ISSN: 2278 – 1323
International Journal of Advanced Research in Computer Engineering & Technology
Volume 1, Issue 4, June 2012

Improved Focused Crawler Using Inverted WAH Bitmap Index

Sanjay Kumar Singh, Sonu Agrawal

Abstract— **Focused Crawlers are software which can traverse the internet and retrieve web pages by hyperlinks according to specific topic. The traditional web crawlers cannot function well to retrieve the relevant pages effectively. The focused crawler is a special-purpose search engine which aims to selectively seek out pages that are relevant. The main characteristic of focused crawling is that the crawler does not need to collect all web pages, but selects and retrieves only the relevant pages. So the major problem is how to retrieve the maximal set of relevant and quality pages. To address this problem, we have designed an Interactive focused crawler which calculates the relevancy of web page. It calculates the URL score for identifying whether a URL is relevant or not for a specific topic. The Interactive Focused Crawler proceeds by gathering pages related to the seed set by using techniques like keyword extraction and search engine query and link neighbourhood expansion. These collected pages are then prompted to the user in a ranked order that facilitates quick elimination of negatives. The user then provides feedback and helps the baseline classifier to be progressively induced using active learning techniques. Once the classifier is in place the crawler can be started on its task of resource discovery.**

Index Terms— classifier, focused crawler, keyword extraction, , URL.

.

I. INTRODUCTION

The World Wide Web is huge. It is popular and interactive medium for resource discovery. The WWW is growing rapidly. The Web resources are well structured by the hypertext and the hypertext can be used to determine the relevance of the document to the search domain. The storage and computational resources are limited and the nature of web is dynamic, therefore search engines cannot index every web page. Therefore it is very important to develop effective agent to conduct real time searches for users. For searching a particular topic in internet, we require a system whose goal is to index web information according to specific topic. Web crawling is one of main component in web information retrieval. The Crawling process of the entire web is an unrealistic and expensive because of required hardware and network resources. A Focused Crawler is a hypertext resource discovery system that targets relevant pages of a particular topic. To Surfing the WWW, focused crawler predict priority based order of visiting hyperlinks for downloading relevant documents.

Sanjay Kumar Singh, Department of Computer Science and Engineering, SSCET Bhilai,. India, Phone : 09575620964 (e-mail: sanjay16kumar@gamil.com)
Sonu Agrawal, Department of Computer Science and Engineering, SSCET Bhilai, India, Phone: 09926848558 *(e-mail: agrawalsonu@gmail.com).*

The search engines use simple crawling strategy to retrieve all website, they expect a high efficient focus crawler, to retrieve the topic – specific web pages. There are two problems about focused crawlers to be presented: 1) selection procedure of webpage 2) Revisiting procedure of webpage. The aim of our project is to develop a focused crawler using Inverted WAH Bitmap Searching User Defined Document Fields to find topic-related webpages for the end users.
In addition, the advantage of this system in Web information retrieval is to produce more comprehensive search in the WWW and improve the performance, efficiency of the search engine.

II. RELATED WORK

Focused crawling was first introduced by Chakrabarti in 1999 [2]. The fish-search algorithm for collecting topic-specific pages is initially proposed by P. DeBra et al. [3]. Based on the improvement of fish-search algorithm where page relevant score is based on two values either 0 (for irrelevant pages) or I (for relevant pages), M. Hersovici et al. proposed the shark-search algorithm [4] which is based on fuzzy set values that fetches more relevant pages than the fish-search approach. An association metric was introduced by S. Ganesh et al. in [5]. This metric estimated the semantic content of the URL based on the domain dependent ontology, which in tum strengthens the metric used for prioritizing the URL queue. In [9], Yunming Ye et al. presented isurfer, a focused crawler that uses an incremental method to learn a page classification model and a link prediction model. It uses an online sample detector to incrementally distill new samples from crawled web pages for online updating for the model learned. The Link Structure-Based method is analyzing the reference information among the pages to evaluate the page value.

These kind of famous algorithms are the Page Rank algorithm [6] and the HITS algorithm [7]. There are some other experiments which measure the similarity of page contents with a specific subject using special metrics and reorder the downloaded URLs for the next crawl. In order to find pages of a particular type or on a particular topic, focused crawlers aim to identity links that are likely to lead to target documents and avoid links to off topic branches. A major problem faced by above focused crawlers are that all focused crawlers measure the relevancy of a page and calculate the URL score based on whole page's content and a Web page usually contains multiple topics, not all of which are related to the given domain. The evaluation on the whole page may cause a lot of irrelevant links crawled first, because most Web pages are dirty from the point of view of content and some noises such as navigation bar, advertisement and logo usually exist in Web pages. Meanwhile, the evaluation

on the context of links may ignore some relevant links which has little information [8]. These features bring difficulties for focused crawlers to compute the relevance of Web pages but in our proposed approach, we calculate the relevancy of block and link score based on relevancy of content blocks of web page. For a block which is irrelevant to topics, we do not calculate the URL score of that URL which belong to this block because there are number of URLs in a single web page. So, we must reduce the time to identity the URL which IS more relevant to topics. A group of researchers proposed an ontology focused crawler in which two cycles are involved in the crawling framework. Based on the ontology and crawling scope, the focused crawler starts to work on retrieving data from those websites. The focused crawler is used to retrieve, cluster and store relevant webpages by linking them to topics. Relevance values between the terms and document texts are computed from both global and local perspective, by means of the topical Page Rank algorithm [7]. From the existing works, we can observe that most of these crawlers utilize various ontology document link analysis technologies to control crawling scope and retrieve web documents according to users' specific interest.

III. METHODOLOGY

We are using Word-aligned hydride code (WAH) bitmap indexing strategy to implement our Interactive Focused Crawler. WAH encodes the bitmap in the words. Bitmap indexes use one bitmap for each distinct value. Our implementation is based on three steps:

1) Setting up Focused Crawler
2) Learning the base line classifier
3) Monitoring the Crawl

WAH bitmap perform the following operation

1) Uses run-length encoding for long sequences of identical bits
2) Encode / decode bitmaps in word size chunks
3) Designed for minimal decoding to gain speed

Figure1 and Figure 2 shows the Fraction of Time Spent in CPU that compared to two implementations of BBC, WAH spends smaller fraction of time in CPU

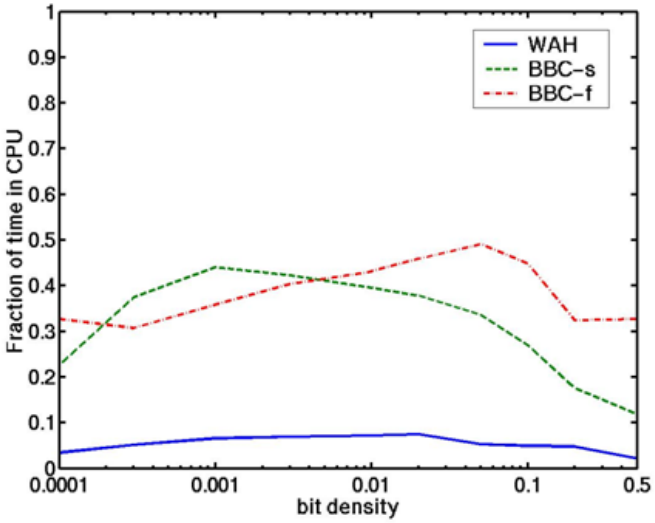

Figure1 : On a 2 MB/s disk system

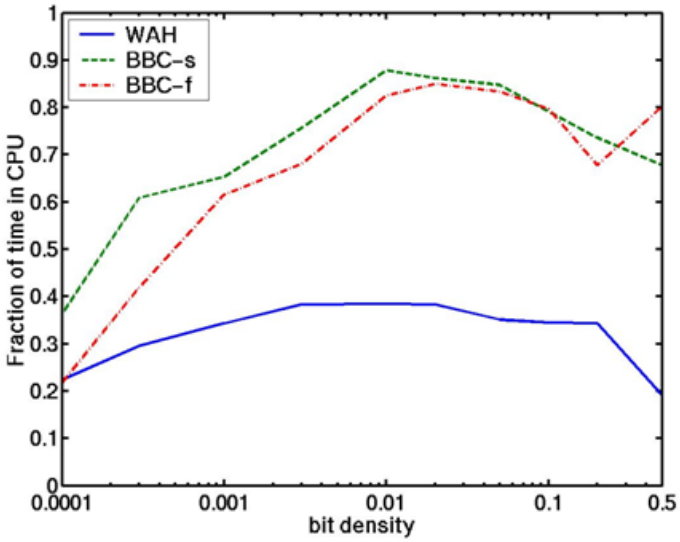

Figure2 : On a 20 MB/s disk system

Logical Operation Time

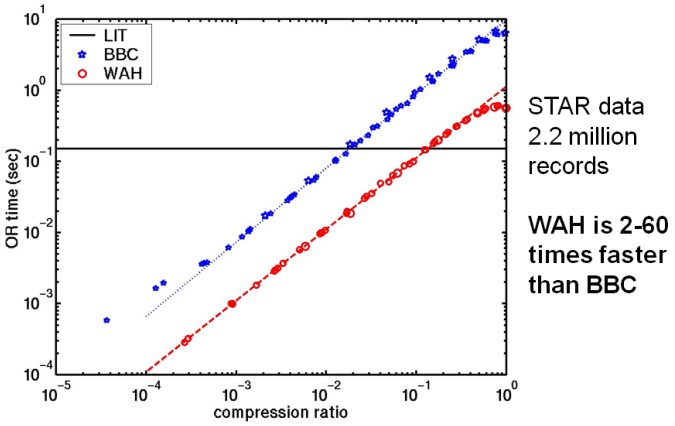

Figure 3 : Logical Operation Time

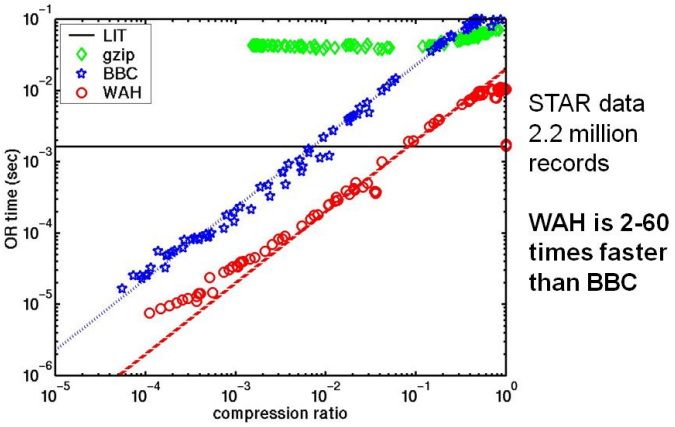

Figure 4 : Logical Operation Time

Advantages of using WAH Bitmap indexing

- Main operations are bitwise logical operations and they are fast.
- Index sizes are small for categorical attributes with low cardinality.
- Each individual bitmap is small and frequently used ones can be cached in memory.
- WAH compressed indexes are faster than BBC compressed indexes (3X) and uncompressed indexes (3X).
- WAH compressed indexes are 10X faster than ORACLE, 5X faster than BBC.
- Compressed bitmap index is more efficient for range queries than B+-tree or no index (p scan).

- A WAH compressed index uses more space than a BBC compressed index, but is more efficient.
- The size of WAH compressed bitmap index is modest even in the worse case.
- The WAH compressed index is efficient on attributes of any cardinality.

Setting the Focused Crawler

Techniques decided to set up the Focused Crawler for effective and efficient resource discovery. Steps carried out in the scout procedure.
A. Collect seed URLs from the user.
B. Gather related pages through keyword extraction and search engine query, consulting directories, and forward/backward link expansion. We denote these gathered related pages as the collected set or the expanded seed set.
C. Order the collected pages in decreasing similarity (w.r.t. the seed URLs) scores and prompt to the user for a threshold.
D. Learn a classifier from the user indicated positive and negative examples using active learning. We elaborate the above steps in more detail in the following sub-sections:

Keyword/Phrase Extraction and Search Engine Query

1. Keyword Extraction
2. Phrase Extraction
3. Constructing Search Engine Queries

Keyword Extraction

In this section we discuss some experimental results from the Keyword Extraction module. The Development Gateway Foundation (DGF) had given us 5568 URLs and wanted us to do topic discovery using these pages. We thought this would be an excellent oppurtunity to pilot run our proposed techniques and observe the gained results. The content in these pages revolve around:

- Applications of IT for development of rural areas in developing countries
- Building a virtual resource network to provide advice/expertise and favor knowledge sharing
- Knowlege creation and sharing with e-governance strategies.
- Digitisation of local content, increased connectivity, more access.
- A Global initiative by world bank to bridge the digital divide.

Phrase Extraction

Phrase Extraction can be regarded as a precision enhancing technique for retrieval responses. It can be used to identify an important concept which may be overlooked by single term queries. Though in the past the results for use of phrases in traditional Information Retrieval(IR) haven't been impressive, it would be helpful to know their behaviour in the context of the web. We plan to use two term syntactic and statistical phrases and see how they fare? We would postpone this feature and add it later as an enhancement since the main goal is to have a working system for quick experimentation and deployment.

Constructing Search Engine Queries

The selected keywords may be high in number. Making a query with all the keywords together may not gather enough responses. Also Google1 doesn't allow more than 10 words in a single query. On the other hand, making single word or two-word queries may fetch irrelevant results. Another problem is which words should be queried together. A good approach would be to construct queries using words that usually appear together in the seed documents. Following on these lines, we cluster the keywords with the document IDs as their attributes and construct a dendogram. Now we start from the root, form
a query constituting of all words at the root node and check if enough responses are retrieved. If not, we progressively move a level down forming queries for each of the nodes at that level and checking till enough responses are retrieved. The final collected responses are added to the expanded seed set.
In these particular steps we will collect seed URLs from the user, gathering related pages through keyword extraction and link expansion. We will order the collected pages based on similarity scores and learn the classifier.

IV. CONCLUSION

The World Wide Web is spreading widely and web content increases tremendously. Hence, there is a great requirement to have system that could list relevant web pages accurately and efficiently on the top of few pages. Mostly search engines used simple focused crawler but users may not get required documents easily. With a view to resolve the existing problems, a new Interactive Focused Crawler has been proposed which produces the index of Web Information resources. This system will improve the order of web pages in the result list so that user may get the relevant pages easily.

REFERENCES

[1] http://www.cs.uiuc.edul-dengcai2NIPSNIPS.html.
[2] S. Chakrabarti, M. van den Berg, B. Dom, "Focused crawling: a new approach to topic-specific Web resource discovery," in 8th International WWWConference, May 1999.
[3] P.M.E. De Bra, R.D.J. Post, "Information Retrieval in the World Wide Web: Making Client-based searching feasible", Computer Networks and ISDN Systems, 27(2) 1994, 183-192.
[4] M. Hersovici, A. Heydon, M. Mitzenmacher, D.pelleg, "The Shark search Algorithm-An application: Tailored Web Site Mapping. Proc of World Wide Conference", Brisbane. Australia, 1998, 317-326.
[5] S. Ganesh, M. Jayaraj, V. Kalyan, S. Murthy and G. Aghila. "Ontologybased Web Crawler", IEEE Computer Society, Las Vegas - Nevada- USA, pp. 337-341, 2004.
[6] S. Bri, L. Page, "The anatomy of large-scale hypertext Web search engine", Proc of World-Wide Web Conference, Brisbane, Australia, 1998, 107-117.
[7] Jon M. Kleinberg, "Authoritative Sources in a Hyperlinked Environment", Journal of the ACM, 1999, 46(5), 604-632.
[8] J. Cho, H. Garcia-Molina, and L. Page, "Efficient crawling through URL ordering," in Proceedings of the Seventh World-Wide Web Conference, 1998.
[9] Y. Ye, F. Ma, Y. Lu, M. Chiu, and J. Huang, "iSurfer: A Focused Web Crawler Based on Incremental Learning from Positive Samples", APWeb, Springer,2004, pp. 122-134.