Available online at www.sciencedirect.com

SciVerse ScienceDirect

Procedia Engineering 15 (2011) 2983 – 2987

Procedia Engineering

www.elsevier.com/locate/procedia

Advanced in Control Engineeringand Information Science

An Improved PageRank Method based on Genetic Algorithm for Web Search

Lili Yan[a], Zhanji Gui[a], Wencai Du[b,]*,Qingju Guo[a]

[a]*Department of Software Engineering ,Hainan College of Software Technology , Qinghai 571400, China*
[b]*College of Information Science & Technology, Hainan University,Haikou 570228, China*

Abstract

Web search engine has become a very important tool for finding information efficiently from the massive Web data. Based on PageRank algorithm, a genetic PageRank algorithm (GPRA) is proposed. With the condition of preserving PageRank algorithm advantages, GPRA takes advantage of genetic algorithm so as to solve web search. Experimental results have shown that GPRA is superior to PageRank algorithm and genetic algorithm on performance.

Selection and/or peer-review under responsibility of [CEIS 2011]

Keywords:Web search, relevance ranking, ageRank, hypertext induced topic search;

1. Introduction

Web search engine has become a very important tool for finding information efficiently from the massive Web data. As the explosive growth of the Web data, the traditional centralized search engines become harder to catch up with the growing step of people's information needs. Search engines have become an important network information navigation tool that helps people in the massive Web data quickly and easily find the information they need. With the Web's continued rapid growth in the amount of data the traditional centralized search engines are increasingly not meet increasing access to information people need. the one hand, a centralized search engine, server processing capacity is limited, currently one of the best search engine Google to use on the million PC cluster consisting of servers, and

* Corresponding author. Tel.:13807680220;
E-mail address:softhainan@yahoo.com.cn.

doi:10.1016/j.proeng.2011.08.561

can only index the entire Web page of the total about 1/10, not including the number of surface Web 400 ~ 500 times the deep web, and the current centralized search engines longer than the data update cycle, it is difficult to meet people timeliness of information needs. On the other hand, by the Web crawler information collection capacity constraints, the depth of traditional search engine is difficult to dig deep web information.

According to different technical principles, the search engine can be divided into: crawler-type search engine, directory search engine, Meta search engine. Crawler is to roam the Web and found the download page of the computer program, also known as spiders, Robot [1]. Crawler automatically crawl the Internet, the search page to automatically download to a local database, the index is to provide a user search service. Directory search engine search by way of manual or semi-automatic, collecting and editing, organizing information, support classification browsing, keyword search. Meta-search engine database is not page when a user submits a query request; it forwards the user query to multiple other search engines to return multiple results, the merge return to the user. Most meta-search engines to extract only the results of each search engine in front of 10 to 50 pieces of information. Features based crawler, search engines can be divided into: General on crawling (Scalable Web Crawler) search engine, topic-based crawling (Focused Web Crawler) search engine, based on the individual crawler (Customized Web Crawler) search engine Based on Intelligent Agent (Agent based Web Crawler) search engine crawler Based Migration (Reloadable Web Crawler) search engine for the deep Web information (Deep Web Crawler) search engine. Fig.1. is the IBM Focused crawler.

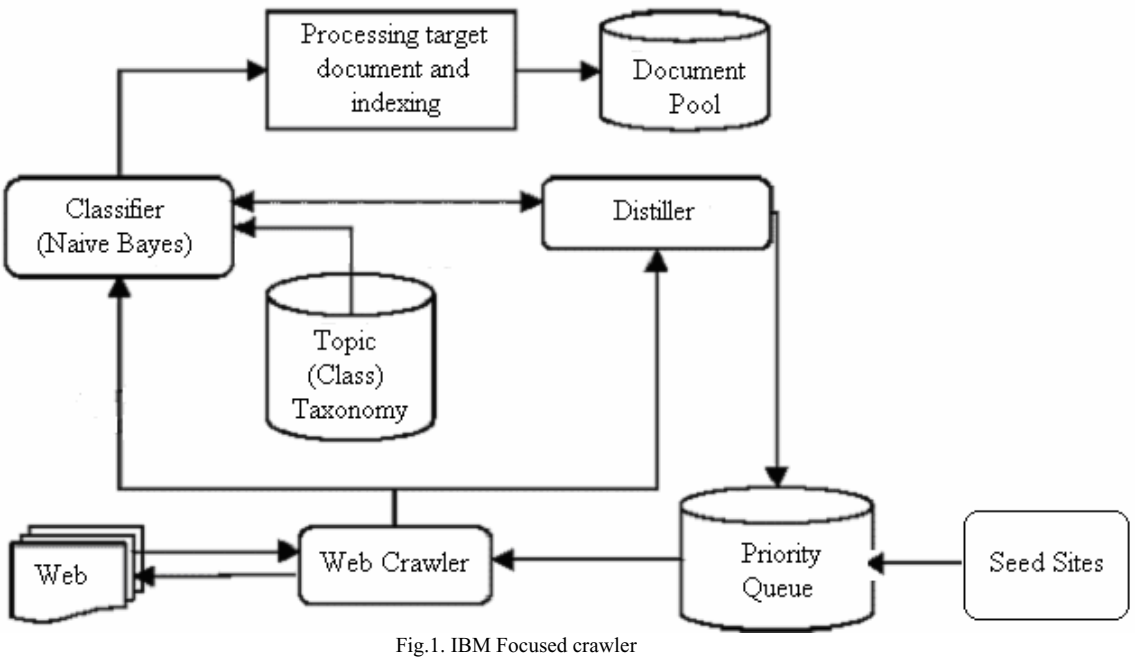


Fig.1. IBM Focused crawler

This paper is organized as follows: a brief introduction to relevance ranking the related works is given in Section 2. Section 3 describes the application of genetic PageRank algorithm (GPRA), and presents the experimental results from using the proposed method. Section 4 summarizes the results and draws a general conclusion.

2. Relevance Ranking

Correlation refers to the search terms and page relevance. Relevance ranking technology generation is mainly determined by the characteristics of search engines [2]. First of all, the modern search engines can access the number of Web pages have reached the scale of billions, even if the user is only one small part of the search, text search technology based search engine can return thousands of pages. Even if these are the results page the user needs, the user is not possible for all of the web browser again, so that will be most interested in the results page the user put in front of the search engines is bound to enhance customer satisfaction. Second, the search engine to retrieve the user's own expertise is usually limited, in the most popular keyword search behavior, users who typically only a few key words. Relevance ranking is based on documents the correlation between the query on the query results sort. In the traditional information retrieval, relevance evaluation model includes Boolean Model, vector space model, and the probability model and developed on this basis a series of extensions model [3].

Relevance ranking technology at this stage are mainly the following: First, traditional information retrieval techniques based on the way, it is mainly the key words in the document itself, the importance of the document with the user query to the requirements of relevance to make measurements, such as Using a Web page keywords the frequency and location. In general, the retrieved web documents containing the query keywords in the number of the more relevant the larger and higher the degree of this key distinction; Second, hyperlink analysis, the use of the representativeness of this technology search engine Google and Baidu, etc. there. And the former than it recognized the importance pages as search results sorted by relevance. From the design point of view, it is more emphasis on third-party recognition of the pages, such as pages with large chain network's website is widely recognized the importance of web pages [4].

Relevance ranking technology is mainly dependent on the hyperlink analysis technology. Hyperlink analysis can provide a variety of functions, including the results of the main function is to solve the problem pages in order of relevance. It is mainly the use of hyperlinks between Web pages point to the existence of various, on page analysis of the relationship between the references, based on the number of how many page chains calculate the importance weight of the page. Generally believed that, if a web page has hyperlinks to point B, the equivalent of page A page B cast one vote, that is, A B recognized the importance of web pages. In-depth understanding of hyperlink analysis algorithms, based on the entire Web page link structure as a collection of documents to the topology, where each page constitutes a node in the figure, the links between pages to form between the end points directed edges, in accordance with this idea, based on each node of the out-degree and degree to evaluate the importance of web pages. For the hyperlink analysis, a representative algorithm is mainly designed PageRank Page, etc. to create the HITS algorithm and Kleinberg algorithm. Which, PageRank algorithm in actual use is better than in the HITS algorithm, which is mainly for the following reasons: First of all, PageRank algorithm can be a one-time, offline and independent of the query operators on Web pages are expected to get an estimate of page importance, then in the specific user query, in conjunction with other query index values associated with the query results are sorted, thus saving the cost of the system queries the operator; Secondly, PageRank algorithm is calculated using the entire collection of pages, unlike HITS algorithm vulnerable to the effects of local links to the trap produce "topic drift" phenomenon, so the technology is now widely used in many search engine, Google search engine wide success to hyperlink analysis also shows that the page is characterized by correlation become more sophisticated sorting algorithm[5].

3. Genetic Pagerank Algorithms

PageRank [6] and HITS (Hypertext Induced Topic Search) [7] is a traditional centralized search engine links between the two most important evaluation algorithm, the sorting of search results they play a very important role. The main idea of PageRank algorithm is a page is referenced several times, then it

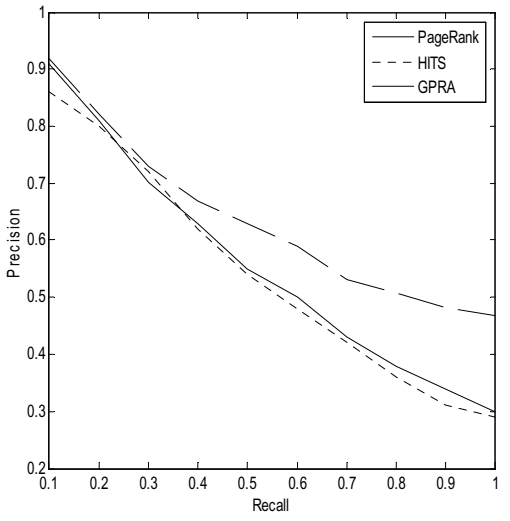
may be very important; a web page reference is important, it may be very important; the importance of a page is passed to it refers to the average page.

$$PR(i) = (1\text{-}d) + d \cdot \sum_{j \in B(i)} \frac{PR(j)}{N(j)}$$

in which , $B(i)$ on behalf of a collection of pages pointing to page i, $N(j)$ indicate that the page j in the number of hyperlinks pointing to other pages, $PR(j)$ the authority of that degree of page j, $d(0 < d < 1)$ is an attenuation factor, the best value is 0.85.

Topic-Sensitive PageRank algorithm basic idea is to compute a PageRank vector off-line collection, the collection with a theme each vector, i.e. the calculation of a page's score on various subjects [8, 9]. Divided into two phases: a collection of themes related to the calculation of PageRank vectors and online inquiries to determine the theme. It is based on the user's query request and the related user queries to determine the context of related topics (the user's interest) to return the query results with high accuracy.

We select such a subset using a genetic algorithm (GA). A genetic algorithm is a search technique used in computing to find exact or approximate solutions to optimization and search problems. Genetic algorithms are categorized as global search heuristics. Genetic algorithms are a particular class of evolutionary algorithms (EA) that use techniques inspired by evolutionary biology such as inheritance, mutation, selection, and crossover. Several reasons motivate this choice. First, the use of met heuristic techniques is well established in optimization problems where the objective function and the constraints do not have a simple mathematical formulation. Second, we have to determine a good solution in a small computing time, where the dimension of the problem may be signi3cantly large. A standard representation of the solution is as an array of bits. Arrays of other types and structures can be used in essentially the same way. The main property that makes these genetic representations convenient is that their parts are easily aligned due to their fixed size, which facilitates simple crossover operations. Variable length representations may also be used, but crossover implementation is more complex in this case. Third, the structure of our problem is straightforward represent able by the data structure commonly used by a GA [10].



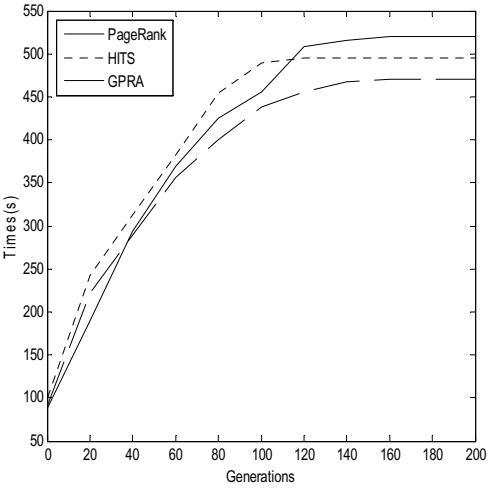


Fig.2. the three algorithms for recall and precision

Fig.3. the three algorithms for execution time

PageRank, a measure of web pages' relevance introduced by Brin and Page, is at the heart of the well known search engine Google [6]. Google classifies the web pages according to the pertinence scores given by PageRank, which are computed from the graph structure of the Web. A page with a high PageRank will appear among the first items in the list of pages corresponding to a particular query. PageRank is a link structure-based algorithm, which gives a rank of importance of all the pages crawled in the Internet by the Google's web crawler. To compute a PageRank is actually to compute the stable distribution of a transition matrix, also called the Google matrix, which is based on the web graph structure. The overwhelming size of the Google matrix pales many popular and intricate algorithms, which may otherwise have excellent performances in those normal scale computing. By comparison, the simple power method stands out for its stable and reliable performances in spite of its low efficiency. Fig.2-3 is the experimental results.

4. Conclusions

With the rapid growth of Internet, the WWW provides an important channel for user to obtain useful information. A new PageRank algorithm is pulled into to improve the original PageRank algorithm. By illustrating examples, we verify the new algorithm's effectiveness. GPRA takes advantage of genetic algorithm so as to solve web search. Experimental results have shown that GPRA is superior to PageRank algorithm and genetic algorithm on performance.

Acknowledgements

This research work was supported by National Natural Science Foundation of China (Grant No. 60963025), the Natural Science Foundation of Hainan Province (Grant No. 610229). The authors would like to thank the anonymous reviewers for their insightful comments and constructive suggestions that have improved the paper.

References

[1] Nie Zaiqing, Kambhampati S, and Nambiar : Effectively mining and using coverage and overlap statistics in data integration. IEEE Transactions on Knowledge and Data Engineering, 2005, 17(5): 638-651.

[2] Hernandez T and Kambhampati S. Improving text collection selection with coverage and overlap statistics. In: Poster Proceedings of the 14th International Conference on World Wide Web (WWW2005). New York: ACM Press, 2005. 1128-1129

[3] Zhong Ming, Moore J, Shen Kai and Murphy A. An evaluation and comparison of current peer-to-peer full-text keyword search techniques. In: Proceedings of the 8th International Workshop on Web and Databases (WebDB2005). New York: ACM Press, 2005. 61-66.

[4] Shi Shuming, Yang Guangwen, Wang Dingxing Yu Jin, Qu Shaogang, and Chen Ming. Making peer-to-peer keyword searching feasible using multi-level partitioning. In: Proceedings of the 3rd International Workshop on Peer-to-Peer Systems (IPTPS'04). Berlin: Springer-Verlag, 2004. 151-161.

[5] Sripanidkulchai K, Maggs B and Zhang Hui. Efficient content location using interest-based locality in peer-to-peer systems. In: Proceedings of the IEEE INFOCOM'03 Conference. San Francisco: IEEE Computer Society, 2003. 2166-2176.

[6] Page L, Brin S, Motwani R, and Winograd T. The pagerank citation ranking: Bringing order to the web. Technical Report. Stanford University, 1998.

[7] Kleinberg J. Authoritative sources in a hyperlinked environment. Journal of the ACM, 1999, 46(5): 604-632.

[8] C. de Kerchove, L. Ninove, Maximizing PageRank via out-links, Lin. Alg. Appl., 429(2008): 1254-1276.

[9] H. Sun, Y. M. Wei, A Note on the PageRank algorithm, Appl.Math. Comp., 179(2006): 799-806.

[10] E. Ward, Market through 'link analysis' to improve, popularity quality. Advertising Age's Business Marketing, 85(2000),32-33.