April 2012

Hybrid Algorithm for Improving Efficiency of Keywords based Search Engine

K. F. Bharati

CSE Department, JNTUA College of Engineering, Anantapur, Andhra Pradesh, India,
kfbharathi@gmail.com

Follow this and additional works at: <https://www.interscience.in/ijcsi>

Part of the [Computer Engineering Commons](#), [Information Security Commons](#), and the [Systems and Communications Commons](#)

Recommended Citation

Bharati, K. F. (2012) "Hybrid Algorithm for Improving Efficiency of Keywords based Search Engine," *International Journal of Computer Science and Informatics*: Vol. 1 : Iss. 4 , Article 9.

DOI: 10.47893/IJCSI.2012.1051

Available at: <https://www.interscience.in/ijcsi/vol1/iss4/9>

This Article is brought to you for free and open access by the Interscience Journals at Interscience Research Network. It has been accepted for inclusion in International Journal of Computer Science and Informatics by an authorized editor of Interscience Research Network. For more information, please contact sritampatnaik@gmail.com.

Hybrid Algorithm for Improving Efficiency of Keywords based Search Engine

K. F. Bharati

CSE Department, JNTUA College of Engineering, Anantapur, Andhra Pradesh, India
E-mail : kfbharathi@gmail.com

Abstract - Web mining is the application of data mining techniques to discover interesting patterns from the Web. Web usage mining is the process of extracting useful information from server logs i.e users history. While discovering interesting patterns in multi agents the efficiency is decreased. In our paper presents hybrid algorithm is designed for the improvement of the efficiency of keywords based search engine. The model divides mining task into several parallel agents which coordinately work together, and the mining efficiency is improved greatly. The hybrid algorithm is Evolved from HITS, algorithm. Hybrid algorithm removes Link Farm pages in the expansion of root set, makes anchor text similarity calculation when crawling link page, and chooses pages by a brief conceptual analysis of page content. With the overcoming of the shortcomings of only text analysis or link analysis, Hybrid enhances the search engine in understanding the user interest and crawling more Web pages to meet the needs of the users.

Keywords - Web mining, Multi -Agents, Efficiency.

I. INTRODUCTION

Web mining is the application of data mining techniques to discover interesting patterns from the Web. Web usage mining is the process of extracting useful information from server logs i.e users history. While searching search engine has some problems. Actually, the search engines are important tools to help people search information on the large Internet. Search engine technology is constantly being improved and developed since the Internet was born. It developed from simple categories at the early stage to nowadays senior application services which use complex algorithms and large cluster systems. Some search engines have introduced intelligent information retrieval technology to improve the quality of information collection and query. Nowadays, Web information expands rapidly and Web search engines are facing severe situation. The latest survey data show that the number of public access Web pages increased by over 7 million every day. 70% of the information people demand comes from search engines, but only 60% of Web information can be indexed by six big search engines such as Yahoo, Google etc., and single search engine can only index one third of Web resources [1]. Web information retrieval has defects such as low coverage, low quality inquires, slow renewal speed, the lack of effective information classification, bad understanding of user behavior based on keywords and so on. Search engines index web pages by Web spiders

(also called web crawler or Web robot). A Web spider gets target Pages back without order. In order to improve searching speed, it rarely filters information when collecting message, even if it does filter, it is only based on some simple key words, so recall-precision and accuracy is low. If a search engine has the concept of understanding user intent, and we design web spiders as multi-agent mode which can work coordinately to filter Web pages, then it is possible to improve the efficiency of Web page gathering [2].

II. MULTI-AGENT TECHNOLOGY

The importance of gaining information from data has still been increasing over the last few decades, as we are literally flooded with large amounts of data. Discovering the patterns in data or generally getting some knowledge from them is therefore one of the most important goals of today's research on the field of artificial intelligence. It usually requires deeper understanding of both the data and the data mining methods, to be able to use them with satisfactory results. It should therefore be the next goal of artificial intelligence research to develop meta learning systems, which would learn from their previous experience, and which would be able to give advice on what methods to use in particular situations.

Research of agent and the multi-agent systems (MAS) has become a hot spot in Distributed Artificial

Intelligence (DAI) [3]. Agent technology provides a new calculation model and problem solving method, as well as a good remote intelligent program design method. The multi-agent system relaxes the limit of centralizing, planning and sequence control, and provides a decentralized control, emergency and parallel processing. It can also reduce the cost of hardware and software. So MAS provides more rapid problem solving speed. Agent has many features such as autonomy, interactivity, responsiveness, initiative, etc. Calculation entity with these characteristics may be UNIX process (or thread), computer system or simulation robot. A web spider extremely has such features so it can be designed as multiagent (often called multi-intelligent agent system) to make full use of the theory and method of intelligent agent.

III. WEB PAGE CRAWLING MODEL

At present, there are three kinds of methods to acquire web information. One kind is pure link analysis method (such as Page Rank [4]) whose defect is that it doesn't make full use of Web text content. Another kind is text analysis method (such as HITS [5]) whose defect is that it doesn't make good use of Web links. The third is a combination of link analysis and text analysis method which combines the advantages of the first method and the second method perhaps to makes better effect.

According to the basic principle of Multi- agents system, we designed a web page crawling model by using hybrid algorithm. For higher search efficiency, it expands the user's concept space and determines the user's interest degree based on the user's search demand and thus gets Web resources more quickly and accurately. It doesn't do the whole text analysis of the result pages but only guarantees that gathered pages are related with users' themes.

3.1. Structure of Hybrid Model

Web page crawler model that is used to increase the efficiency of keyword based search engines. Figure 1 is the kernel structure of IG model. The model includes four agent types: Manager Agent, URL agent, Spider agent and Hybrid agent.

The Manager agent is the main body which is responsible for generating and terminating other parallel running agents as well as managing and distributing system resources. As the URL task management center, URL agent manages URL database and is responsible mainly for URL task access, distribution and tidying. In order to improve efficiency, initial URL database only contains seeds URL (URL related to profession or domain). On one hand, URL agent reads a series of URL into its task queue and assigns to other spider

agents. On the other hand, it accepts new URL analyzed from other spiders and stores and updates database. In addition, Hybrid agent also analyzes anchor text links in page context to get its similarity with the user's query. When similarity is lower than the threshold, URL links will be removed from database. If the link is being crawled by a spider, that spider will be notified to stop working.

Manager agent starts up many spiders at the beginning. When the spider gets URL task, it begins to work, gathers all required web pages, analyzes URL information and sends them to URL agent and at last sends crawled contexts to Hybrid agent. The spider will ask for new task actively after it finishes one operation. Manager can adjust the number of generated spiders according to system status.

Hybrid agent is responsible for analyzing download Web pages, which is the kernel of IG. We do Web page analysis based on improved HITS algorithm to decide if the Web page is removed or not. Web pages left behind will be saved into database for the next process.

In figure 1, dotted line shows Manager Agent's control flow. Fine arrow shows URL flow including URL tasks and new URL. Thick arrow shows web data stream.

Fig. 1 : Web Page Crawling Model

In the web page crawling model, there are two kinds of operation mechanisms: parallel mechanism and distributed mechanism. If all agents run on one server, a parallel spider system is then composed. All agents share CPU and memory resources. Otherwise, it is a distributed spider system and all agents work synchronously and make full use of all the computer's resources.

Since agent has good communication ability, autonomous and cooperative ability, it could make good use of resources no matter running on single or multiple computers. When running on multiple computers, agents need simple communication between each other such as notification, request, commitment, and refusal. All other things will be done by agent itself. It is quite similar to people's interactive cooperation. There is no need to have the same operating system in each node since the system framework is simple and reliable.

3.2. Hybrid algorithm in IG model

Hybrid algorithm is the kernel part in IR model, which has improved HITS algorithm. It is shown in figure 2, a , h , S , R , $Sim1$, $Sim2$ and t are elements in Hybrid algorithm. Something should be noted as follows:

The algorithm gets root set using multi-intelligent agent spider in the form of concept subject. As concept subject is superior to key words, root set based on concept is also superior to that based on keywords.

When expanding the root set R , ordinary HITS put URL in pages to basic set S which may include some irrelevant pages to subject. Algorithm firstly removes link farm pages [6] and then analyzes anchor text of URL. If it is related to subject, it will be put into S directly. Otherwise page concept similarity will be calculated.

S is the base set including all of the pages that root set pages link to, a is authority weight vector, h is hub weight vector. Finally, the algorithm outputs a list of pages with large hub weights, and the pages with large authority weights for the given search topic.

Fig. 2 : Flowchart of Hybrid Algorithm

IV. EXPERIMENT AND CONCLUSION

Using two different algorithms (old algorithm and newly multi-agent designed algorithm, i.e. Hybrid) to crawl the page, since new algorithm is multi-agent designed, system performance has been improved greatly than before. Specific performance and conclusions are as follows:

Parallel execution between multiple agents has improved system efficiency. The average time mining seeds website has been dropped by 20%. The conclusion can be drawn that intelligent multi-agent technique can significantly improve the operation efficiency of search engines.

Though Hybrid algorithm' brief analysis of page content is at the cost of time, it improves the quality of page extraction. Assume that the length of a page is n , the handling of the page is limited to extraction of key words set which is not big, and time complexity of this algorithm is close to.

Mining algorithm of search engines needs further improvement, especially the improvement of mining intelligent degree. In the following research, we will give more domain knowledge to search engines based on clothing knowledge base. In that way, a spider will remove more junk web pages when crawling web sites and the recall effectiveness and accuracy will be greatly improved.

REFERENCES

- [1] Pierre Baldi, Paolo Frasconi, Padhraic Smyth. Modeling the Internet and the Web[M]. Chichester: John Wiley & Sons Ltd, 2003.
- [2] Wang Chuan, Chang Gui-ran, Ma Yan, et al. Research on search engines based on the user behavior model [J]. Journal of Information.
- [3] Shi Zhongzhi. Intelligent agent and its application [M]. Beijing: science press, 2000.
- [4] Brin S, Page L. The anatomy of a large-scale hypertextual Websearch engine[A]. In: Thistlewaite P, et al., eds. Proceedings of the 7th ACM-WWW International Conference[C]. Brisbane: ACM Press, 1998, pp. 107-117.
- [5] Kleinberg J. Authoritative sources in a hyperlinked environment [A]. In: Tarjan RE, et al., eds. Proceedings of the 9th ACM-SIAM Symposium on Discrete Algorithms[C]. New Orleans: ACM Press, 1997, pp. 668-677.
- [6] Dong Zhendong, Dong Qiang. Construction of a knowledge system and its impact on Chinese research [J]. Contemporary Linguistics, 2001, 3(1), pp.33-77.