, and total search house by search engine [Hines 2002, SearchEngine Watch 2003].

Specific requirements used by such search engines as Alta Vista, Google, HotBot and Lycos can be seen in searchengines.com [Searchengines 2003a]. These search engines are can be very sophisticated or not as sophisticated as users would like or expect. AltaVista, for example, indexes only the first two sentences in a document. When a search robot looks at a page, via Alta Vista, the top of page tags are looked at, the document length, keywords in the html title, and themes are considered [Hines 2002]. AltaVista scores the retrieved information and returns the results. The way the search engines score Web pages can often cause unexpected results [Jensen 2002]. In some case some relevant Web pages for a given search query are not shown on via one search engine query while a different search engine might give the page a low ranking.

There are many online information retrieval tools that are powerful in their ability to locate matching term(s) and phrases. These search engines are largely passive systems. Eyeballz in New Zealand did a survey and found that 92% of Internet information requests begin at a search engine and only 90% of the users only looked at the first 30 results [Eyeballz 2002]. Given the high level of searching and the low level of results accessed there remains the question of the ability of search engines to provide users with what they are asking for.

SEARCH ENGINES

There is relatively limited information on the specific details of the algorithms that search engines employ to achieve their particular results. A Web site can be indexed, scored and ranked using many different methods and algorithms depending on the particular search engine [Searchengines.com 2003a]. Web managers have become concerned with ranking and the scores allocated to their organization's Web site. As such, Web managers have become more familiar with the way search engines work, thus designing their Web pages based on that knowledge. On the other hand, search engines are becoming more familiar with the ways that Web managers are designing their Web pages [Searchengines.com 2003b]. Search engines usually provide the users with the top ten to top twenty relevant hits. The search engines ranking algorithm most often are based on the use of the position and frequency of keywords for their search. The Web pages with most instances of a keyword and the position of the keywords in the Web page can determine the higher document ranking [Jensen 2002, Lucas and Nissenbaum 2000, Searchengines.com 2003a].

It is interesting to note that search results obtained from search engines may be biased towards certain sites and rank low a site that may offer just as much value as do those who appear on the top ranked Web site [Lucas and Nissenbaum 2000]. There has often been questions asked without substantial responses in this area.

INTELLIGENT AGENT

Intelligent agents [Bigus and Bigus 1998, Watson 1997] may prove to be the needed element in transforming the passive search and retrieval systems into active, personal research assistants. Using a combination of effective information retrieval and intelligent agent techniques, the combination provides promising results in improving the performance of existing search engine queries.

Intelligent agents are computer software that can assist the user with computer applications. Intelligent agents can employ several techniques. Agents are created to act on behalf of its users(s) (another type of agent) to carry out difficult and time-consuming tasks [Bigus and Bigus 1998, Jensen 2002, Watson 1997]. Most agents today employ some type of artificial intelligence techniques to assist the users with their computer related tasks, such as reading e-mail [Watson 1997], maintaining a calendar, and filing information [Bigus and Bigus 1998]. Some agents can be trained to learn by example in order to improve the performance of the tasks they are given [Bigus and Bigus 1998, Watson 1997]. There are several ways that agents can be trained to better understand user preferences, such as (not limited to) using evolutionary computing systems, neural networks, and adaptive fuzzy logic expert systems. The combination of search and retrieval engines, the agent, the user preference, and the information retrieval algorithm can provide users with the confidence and trust they require from intelligent agents. A modified version of this approach is used throughout this study for intelligent information retrieval from WWW.

Intelligent Agents (i-agents) may be on the Internet or they can be on mobile wireless architectures. In the context of this research, however, the tasks that we are primarily concerned with include reading, filtering and sorting, maintaining information currency and informing users about certain events.

The user who is seeking information from the WWW is an agent. The user agent may teach the i-agent by example, or by employing a set of criteria for the i-agent to follow. Some i-agents have certain knowledge (expressed as rules) embedded in them to improve their filtering and sorting performance. For an agent to be considered intelligent, the agents should be able to sense and act autonomously in its environment. To some degree i-agents are designed to be adaptive to their environment and the changes in their environment [Bigus and Bigus 1998, Jensen 2002, Watson 1997].

This research considers i-agents for transforming the passive search and retrieval engines into more active, personal user assistants. By playing this role i-agents can be considered to be collaborative with existing search engines as a more effective information retrieval and filtering technique.

For a more complete definition of “What is an agent?” check Wooldridge and Jennings 1995, Nwana 1996, or Franklin and Graesser 1996 [Franklin and Graesser 1996, Nwana 1996, Wooldridge and Jennings 1995].

INTELLIGENT AGENTS FOR DATA MINING AND INFORMATION RETRIEVAL

Finding information has always been a labor intensive task. With the advent of the Internet, intranet and World Wide Web (WWW) there has been an ever increasing amount of information available to those who have WWW access, time to access it, and time to read through what is presented to them. The number of Internet site has grown exponentially since the introduction of the WWW infrastructure and Web browsers. Web browsers such as Opera, Netscape Navigator, and Windows Explorer and Web search engines (some listed in reference list) have made it possible for many new users of the WWW to have access to more readily available information then ever before. But with growth new problems are created. This is especially true due to the speed and complexity that has accompanied the WWW for information seekers. I-agents are one way in which these seekers can deal with the complexities of first finding and second filtering the information that is available and desirable to the seeker.

In this research we have aimed to investigate and create intelligent agents that are able to perceive the world around them. As an intelligent agent, they are required to recognize and evaluate events as they occur, determine the meaning of those events and then take one or more actions as directed. An event is a change of state within that agent’s environment for example when a information is added to changed in a database or on a Web page [Bigus and Bigus 1998, Watson 1997].

Intelligent agents must be able to process data and may use several processing strategies to get the right data. They may be designed to use simple strategies (algorithms) or they could use complex reasoning and learning strategies to achieve their tasks. The success of i-agents depends on how much value they provide to their users [Bigus and Bigus 1998, Jensen 2002, Lucas and Nissenbaum 2000, Watson 1997].

The Java programming language was used in this research to develop an intelligent agent using a fuzzy logic approach for information finding and filtering. The i-agent that was developed is used to actively search out desired information on the WW and filter out unwanted information for the searcher.

The procedures for information retrieval from WWW that was used is described below:

Phase 1:

The searcher tells the intelligent agent which search engine’s such as AltaVista, Excite, Google, HotBot, Dogpile, and Northernlight etc… are to be used in the search query.

The searcher identifies one or more (or combinations of) keywords (k_1, k_2, … , k_n) in a form understandable to the selected search engine’s. Then the agent is initialized to obtain the results of search query from the selected search engine’s. The host machine (of the search engine) returns the requested information with no specific format or acknowledgment.

Phase 2:

The intelligent agent then calls its routines to identify all related URLs obtained from search engine(s) and inserts them into a temporary list referred to as “TempList”. The TempList is restricted to at most 600 URLs, this is more of a limitation of the platform that was used than any statistical or particular reason. We doubt that any one searching would look at more than 600 (if that many) URLs.

For each URL in the TempList the following tasks are performed:

2.1 Given an URL address from TempList connect to that Web page;
2.2 Once the connection is established then:

Read the Web page content and calculate the number of times (keyword frequency from phase 1) the required keyword(s) in the list (k_1, k_2, … , k_n) appear on that page.

Both position and frequency of keywords are used to assign a position and frequency score to a Web page. Where the instances of the keywords on the Web page are more frequent and the relative position earlier on the Web page, then other keywords with the other occurrence instances, the Web page is given a higher ranking. In other words, a higher weighting is assigned to the search query term first with a higher frequency of occurrence than the other keyword terms (k_1, k_2, … , k_n). The frequency is considered twice are relevant to other criteria do to its relevance to the keyword terms identified by the searcher. The following fuzzy rules are used in evaluating and assigning a score to a Web page:

If Frequency_of_keywords = High then Frequency_Score = High;
If Frequency_of_keywords = Medium then Frequency_Score = Medium;
If Frequency_of_keywords = Low then Frequency_Score = Low;

The score obtained from applying these fuzzy rules is referred to as the ***Frequency_Score***. The frequency_score is balanced against the relative position of the keyword term to the Web page. This helps in separating out two or more keyword terms that have equal or near equal frequency of occurrence as well as relevancy to the searcher request. The position of keywords on a Web page is used to assign a relative position score for the Web page. The following fuzzy rules are used in evaluating and assigning a position score to a Web page:

If Position_of_keywords = Close_To_Top then Position_Score = High;
If Position_of_keywords = More_&_Less_Close_To_Top then Position_Score = Medium;
If Position_of_keywords = Far_From_Top then Position_Score = Low;

The score obtained from the above fuzzy rules is referred to as the ***Position_Score***.

The number of links on a Web page is used to assign a link score for the Web page. The following fuzzy rules are used in evaluating and assigning a link score to a Web page:

If Number_of_Links = Large then Link_Score = High;
If Number_of_Links = Medium then Link_Score = Medium;
If Number_of_Links = Small then Link_Score = Low;

The score obtained from the above fuzzy rules is referred to as the ***Link_Score***.

A final calculation, based on the scores for each page, is done by aggregating all scores frequency_score, position_score, and link_score, as described above. That is for each Web page a score according to the following is derived.

Score = (2*Frequency_Score) + Position_Score + Links_Score

2.2.2 For Web pages with high scores, identify all URLs on the Web page and make a list of these URLs;
2.2.3 For each URL found on the Web page, connect to that Web page evaluate and assign a score as described in 2.2.1. If no keywords appeared in that page, then abandon this page and store the results, otherwise store the URLs with its score in a list called FitURLs;
2.2.4 Continue processing the information based on the scores and save the results in the FitURLs list.

Notice that steps 2.2.2 and 2.2.3 not only the Web page's content that are returned as relevant Web pages by search engines for given keyword's but these steps are used to identify any Web pages linked to and from (Web link) that Web page. A Web link is a link from a Web page on WWW to another Web page on WWW.

That is choosing a Web link L_1 in a Web page P_1, it may follow L_1 to the second Web page $P_{2,1}$ linked by L_1 based on the score of the page P_1 and decision made by the fuzzy logic system. The intelligent agent then performs keyword's count in Web page $P_{2,1}$ and may also investigates all the links ($L_{2,1}$ $L_{2,2}$ …$L_{2,n}$) in Web page P_2 collecting the statistics about the appearance of keyword's in those pages based on the score of page $P_{2,1}$. If no keyword(s) appeared in that page then the Web page is not immediately abandoned. In this approach a Web page link to a high scoring Web page could discarded because it does not contain the required keyword(s). This abandoned page may however contain Web links that are very relevant the searchers request. Therefore an additional feature has been added to the i-agent for deciding to continue or abandon a Web page based on the historical information of the search query. This feature is described below.

Initial simulation and testing of the i-agent using the above described approach has been an ongoing process throughout all phases of this research project. It was decided to use the University of Canberra's (UC) search engine for simulation and validation of this system due to the moderate number (compared to the WWW) of Web pages available on this website. With the number of Web pages available on the (University of Canberra) UC website, it is possible to conduct this research and validate the results. It also helps to ensure that the intelligent agent is doing what it is designed to do.

Search query: Fuzzy Logic
Search results – as of May 2003

Search Engines	Number of Links Visited	Number of Links Abandoned	Number of pages downloaded
UC	39	27	12
Yahoo	1873	1749	124

Search results by i-agent

Follow up testing was done using such sites as Google, AltaVista, and Yahoo. It was found that Yahoo and other search engines such as those listed in the references, return very large number of pages for most keyword searched. It was also found that the time it took to search using multiple search engines did not always justify the results. Therefore it is very difficult to objectively validate the results of the i-agent using multiple search engines. The index of these search engines are fairly dynamic and the result of a search using for example "fuzzy logic" as the required combined keywords returns different results on different days using these search engines. For example in the table above we had to visit all pages returned from the yahoo search before we could validate the results of the i-agent.

Finally in this research study we have investigated the possibility of recording the number of keywords located in pages visited by i-agent. The aim of this part of the research project is to examine the idea that if a series of pages visited have the keyword required then the links from these pages should not be abandoned if one of the pages doesn't have any occurrences of the required keyword. In this way the number of pages searched will however increase. For example assume that a user is looking for the pages related to the keywords "fuzzy logic". Assume also that a search engine returned a number of pages for this keyword. Assume also that the first page returned has one occurrence of the combined keyword "fuzzy logic" and it has also a link to another webpage. See below:

Fuzzy logic: http://www.ise.canberra.edu.au/masoud/index.html

Then according to i-agent, the links on this page are to be investigated. Now i-agent will visit the page http://www.ise.canberra.edu.au/masoud/index.html

Now assume that this page (http://www.ise.canberra.edu.au/masoud/index.html) has the following details:

Fuzzy logic is useful. Research is being conducted into fuzzy logic applicationhttp://www.ise.canberra.edu.au/masoud/project1.html

Now according to the i-agent the search will continue to the page: http://www.ise.canberra.edu.au/masoud/project1.html

This page is discusses research for project 1.http://www.ise.canberra.edu.au/masoud/project2.html

The above Web page (http://www.ise.canberra.edu.au/masoud/project2.html) does contain any occurrence of the keywords "fuzzy logic". Therefore i-agent abandoned this Web page and it will not search the links in this webpage.

The additional feature added to the i-agent keeps track of the number of keywords found in all Web pages that are linked together and then decides to abandon or continue with its search (by calculating the average number of keywords found so far in all the linked Web pages visited). The average calculated using the following formula:

$$\text{Average} = \sum_{i=1}^{n} \frac{NumberOfKeywordFoundInPage(Pi)}{i}$$

A fuzzy logic system is also employed to make a decision to continue the search query or to stop searching. The knowledge base of the fuzzy logic system is illustrated below:

Averages

Score	Very Few	Few	Average	Large	Very Large
Very Small	VL	VL	LW	MD	MD
Small	LW	LW	MD	HI	HI
Normal	MD	HI	HI	HI	VH
Large	HI	HI	HI	VH	VH
Very Large	HI	HI	VH	VH	VH

Depth Level

Here will notice that the fuzzy logic has been extended to five classifications. The input to this system is the Score and Average and the output the Depth level (the level of search that i-agent will perform). Comparison of the results obtained using i-agent, and i-agent with historical fuzzy logic is given below:

UC Search	No. of Pages Return	User	i-agent and fuzzy logic	i-agent and historical fuzzy logic
Masoud Mohammadian	25	8	7	20
Ric Jentzsch	51	7	5	18
Fuzzy Logic	79	22	9	23
Intelligent Agent	3	2	3	3

Results obtained from i-agent, and i-agent with historical fuzzy logic (FL) on UC Website

The details of the number of links visited, abandoned and depth level of search as well as the number of pages returned is shown below:

UC Search	No. of Links Visited	No. of Links Abandon	Depth level	No. of Pages Return
Masoud Mohammandian	20	13	2	7
Ric Jentzsch	15	10	2	5
Fuzzy Logic	18	9	2	9
Intelligent Agent	5	2	1	3

Search result from i-agent with fuzzy logic:

UC Search	No. of Links Visited	No. of Links Abandon	Depth Level	No. of Pages Return
Masoud Mohammandian	460	440	2	20
Ric Jentzsch	70	51	1	19
Fuzzy Logic	35	12	1	23
Intelligent Agent	5	2	1	3

Search result from i-agent and historical fuzzy logic:

From the information provided, the keyword "intelligent agent" is returned the same results with all other strategies. The result of the i-agent with historical fuzzy logic returns more pages.

The number in the parenthesis in this table shows the evaluated links by user. The user evaluated each page and rank them as relevant or not relevant.

UC Search	No. of Pages Return	User	I-agent and Fuzzy Logic	I-agent and historical Fuzzy Logic
Masoud Mohammadian	25	8	7 (4)	20 (8)
Ric Jentzsch	51	7	5 (3)	18 (7)
Fuzzy Fogic	79	22	9 (6)	23 (11)
Intelligent Agent	3	2	3 (2)	3 (2)

It is interesting to note that the i-agent was able to retrieve the relevant Web pages and ignore the Web pages that are not relevant to the search query. For example, the i-agent and fuzzy logic has returned nine pages for the search keyword "Fuzzy Logic". Six of these three pages are relevant to the search (evaluated by the user). In compression the i-agent with historical fuzzy logic has returned twenty three pages from which eleven pages are ranked by the user to be relevant.

CONCLUSION

In this project, the intelligent agents are combined with fuzzy logic to improve search and information retrieval. The inclusion of fuzzy logic would enhance the system to find sites and documents that are more related to the keywords being searched.

Using a combination of search engines, the i-agent and fuzzy logic has shown to be an effective agent in filtering and retrieval of information from the World Wide Web. Simulation results conducted up to September 2003 show that the performance of the system using intelligent agents and fuzzy logic has improved the search results (in comparison to conventional search engines). It is not very surprising as agents use the results of search engines and then filter the irrelevant Web pages. The combination of fuzzy logic and i-agent in filtering and information retrieval provided promising results. The performance of the proposed system, in retrieving and filtering information, has shown to be generally good.

There are several drawbacks and limitations that are needed to be overcome. Additional simulations and experimentations are needed. What happens when a search engine is updated, modified, or changed? What effect does this have on the filtering of the identified information? What happens when a Web page is old, and the information is no longer current or accurate, yet the "score" remains high due to out-of-date inaccurate Web page content?

The application is quite stable and does perform its tasks well. However, there is a need to compare the use of one or more other programming languages for intelligent agents. Each might provide performance enhancement or degradation.

Future research includes identifying the association between the relevant link_scores and the use of meta-data for Web pages and the data semantics of the Web page contents. There are many questions that can be asked and investigated. What has been described herein is only the beginning.

ACKNOWLEDGEMENT

The authors would like to acknowledge the assistance of Mr. Long Tan and Mr. Thien Huynh in their programming assistance in this research project.

REFERENCES

Bigus, J. P. and Bigus, J. (1998). "Constructing Intelligent Agents with Java – A Programmers Guide to Smarter Applications", Wiley, ISBN: 0-471-19135-3.

Biology Internet Gateway. (2003). http://www.it-tallaght.ie/library/new/asb.htm 10 June, last accessed 18 November 2003.

Franklin S. and Graesser, A. (1996). "Is it an Agent, or just a Program? A Taxonomy for Autonomous Agents". http://www.msci.memphis.edu/~franklin/AgentProg.html last accessed 18 November 2003.

Goldberg, D. (1989). "Genetic Algorithms in Search, Optimisation and Machine Learning", Reading, Massachusetts: Addison Wesley Publishing.

Hines, M. (2002). The Information Architect. "The Problem with Petra bytes." http://www.techtarget.com.au last accessed 05 June 2002.

Internet World Stats. (2003). "The Big Picture". http://www.internetworldstats.com/stats.htm last accessed 02 December 2003.

Jensen, J. (2002). "Using An Intelligent Agent To Enhance Search Engine Performance". http://www.firstmonday.dk/issues/issue2_3/jansen/index.htm. last accessed 11 January 2002.

Lawrence, S. and Giles Lee, C. (1999). "Accessibility of Information on the Web". Nature, 400, pp107 – 109.

Lucas, I. and Nissenbaum, H. (2000). "The Politics Of Search Engine", IEEE Spectrum, June.

Nwana, H. S. (1996). "Software Agents: An Overview." The Knowledge Engineering Review. vol. 11, no 3, pp 1-40.

OCLC Online Computer Library Center. (2003). (http://www.oclc.org/research/projects/archive/wcp'stats'size.htm last accessed 27 November 2003.

Search engine marketing in New Zealand by Eyeballz. (2002) http://www.eyeballz.co.nz'search_engine_statistics.htm, last accessed 24 May 2002.

Searchengines.com. (2003a). http://www.searchengines.com'search_engine_statistics.html, and http://www.searchengines.com'searchEnginesRankings.html, last accessed 10 November 2003.

Searchengines.com. (2003b). "What Search Engines Don't' Like". http://www.searchengines.com/mistakes.html last accessed 05 December 2003

Search Engine Watch. (2003). SearchEngineWatch.com. http://www.searchenginewatch.com/ last accessed 05 December 2003.

Sullivan, D.(editor). Nielsen/NetRatings Search Engine Ratings. (2003) Current Audience Reach, http://www.searchenginewatch.com/reports/netratings.html last accessed November 22, 2003.

Watson, M. (1997). "Intelligent Java Applications for the Internet and Intranets". Morgan Kaufmann, ISBN: 1-55860-420-0.

Search Engine List 2003. Some search engines and URLs as of November 04, 2003:

AltaVista	www.altavista.com
America On Line	search.aol.com/aolcom/index.jsp (enhanced by Google)
Excite	www.excite.com/
Dogpile	www.dogpile.com
Google	www.google.com
InfoSeek	now powered by Google
Lycos	www.lycos.com/
MSN	www.msn.com/
Northernlight	www.northernlight.com
Yahoo	www.yahoo.com/

Wooldridge, M. and Jennings, N. (1995). "Intelligent Agents: Theory and Practice". The Knowledge Engineering Review. vol. 10, no 2, pp 115-152.

Related Content