Volume 4, Issue 1, January 2014 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

Keyword Optimization for Web Crawler: A Review

Balinder Singh **Sanjeev Kumar***

Abstract— In this paper we provide several dimensions of web information retrieval using internet search engine. Increasing growth and updating in web world yield big challenges for search engine to retrieve the required information. The large size and dynamic nature of web highlights the need of continuous support and updating of web based information retrieval system. Keyword optimization supports the search engine to find effective information retrieval from dynamically changing web environment. This paper provides study of existing information retrieval techniques and explores new methods of information retrieval. Keyword optimization will be implemented by using genetic algorithms of proposed information retrieval techniques.

Keywords—Keyword Optimization, Crawler, Genetic Algorithm and Similarity Measure.

I. INTRODUCTION

Search engine is a tool that accesses the information from the web. This provides the preferred entry point of pages on the web. The World Wide Web is a vast source of information and this information is often scattered among many web servers and hosts. This information is retrieved using many different formats and search engines. Some basics search engines are Google, Yahoo, AltaVista, Ask.Com etc.

A. Keyword Optimizations

It would not be a very wise decision to rush into search engine marketing without having a set of relevant keywords ready beforehand. Search engines search information on web related to the keywords type in search toolbar. The web surfer type keywords and the search engines show results after analyzing the relevance of those keywords in the web world [4]. The systems first extract keywords from documents and then assign weights to the keywords by using different approaches. In this system only two things is important: One is how to extract keywords precisely and the other is how to decide the weight of each keyword [2].

B. Web Crawlers

A web crawler is one type of software agent also called a component of search engine. In general, the crawler create a list of URLs to visit, called the seeds. When the crawler visits these URLs, it matches all the hyperlinks in the page and become a list of URLs to visit, called the crawl frontier. Crawler is one of the most critical elements in a search engine [17]. It traverses the web by following the hyperlinks and storing downloaded documents in a large database that will later be indexed by search engine for efficient responses to users' queries [18].

Issues to be Consider: According to the keywords of information retrieval, there are following issues:

a. Where to start crawling?
b. Which link do you crawl next?
c. What pages should crawler download?
d. How to minimize the load on visited pages?

We can conclude the similarity of web pages and how much keyword is According to these issues [5]. Starting with an initial set of keywords, our system expand the set by adding the most suitable term that intelligently selected during the crawling process by genetics algorithm [6]. Then system will gather, sort and organize our keyword into a high performance SEO [7]. As keyword based relevance ranking does not guarantee relevance in meanings, different semantic models have been introduced to improve the relevance ranking [9].

To improve the efficiency of retrieval system, it has been proposed that the documents which are generally retrieved together in response to some query should be kept close together within the system in the form of clusters [10, 11]. A cluster based search proceeds to satisfy a query efficiently by identifying and retrieving only those clusters which exhibit a sufficiently high degree of match with the query[13]. Clustering improve the effectiveness of retrieval as it results in the retrieval of a higher number of relevant documents for a given amount of effort [12]. Due to the increasing capacity of storage devices and their decreasing cost, there is tremendous growth in database of all sorts. This explosive growth has led to huge, fragmented database and it makes extremely difficult to retrieve relevant information from these large document collections [1]. Information Retrieval is a system which gets the information from the web related to a specific keyword in an order according to their rank. Information Retrieval System is a system used to store items of information that need to be processed, search and retrieved corresponding to the user's query [2]. The general objective of

Information Retrieval System is to minimize the overhead that can be express at the time a user spend in all of the steps leading to reading an item containing the needed information [3]. This document is a template. An electronic copy can be downloaded from the Journal website. For questions on paper guidelines, please contact the journal publications committee as indicated on the journal website. Information about final paper submission is available from the conference website.

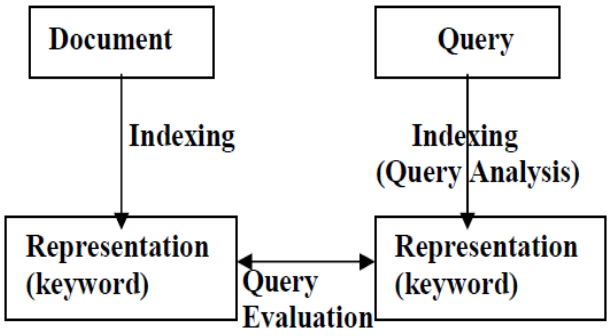

Fig. 1 Indexing based IR

Fig.1 show that how to best represent their contents and query evaluation show that to what extent does a document corresponds to a query.

The goal of an Information Retrieval System (IRS) is to help a user to locate the most similar document that have the potential to satisfy the user information needs. The focus of information retrieval is the ability to search for information relevant to a user's needs within a collection of data which is relevant to the user's query [8].

There are three types of model use for information retrieval: Boolean Model, Probabilistic model and Vector space Model we will use Boolean model for IR.

1) Boolean Model: Boolean model is a first information retrieval model. Boolean model is a most common exact match model. In a Boolean model the query term and their corresponding set of documents can be combined in a form of new set of documents. In a document representation these document is convert into a binary form and use logical operator like AND, OR, and NOT [2]. A document is represented as a set of keywords. Queries convert into a Boolean expressions of keywords, connected by AND, OR, and NOT, including the use of brackets to indicate scope. Boolean retrieval gives the output as the document is matching fully or not but no partial matching. Boolean models can be extended to include ranking.

IR Challenges:-

I. User and context sensitive retrieval.
II. Multi-lingual and multi-media issues.
III. Better target tasks.
IV. Improved objective evaluations.
V. Substantially more labeled data.
VI. Greater variety of data sources.
VII. Improved formal models.

II. BACKGROUND

A. Information Retrieval

A Ahmed A. A Radwan [2] represent a new fitness method gives more sophisticated results than a cosine fitness function in the test collection. Parveen Pathak [1] described that it is difficult to retrieve relevant information from large document collections. It was present three important paradigms of research in the area of information retrieval (IR): Probabilistic IR, Knowledge-based IR, and, Artificial Intelligence based techniques like neural networks and symbolic learning. Ao-Jan Su [14] described that a focus on the Google ranking algorithm and design, implement and evaluating the ranking system to systematically validate assumptions others have made about this popular ranking algorithm. S. Siva Sathya [3] described three features to get the effective information retrieval: First is to extract keywords and other information from the database by a document crawler. Second is to generate the combination terms using genetic algorithm. Third results generated from the GA are applied to information retrieval system to generate better results. Nguyen Quoc Nhan [6] described that a focused crawler also give the result related to the keyword. From the experimental results, it was presented that Focus crawling approach is better when the number of Web pages is small.

B. Web Crawler

Milad shokouhi [15] described that Crawler estimate the best path for crawling on one hand expands its initial keywords by using a genetic algorithm during the crawling on the other hand. Cui Xiaoqing described that Adjust jaccard algorithm is better in compare to cosine and jaccard method. Carlos Castillo [16] proposed a comparative study of strategies for Web Crawling , that is combination of breadth first order with the largest site first is a practical alternative since it is fast, simple to implement, and able to retrieve the best ranked pages at a rate that is closer to the optimal than

other alternatives. Different author use different techniques to increase the relevancy. Some use traditional method and some use genetics algorithm to increase the relevancy. Basically similarity measures are used to find out the relevancy between documents which are downloading by crawlers in response to query and user query.

III. METHODOLOGY

First we make background study of fundamentals of IR. Then we use SEO tool for IR. Then we compare the relevancy of documents. Simulation Modeling and Analytical Modeling will be considered for information retrieval.

A. Genetic Algorithm

GA is a probabilistic algorithm based on the principle of heredity and evolution which claims in each generation stronger individual survives and weaker dies. In a Genetic Algorithm problem, the keyword use as a search space and represented as a number of individuals called chromosomes these are the initial population and the aim is to obtain a set of qualified chromosomes after some generations.

B. Similarity Measures

Similarity measures are functions which are used to find out the similarity between two objects and documents. In information retrieval system similarity measures are used to find out the relevancy between the user query and documents which are extracted by crawler.

Similarity measures:

Jaccard coefficient: $sim(x,y) = (\left| x \cap y \right|) \div |xUy \left|$

C. GA Design

1) Encoding: In GA, search space is composed of candidate solution to the problem each represented by a string is termed as a chromosome. Representation of chromosome is called encoding Binary Encoding Scheme.

2) Fitness Function: Fitness function is a performance measure which evaluates how each solution is good. Similarity measure is used as a fitness function and jaccard coefficient is used as a fitness function.

3) GA operators: How to select individuals for the next generation is one of the important steps in GA design.

- *Selection:* Selection operator selects these individuals which have higher fitness value.
  - Tournament selection
  - Roulette wheel selection
- *Crossover:* After selecting best individuals crossover operator is applied. Crossover is the Genetic operator that mixes two or more chromosomes to form new offspring.
  - One point crossover
  - Multi-parent crossover
- *Mutation:* It produces new genetic structure by randomly changing some of its building blocks or by altering a random bit.

IV. CONCLUSION AND PROPOSED WORK

Due to the change of web day by day web information retrieval is become a major problem and using keyword optimization this problem can be removed. The future work of keyword optimization is effective information retrieval using genetics algorithm and similarity function. Initially we download a set of document from seed URLs. Based on weight scheme, we mark all the document in the above set and select n document at highest marks. Then we take m words from each document basis of maximum word frequency in a document. Then combine all words of n document and become a single keyword and convert these words into Boolean model 0 and 1 form. Using the genetics algorithm and jaccard similarity function, we find the fitness value of each document.

REFERENCES

[1] Parveen Pathak, Michael Gordon, Weiguo Fan. "*Effective information retrieval using genetic algorithms based matching functions adaption*" proceedings of the 33rd Hawaii International Conference on System Sciences- 2000.

[2] Ahmed A.A. Radwan, Bahgat A. Abdel Latef, Abdel Mgeid A. Ali, Osman A. Sadek. "*Using Genetic Algorithm to improve information retrieval systems*" proceeding of the world academy of Science, Engineering and Technology 17,2006.

[3] A. S. Siva Sathya, B.Philomina Simon. "*A document retrieval system with combination terms using genetic algorithm*", proceeding of the international journal of computer and electrical engineering, Vol. 2, No.1, February, 2010.

[4] Keyword optimization, *http://en.wikipedia.org/wiki/Search_engine_optimization.*

[5] Huo Ling Yu, Liu Bingwu, Yan Fang. "*Similarity computation of web pages of focus crawler*" proceeding in the 2010 International forum on Information Technology and Applications 2010 IEEE.

[6] Nguyn Quoc Nhan, Huynh Thi Thanh Binh, Vu Tuan Son, Tran Duc Khanh. "*Crawl topical Vietnamese web pages using genetic algorithm*" proceeding of the 2010 Second international conference on knowledge and systems engineering.

[7] Huilian Fan, Guangpu Zeng, Xianli Li. "*Crawling strategy of focused crawler based on Niche genetic algorithm*" proceeding of the 2009 eighth IEEE international conference on dependable, autonomic and secure computing.

[8] Stoney G degeyter, Jason Green. "*Keyword Research and Selection*" www.polepositionmarket.com , www.emarketingperformance.com .

[9] Rui Huang, Fen Lin, Zhongzhi Shi, "*Focused Crawling with heterogeneous Semantic Information*" 2008 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology.

[10] E.M. Voorhees. "*The cluster Hyppthesis Revisited*" Proceedings of the Eighth ACM SIGIR, pages 188-196. Montreal, Quebec, Canada, 1985.

[11] C.T. Yu, Y.T. Wang, and C.H. Chen. "*Adaptive Document Clustering*". In Proceeding of the Eighth ACM SIGIR, pages 197-203, Montreal, Quebec, Canada,1985.

[12] C.J. van Rijsbergen. "*Information Retrieval*" Butterworth Publishers, Boston, MA, end edition, 1981.

[13] Jay N. Bhuyan, Jitender S. Deogun, Vijay V. Raghavan, "*Cluster-Based Adaptive Information Retrieval*", proceeding in 0073-1129/91/0000/1991 IEEE.

[14] Ao-Jan Su, Y.Charlie Hu, Aleksandar Kuzmanovic, and Cheng-Kok Koh, "*How to Improve Your Google Ranking: Myths and Reality*", 2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology.

[15] Milad shokouhi, Pirooz Chubak, Zaynab Raeesy, "*Enhancing Focused Crawling with Genetic Algorithms*" Proceedings of the International Conference on Information Technology: Coding and Computing(ITCC'05) 0-7695-2315-3/05 IEEE.

[16] Carlos Castillo, Mauricio Marin, Andrea Rodriguez, "*Scheduling Algorithms for Web Crawling*" Proceedings of the WebMedia & LA-Web 2004 Joint Conference 10th Brazilian Symposium on Multimedia and the Web 2nd Latin American Web Congress (LA-Webmedia'04) 0-7695-2237-8/04 IEEE.

[17] S. Brin, L. Page, "*The anatomy of a Large Scale Hypertextual Web Search Engine*", Computer Science Department, Standford University, Standford, USA, 1998.

[18] T. Suel, V.Shkapenyuk, "*Design and Implementation of a High-Performance distributed Web Crawler*", In Proceedings of the 18th International Conference On Data Engineering, San Jose, CA,2002.

