

GENERAL PURPOSE SEARCH AND CLASSIFICATION ALGORITHMS FOR DISTRIBUTED INFORMATION ACCESS

Fadzilah Siraj, Wan Hussain Wan Ishak

Sekolah Teknologi Maklumat, Universiti Utara Malaysia, 06010 Sintok, Kedah

Email: {fad173; hussain}@webmail.uum.edu.my

Abstract

Networks protocol and software such as HTTP, Netscape and Internet Explorer have significantly ease importation and fetching of on-line information sources available on the Internet for the public, personal and organizational access. These information need to be organized so that users can access and search the information in the convenient way. One of the most popular tools for searching and finding information on the Internet is using the search engine. Many search engines have been made available on the Internet, yet users are not able to efficiently explore and find what they want in information overload. Some search engines displayed hundred of results each time the user requests for specific keywords or queries. Users have to visit each URLs even though the page does not suit their requests. Therefore, it is vital to provide efficient and timely Internet search algorithm to ensure the success of distributed services, particularly to facilitate the current and future applications of national information superhighway. Previous and current research on search engines and information retrieval show that Artificial Intelligence technique is one of the recommended solutions in providing the most relevant, timely and efficient search results for the user. This paper proposed a search engine model that incorporated Artificial Intelligence techniques.

1.0 Introduction

Internet is a major medium for information delivery. It is believed that Internet is anticipated to expend 1000% in the year 2000 and above (Lawrence and Giles, 1998a). Almost everything could be found in the World Wide Web (WWW). Information from all over the world covering historical, current and the latest news or event have been made available over the WWW for public, personal and organizational access. However, because of too many information, user suffered from the information overload problem. Users have to browse all the information in order to find the most relevant pages that suite their needs. This problem cause unpleasant feeling to users, especially those who are having an important thing to do in a limited time.

Another issues arises in searching on the WWW is the organization of the information. Well organized information is the most important feature in Internet surfing and searching. Well organized information helps user in retrieving the most related document as they search for certain specific documents. However, the characteristics of the web itself are limitation in managing and structuring the information on the Internet. WWW is not centralized, dynamic and diverse, therefore, organizing WWW content is very difficult and finding information on the web is very challenging (Pinkerton, 1994).

Search engines are the most powerful search tools on the World Wide Web, with the most popular being those that access the largest databases. A number of useful and popular search engines attempt to maintain full text indexes of the World Wide Web such as Altavista (www.altavista.digital.com), Excite (www.excite.com), Hotbot (www.hotbot.com), Infoseek (www.infoseek.com) and Lycos (www.lycos.com). Each search engine grouped information into several categories such as sports, education, music, news and etc. Users can simply click the hyperlink and another set of sub-category will appear. These facilities help user in narrowing their search to a specific page or sub-category. However, searching the web can still be a slow and tedious process. Limitations of the search services have led to the introduction of Metasearch engines such as Inquirus (Lawrence and Giles, 1998b), SavvySearch (Howe and Dreilinger, 1997) and STARTS (Gravano *et al.*, 1997). Meta search engine searches the web by sending query to multiple search engines (Lawrence and Giles, 1998b).

Artificial Intelligence (AI), a science of transferring human intelligence into a computer or machine, is an adequate approach in enhancing information retrieval on the WWW. The potential of AI have been explored in various area such as medical, business, finance, education, services and many more. It has been successfully studied in the data integration problem by combining with databases technology (Levy, 1998). AI techniques such as Machine Learning, Neural Networks, Genetic Algorithm and Intelligent Agent have been studied in conjunction to current or conventional information retrieval systems (*see for instance* McCallum *et al.*, 1999; Moukas, 1996; Bloedorn, *et al.*, 1996; Shavlik and Eliassi-Rad, 1998; Mohamed *et al.*, 2000; Nur Izura *et al.*, 1997).

This paper discusses AI technique, namely neural network in providing relevance feedback to user based on previous query or "experience". WSearch (**Web Search**) a search engine for general purposes was developed by integrating neural network in providing the most relevance results to users. In addition, search engine also provides several related keywords to enhance user's query.

2.0 Information Retrieval and WWW

Information retrieval is a process of retrieving documents from a set of documents (databases). Advancements in information storage and retrieval enable users to store and access documents on the Internet. Databases of documents could be stored in one place (a server) and accessed by users from any locations. Such characteristics is one of the benefits that encourages organizations to store information on the Internet with WWW as the medium for the information access. The main mechanisms for information retrieval on WWW are by either keyword search or hypertext browsing (Chen *et al.*, 1996). Nevertheless, more advanced research such as retrieving multilingual information on the WWW has also been introduced (Saidah *et al.*, 2000). Most of databases have their own simple search engine to allow users to retrieved certain documents from their databases (*see for instance* Riza *et al.*, 2000).

Common database system allows users to upload documents and store them into the databases using the upload facilities. The techniques are inconvenient to users and system administrator as they have to monitor the databases and update the information

manually. Search engines differ from common databases as they required up-to-date information and frequent update of information. Therefore, search engines require special engine to search the web and retrieve current and up-to-date documents on the web. Using specific keywords or query, the documents are then indexed and stored into the database. An engine is developed to assist users in retrieving information in the databases using specific keywords or query.

On the Internet, a search engine has three parts:

- A spider (also called a "crawler" or a "bot") that goes to every page or representative page on every Web site that wants to be searchable and reads it using hypertext links on each page to discover and read a sites's other pages (Figure 1).
- A program that creates a huge index (sometimes called a "catalog") from the pages that have been read.
- A program that receives user search request, compares it to the entries in the index, and return results to user.

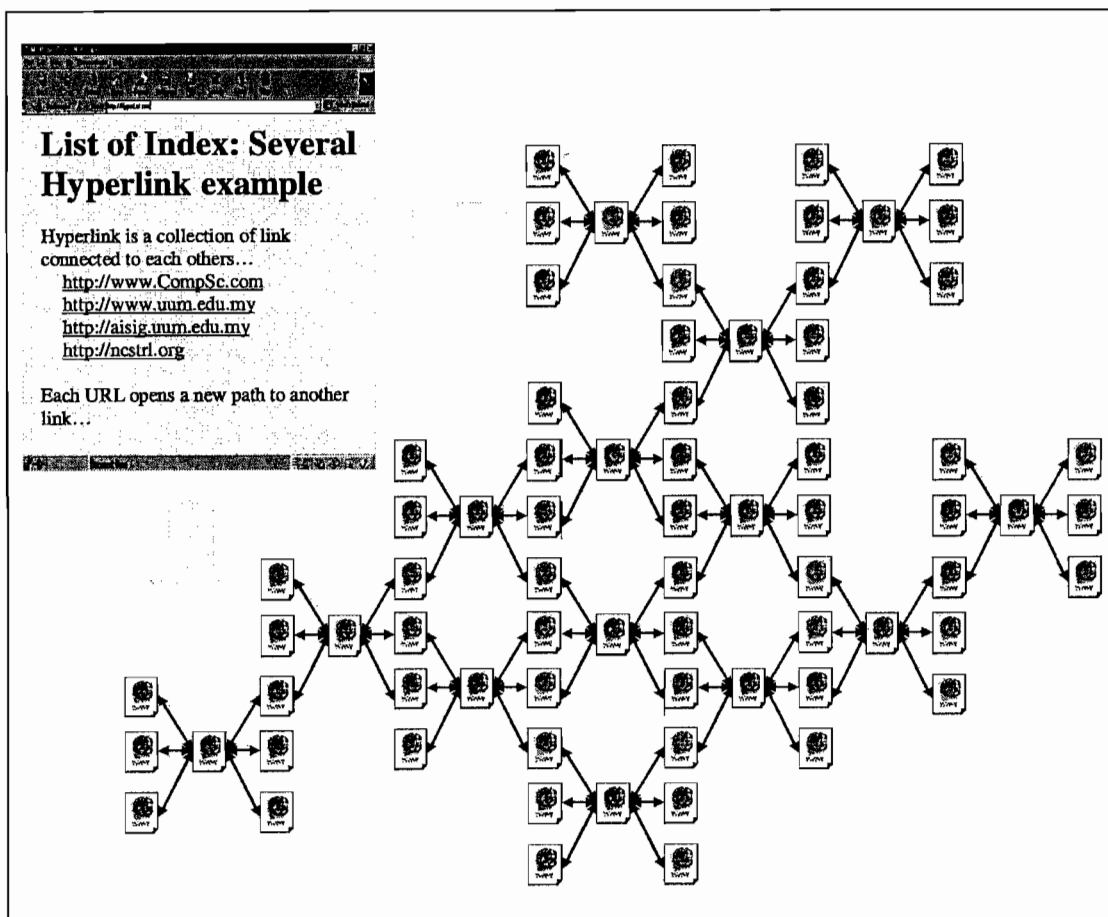


Figure 1: Hyperlink Structure

Currently, available search engines score queries against documents, compute the highest scoring documents, and present user with a set of document surrogates in ranked order. While the display of surrogates should allow the user to make a quick and possibly accurate judgement about the relevance of retrieved documents without downloading the full documents, its utility is reduced by the lack of effectiveness of

current retrieval engines interfaces. To improve the interface, AI techniques such as Natural Language Processing (NLP) can be integrated to process user's query and represent the keywords to the search engine (Strzalkowski *et al.*, 1996).

3.0 Neural Networks

Neural Network (NN) is a network of many simple processors or units (Sarle, 2000). It simulates the function of human brain model. In some cases, NN was viewed as a class of algorithm for statistical modeling and prediction (Jordan and Bishop, 1996). Differ to the usual computer programs, NN "learn" from a set of examples rather than being programmed (Kappen, 1996). In medical for example, NN has several potential over conventional computation and manual analysis. Such potentials are implementing using data instead of possibly ill defined rules, noise and novel situations are handled automatically via data generalization, predictability of future indicator values based on past data and trend recognition, automated real-time analysis and diagnosis, enables rapid identification and classification of input data and eliminates error associated with human fatigue and habituation (Patridge *et al.*, 1996).

Internet categorization and search presented by Chen *et al* (1996) employed self-organizing feature map in the process of categorizing the content of Internet's documents into subject-specific categories. The results of the initial study indicate that the technique produces meaningful results for small-scale applications and potentially useful concept maps for serendipitous browsing for large-scale applications. Nur Izura *et al* (1997) employed Counterpropagation network in the process of retrieving information from database. The experiments reveals NN capability in recognizing data from the database.

WAWA (Wisconsin Adaptive Web Assistant) is an instructional software agents that assists users in locating specific, current and relevant information on web (Shavlik *et al.*, 1999; Shavlik and Eliassi-Rad, 1998). WAWA complies users instructions and how to go about looking for the information into two NN. The tasks of NN were performs by two special function that are SCORETHISLINK and SCORETHISPAGE (Shavlik and Eliassi-Rad, 1998). Both functions are used to guide its autonomous navigation of the web.

4.0 Methodologies

The study on current search engine shows that there are two different methodologies to obtain information. The first method is to use one search engine such as Altavista and the second method is to use multisearch engine that is known as metasearch. The drawbacks of one search engine are that it has limited coverage, outdated databases and sometimes unavailable due to problems with the network and engine itself (Notess, 1996). The precision of search engines results may also vary because it generally focuses on handling queries quickly and users relatively simple ranking schemes (King, 1992). Rankings can be further muddled by keyword "spamming" to increase a page's rank order. The relevancy of a particular page is obvious only after the page is loaded and the query terms were found (Lawrence and Giles, 1998).

Metasearch engines such as MetaCrawler and SavvySearch overcome the problem of limited coverage by submitting queries to several search engines at once (Hawking and Thistlewaite, 1995). A Metasearch engine searches the web by making requests to multiple search engines such as Altavista or Infoseek (Lawrence, 1995). The main advantage of metasearch engine is that they combined the results of several search engines and presented consistent user interfaces (Boyan *et al.*, 1996). Combining results from six search engines for example yields about 3.5 times compared to one search engine (Lawrence and Giles, 1998a). Most metasearch engines rely on the document and search results returned by other search engines and thus enhance their limited precision and vulnerability to keyword spamming (Lawrence and Giles, 1998a). However, it inherit some difficulties and may also introduce its own problems (Lawrence and Giles, 1998a).

In this study, a metasearch technique is incorporated in WSearch. The following steps would be performed by WSearch in responding to user's query:

- Obtain the user requests.
- Obtaining result from several search engines.
- Ranking the document according to its relevancy.
- Presenting the results to the user.

Initially, each query will be passed through a stopword filter to remove words that have no meaning when taken out of context (Rijsbergen, 1979). After each query has passed through a stopword filter, the remaining words were then conflated by stemming to remove inflectional suffixes (Porter, 1980). The words will be reduced to their 'stems' and hence reduce redundancy. Suffix-stripping algorithm introduced by Porter will be employed in the study (Porter, 1980; Frakes, 1992). Having obtained the results from different search engines, the search results are ranked according to their relevancy to the user's request. Relevancy ranking was performed by using several heuristics such as search term frequency, position of keywords, the number of links to the page ranked, relevance weights on terms, inverse document frequency and density (Feinstein *et al.*, 1997). The user interface was developed to present the search results to the user.

The benefit of using heuristics ranking is to improve search results. The use of agent in the WSearch generally improves searching since the pages of information retrieved were not thrown away but the agent will rank them according to its importance. WSearch is succeeded in retrieving the document that contains the precise keyword stated in the user's request. In addition, WSearch also suggest some other relevant keywords to be used as queries. This distinctive feature of WSearch will certainly narrow down searching and reduce searching time.

WSearch architecture (Figure 2) consists of three main features, namely searching facility, database administration and searching agent. Searching facility helps users to search for the relevance documents using keyword search. Apart from searching, WSearch will produce a list of suggested keywords to assist users to narrow down their search. Suggested keywords are grouped in two categories; the same meaning keywords and similar keywords. The same meaning keywords are keywords that have the same meaning with users query (e.g. appreciation → gratefulness, esteem → respect, judge → court). Whereas, similar keywords are keywords that being

truncated from the original query (e.g. eaten → eat, books → book, searching → search, publication → publish).

Database administering is a process of managing information stored in the database. Database admin manages the database by deleting unwanted information (rubbish), updating knowledge based system and frequently restructuring database contents. New information is updated using an agent which craws the web into the hyperlink structure (Figure 1). Agent read hyperlink on each page and traverses the document for the next link. Several information from the document are extracted and stored into the database. Such information are page title, URL, keywords and a short description or abstract.

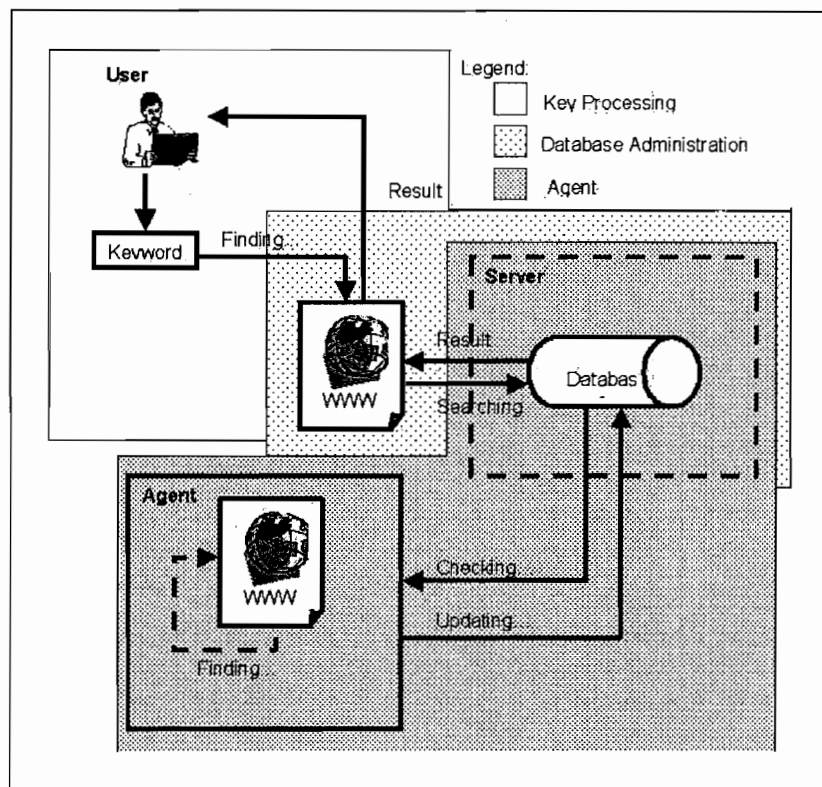


Figure 2: WSearch General Architecture

5.0 Conclusion

Information are world wide and it came from all over the world. Everybody who has access to Internet and WWW can publish their documents and provide it for public, organizational or even personal access. However, too much information on the Internet cause information overload problems. Users have to browse through all documents in order to find the documents with high ranking. Such ranking system facilitates current search engines. Yet, the relevance of the documents are still in debt.

Structured information retrieval system helps in organizing the information on the web into a structured and organized electronic record. Structured and organized information is the key that determines the usability of the information. The usability of the information practically means that the information can be achieved, download,

copied or manipulate (limited to the policy of the documents published) by users to fit their interest.

Facilities incorporated in WSearch is able to assist users in searching for the most relevance documents. Specific search feature in WSearch helps users in determining specific information based on the query. In spite of that, user's query will be pre-processed into root word to narrow down the search into more specific documents. Keywords or phrases that related to users query will be suggested for further searching.

References

- Bloedorn, E., Mani, I., and MacMillan, R. (1996). Representational Issues in Machine Learning of User Profiles. AAAI Spring Symposium on Machine Learning in Information Access.
- Chen, H., Schuffels, C., and Orwig, R. (1996). Internet Categorization and search: A Self-Organizing Approach. *Journal of Visual Communication and Image Representation*, Vol. 7, No. 1, pp. 88-102.
- Gravano, L., Chang, C. C., K., and Molina, H. G. (1997). STARTS: Stanford Proposal for Internet for Internet Meta-Searching. *In Proceedings of the 1997 ACM Sigmod Conference*.
- Howe, A., and Dreilinger, D. (1997). SavvySearch: A Meta-Search Engine that Learns which Search Engines to Query. *AI Magazine*.
- Jordan, M. I., and Bishop, C. M. (1996). Neural Networks. *Technical report No. A.I.Memo No. 1562*, Artificial Intelligence Laboratory: Massachusetts.
- Kappen, H. J. (1996). An Overview of Neural Network Applications. *Proceedings 6th ICCTA*, Wageningen, the Netherlands, pp. 75-79.
- Lawrence, S., and Giles, C. L. (1998a). Searching the World Wide Web. *Science*. Vol. 280, pp. 98-100.
- Lawrence, S., and Giles, C. L. (1998b). Inquirus, the NECI Meta Search Engine. *Appears in the Seventh International World Wide Web Conference*, Elsevier Science, pp. 95-105.
- McCallum, A. K., Nigam, K., Rennie, J., and Seymore, K. (1999). Automating the Construction of Internet Portals with Machine Learning. *Appear in Information Retrieval*, Klumer Academic Publishers.
- Mohamed Yusof, Shahrul Azman Mohd Noah, Azman Yasin, Azizi Zakaria, Fadzilah Siraj, and Shamrie Sainin (2000). Capaian Maklumat Teks Menggunakan Teknik Genetik Algoritma. *Presented at Persidangan Ulang Tahun Ke-30 UKM (5-6 September)*. Universiti Kebangsaan Malaysia.
- Moukas, A. (1996). Amalthaea: Information Discovery and Filtering using a Multiagent Evolving Ecosystem. *Appeared in the Proceedings of the Conference on Practical Applications of Agents and Multiagent Technology*, London.
- Nur Izura Udzir, MD Nasir Sulaiman, Ali mamat, Ramlan Mahmod ad fatimah Ahmad (1996). Rangkaian Neural Dalam Dapatan Pangkalan Data. *National Conference on Research and Development in Computer Science and its Applications*. Universiti Sains Malaysia.
- Partridge, D., Abidi, S. S. R., and Goh, A. (1996), Neural Network Applications in Medicine, *Proceedings of National Conference on Research and Development*

- in Computer Science and Its Applications (REDECS'96)*, Universiti Pertanian Malaysia; Kuala Lumpur, pp. 20-23.
- Pinkerton, B. (1994). Finding What People Want: Experiences with the WebCrawler. <http://info.webcrawler.com/bp/WWW94.html>
- Riza Sulaiman, Mohammad Khatim Hasan, Khairudin Omar, Mohd Juzaidin Ab Aziz, Zulkarnain Md. Ali, Suhaila Zainudin, Tengku Siti Mariam, Zaihosnita Hood, Zurina Muda, and Saidah Saad (2000). Pembangunan Jurnal Berasaskan Web: Pengalaman FTSM. *Simposium Teknologi Maklumat 2000 (11-12 April)*, Fakulti Teknologi dan Sains Maklumat, Universiti Kebangsaan Malaysia., pp. 166-170.
- Saidah Saad, Tengku Mohd Tengku Sembok, and Mohammed Yusof (2000). Pembangunan dan Eksperimen ke atas Sistem Capaian Maklumat Al-Quran Dwi Bahasa Berasaskan Web. *Simposium Teknologi Maklumat 2000 (11-12 April)*, Fakulti Teknologi dan Sains Maklumat, Universiti Kebangsaan Malaysia., pp. 160-165.
- Sarle, W. S. (2000). Neural Network FAQ, part 1 of 7: Introduction. *Periodic Posting to the Usenet Newsgroup comp.ai.neural-nets*, <ftp://ftp.sas.com/pub/neurl/FAQ.html>
- Shavlik, J., and Eliassi-Rad, T (1998). Intelligent Agents for Web-Based Tasks: An Advice-Taking Approach. *Appears in the Working Notes of the AAAI/ICML Workshop on Learning for Text categorization.*
- Shavlik, J., and Eliassi-Rad, T. (1998). Intelligent Agents for Web-Based tasks; An Advice-Taking Approach. *Appears in the Working Notes of the AAAI/ICML Workshop on Learning for Text Categorization.*
- Shavlik, J., Calcari, S., Eliassi-Rad, T., and Solock, J. (1999). An Instructable, Adaptive Interface for Discovering and Monitoring Information on the World-Wide Web. *Appears in the Proceedings of the 1999 International Conference on Intelligent User Interface*, Los Angeles, CA, USA.
- Strzalkowski, T., Guthrie, L., Karigren, J., Leistensnider, J., Lin, F., Perez-Carballo, J., Straszheim, T., Wang., and Wilding, J. (1996). Natural Language Information Retrieval: Trec-5 Report. *In Voorhees, E. M. and Harman, D. K. (Eds.), Proceedings of the Fifth Text Retrieval Conference (TREC-5) 20-22 November*, pp. 291-314.