ISSN:0975-9646

Vruksha Shah et al, / (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 5 (5) , 2014, 6849-6852

Understanding Focused Crawler

Vruksha Shah, Riya Patni , Vivek Patani, Rhythm Shah

Department of Computer Engineering
K.J.Somaiya College of Engineering
Mumbai, India

Abstract— A basic web crawler can be thought of as a web robot which scans through the web and downloads the pages which can be reached by the links and thus work as an automated program. This leads to a lot of irrelevant information being generated increasing memory overhead. However, a type of crawler which aims to search only the subset of the web related to a specific topic is called a focused crawler. It is comparatively complex but extremely efficient. For predefined topic search, focused crawlers use classifiers and distillers, which help the crawler in collecting the most relevant information. This paper explains the importance of classification and distillation in crawling process.

Keywords— focused crawler, classifier, distiller, crawl, relevant context links.

1. INTRODUCTION

World Wide Web (WWW) contains a large amount of information and every second new information is added such that the size of Web is in the order of tens of billions of pages. To retrieve particular pages from the web, following strategies may be used-

a. Navigate through the web by following the links
b. Search the topic taxonomies and
c. Throw a query using search engine.

Web Crawler is the main component of search engine. It continuously downloads pages and these pages are indexed and stored in database. However, it becomes impossible for a crawler to crawl entire web and keep its index fresh. Thus what one needs is a crawler which aims to search only the subset of the web related to a specific topic. This is called a FOCUSED CRAWLER.[6]

In this paper a survey of different approaches of focused crawling has been described along with importance of classifier and distiller. The outline of this paper is as follows: section 2 describes a brief description of focused crawler and their design issues. Section 3 shows different classification techniques and comparison between these techniques . Section 4 shows how distillation helps to improve the results and in section 5, conclusion is presented.

2. BACKGROUND

2.1. Design Issues

The challenges involved in designing the Focused Crawler are as follows:

- Overloading of websites by the crawler
- Handling large amount of data at any particular time
- Web pages are dynamic in nature
- Crawler should keep a count of how frequently it should revisit a page. (Revisit policy).[2].

So there is a need of a focused crawler which effectively overcomes these design issues and also gives appreciable results.

2.2. Focused Crawler Approaches

A focused crawler can be implemented in various ways.[6] Some of the approaches are shown below:

2.2.1 Priority based focused crawler

In a priority based focused crawler, the retrieved pages are stored in a priority queue instead of a normal queue. The priority is assigned to each page based on a function which uses various factors to score a page. Thus in every iteration, a more relevant page is returned. This is mainly useful in distinguishing between important and unimportant information, wherein priority is given to a more important page.

2.2.2 Structure based focused crawler

Structure base focused crawlers take in account the web page structure when evaluating the page relevance. Its strategy is to compute the relevance score of the page with a predefined formula, then predict the relevance-score of the link, and compute the authority-score of URLs in the queue to be crawled and determine their priority according to the comprehensive value of relevance-score and authority-score namely first crawl relevant and quality page.

2.2.3 Context based focused crawler

Many a times, when a user searches a particular topic on the web, the search system is unaware of the user's needs. For e.g.: If a user is looking for a college university, the search results may include references of that university even on a news portal. Such information becomes irrelevant to the user. This increases the work for the user to filter out unwanted data. To avoid this, we implement a context based focused crawler, which tries to understand the context of the user's needs by interacting with user and comprehending the user profile. The crawler then gets adapted to such contexts and uses them for future search requests.

2.2.4 Learning based focused crawler

Learning based focused crawler is a new learning based approach to improve relevance prediction in focused web crawler. Firstly, training set is built to train the

system. Training set contains value of four relevance attributes:

a. URL word Relevancy
b. Anchor text relevancy
c. Parent page relevancy
d. Surrounding text relevancy.

Secondly ,they train the classifier (NB) using training set. After that trained classifier is used to predict the relevancy of unvisited URL.[4][21]

2.3 Relevancy Calculation Techniques

2.3.1 Weighted Page Rank

In this, weight of web page is calculated on the basis of input and outgoing links and on the basis of weight the importance of page is decided. The relevancy using this technique is less as ranking is based on the calculation of weight of the web page at the time of indexing.[22]

2.3.2 HITS

HITS stands for Hyperlink-Induced Topic Search.

It computes the hubs and authority values of the relevant pages. It gives relevant as well as important page as the result.

When comparing two pages which have received roughly the same number of citations, if one of these journals has received many citations from P1 and P2, which are regarded as important or prestigious pages, this pages needs to be ranked higher. In other words, it is better to receive citations from an important page than from an unimportant one.[22][20]

2.3.3 Eigen Rumor Algorithm

Owing to the increasing number of blogs on the web, it is a challenge to the service providers to display quality blogs to users. When page rank decides rank scores, it gives low scores to blogs and thus such scores cannot be used to decide about the importance of a blog. To overcome this problem, an algorithm was proposed by Fujimura, Inoue and Sugisaki[1] for ranking the blogs. This algorithm called Eigen Rumor Algorithm provides a rank score to every blog by weighting the scores of the hub and authority of the bloggers depending on the calculation of eigen vector.[15][22]

3. CLASSIFIERS.

The most important module of a focused crawler is the Classifier which directly affects the working efficiency of a crawler .Higher accuracy of a classifier leads to higher accurate results. Crawling can be done on a full page content basis or link context basis.

3.1 Types of Classifiers :

3.1.1 Support Vector Machine(SVM)

Support Vector Machines are used to classify data set into distinct classes. It uses a training data set to develop patterns which are represented as points in space. It then uses a mathematical algorithm to assign new examples to specific classes. In SVM, points are mapped such that separate classes are divided distinctly by a clear gap.

3.1.2 Naïve Bayes Classifier

Naïve Bayes classifier is a probabilistic classifier which uses Bayes' Theorem with an independence assumption. It assumes that the presence or absence of a certain feature is independent to that of any other feature. It also incorporates a method of maximum likelihood which estimates the parameters.[25]

3.1.3 Decision Trees based Classification

Decision tree learning is the widespread classification technique. It aims at creating a model that predicts value of a target variable. It first creates a decision tree based on training data set. Once the tree is generated, one simply has to traverse the tree to reach to the leaf and predict a yes or no(In case of Boolean classification). The limitation of decision trees is that it creates a complex model which cannot be generalized well (over fitting) and to overcome this we need to implement pruning.[26]

3.2 Comparing Classification

Pant and Srinivasan [4] compared different classification methods for focused crawling using the full-page content. Their experiment did a comparative analysis between naïve bayes classifier, support vector machines (SVM) and neural network. Naïve Bayes classifier is outperformed by SVM and Neural Network considerably. SVM is better than neural network in the sense that it gets trained faster. The authors suggest that combination of classification methods give better accuracy

Çalışkan and Ozcan[3] implemented a focused crawler using the crawler4j[19], an open source crawler implemented in Java. Their work used the Jsoup library to parse HTML documents. In the experiments, target topic was selected to be the sport news.

The authors used the Weka machine learning library to train topic classifiers. In the initial experiments, Naïve bayes, decision tree (J48 in Weka), and support vector machine are selected. The average length of a news in the dataset was around 20 words.

As the first experiment, the study evaluated naïve bayes, decision tree, and SVM classifiers on the training dataset using 5-folds cross validation. It was seen that while SVM performs the best, Naïve bayes classifier is the worst one. This can be seen from figure 1.

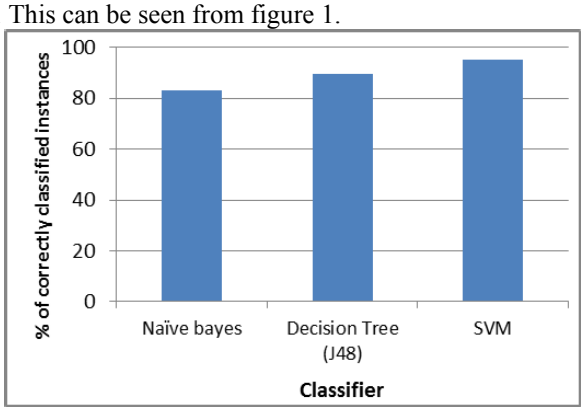

Figure 1

As the second experiment, a different test set was used which consisted of around 10,20 and 40 words link context and the aim was to see the effect of link context size on classifier performance. Figure 2 shows the result of this experiment. Results showed that Support vector machines perform the best among the three. This is shown in figure 2.

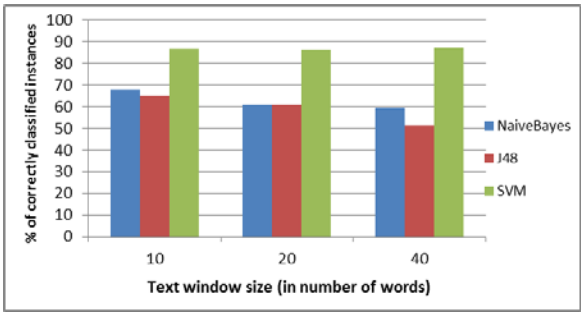

Figure 2

Their experimental results show that SVM classifier performs best compared to Naïve bayes and decision tree classifier. Text window size with 10 words is found as the optimum link context size across different classifiers

4. DISTILLER

Relevance is not the only attribute used to evaluate a page while crawling. Information relevant to the topic but having no outbound links becomes a dead end for a crawler. Here we introduce the concept of Hubs and Authority pages. A good hub page is one that points to many good authorities; a good authority page is one that is pointed to by many good hub pages.

If we talk about social networks, Prestige becomes an important attribute of nodes, especially in the context of academic papers and web documents. Prestige p(u) cannot be solely calculated on the basis of number of back-links. What we need is weighted back-links, that tells us how many important pages point to a particular page. Each node v has two corresponding scores, $h(v)$ and $a(v)$. Then the following iterations are repeated on the edge set E a suitable number of times,

$$a(v) \leftarrow \sum_{(u,v) \in E} h(u) \qquad h(u) \leftarrow \sum_{(u,v) \in E} a(v)$$

Interspersed with scaling the vectors **h** and **a** to unit length. This iteration embodies the circular definition that important hubs point to important authorities and vice versa.

Distillation is not just used as an intermediate component, but it is also an enhancement to the process. In a situation where a highly relevant page is missed out due to improper classification, a distiller comes in handy. For e.g.: a page which contains more images than text is likely to be missed by the crawler(Since crawler mostly relies on textual content). After we use a distiller, we realize that a certain page has a very high prestige p(u), and such a page may then be visited by the crawler. This leads to a more careful retrieval of information. We realize that many of such unvisited links are actually of great importance and worthy of crawling. This can be automated to go parallel with the normal crawling process, thereby saving time and efforts and enhancing the overall performance of the focused crawler. [27]

5. CONCLUSION

A focused crawler is essential for a topic based search. Various types of crawlers are implemented which caters to individual user requirements. Of these, context based focused crawler is useful but difficult to implement, whereas a priority based focused crawler is comparatively easy to implement and is reasonably efficient. Classification is an important step in the crawling process, which can be carried out using three techniques viz. Naïve bayes classification. Decision trees, and support vector machines. Among these, support vector machines prove to be the best compared to the other two with an optimum link size of 10 words. In order to further improve the performance of a crawler, a distiller is used which helps us to re-check whether the chosen page has a high prestige or not, as well to see if any important pages have been missed out, wherein we add such pages to the list of pages to be visited.

6. ACKNOWLEDGMENT

The authors gratefully acknowledge the contributions of Prof. Swati Mali and Prof. Nirmala Shinde and thank her for her endless support and motivation, the college for providing the necessary infrastructure and platform for doing this research.

REFERENCES

[1] Ko Fujimura, Takafumi Inoue and Masayuki Sugisaki,, "The EigenRumor Algorithm for Ranking Blogs", In WWW 2005 2nd Annual Workshop on the Weblogging Ecosystem, 2005.

[2] Swati Mali, B.B.Meshram, "Focused Web Crawler with Page Change Detection Policy" in 2nd International Conference and workshop on Emerging Trends in Technology (ICWET) Proceedings published by International Journal of Computer Applications® (IJCA), 2011

[3] Kamil Caliskan, Rifat Ozcan, "Comparing Classification Methods For Link Context Based Focused Crawlers" , IEEE , 2013

[4] G. PANT AND P. SRINIVASAN, "LINK CONTEXTS IN CLASSIFIER-GUIDED",TOPICAL CRAWLERS," KNOWLEDGE AND DATA ENGINEERING, IEEE,TRANSACTIONS ON , VOL.18, NO.1, PP.107-122, JAN. 2006

[5] I.Partalas, G. Paliouras, and I. Vlahavas. Reinforcement Learning with Classifier Selection for Focused Crawling. In Proceedings of the 2008 conference on ECAI 2008: 18th European Conference on Artificial Intelligence, pp. 759-760, 2008

[6] MEENU,RAKESH BATRA, "A REVIEW OF FOCUSED CRAWLER APPROACHES", IJARCSSE VOLUME 4, ISSUE 7, JULY 2014

[7] S. Chakrabarti, M. Van Den Berg, B. Dom, "Focused Crawling: A New Approach to Topic specific web resource discovery", Proc. Of 8th International WWW conference, Toronto, Canada, May,1999.

[8] Debashis Hati , Amritesh Kumar ," An Approach for Identifying URLs Based on Division Score and Link Score in Focused Crawler" ,International Journal of Computer Applications , Volume 2 – No.3, May 2010.

[9] Jaytrilok Choudhary and Devshri Roy ," A Priority Based Focused Web Crawler" , International Journal of Computer Engineering and Technology , Volume 4 ,Issue 4, july-august 2013.

[10] Sushil Kumar ,Naresh Chauhan , "A Context Model For Focused Web Search", International Journal of Computers & Technology Volume 2 No. 3, June, 2012.

[11] Wenpu Xing and Ali Ghorbani, "Weighted PageRank Algorithm", In proceedings of the 2rd Annual Conference on Communication Networks & Services Research, PP. 305-314, 2004.

[12] Jon Kleinberg, "Authoritative Sources in a Hyperlinked Environment", In Proceedings of the ACM-SIAM Symposium on Discrete Algorithms, 1998.

[13] S. Chakrabarti, B. E. Dom, S. R. Kumar, P. Raghavan, S. Rajagopalan, A. Tomkins, D. Gibson, and J. Kleinberg, "Mining the Web's Link Structure", Computer, 32(8), PP.60–67, 1999.
[14] D. Cohn and H. Chang, "Learning to Probabilistically Identify Authoritative Documents",. In Proceedings of 17th International Conference on Machine Learning, PP. 167–174.Morgan Kaufmann,San Francisco, CA, 2000
[15] Ko Fujimura, Takafumi Inoue and Masayuki Sugisaki,, "The EigenRumor Algorithm for Ranking Blogs", In WWW 2005 2nd.Annual Workshop on the Weblogging Ecosystem, 2005.
[16] E. Gatial, Z. Balogh, M. Laclavik, M. Ciglan, L.Hluchy, "FocusedWeb Crawling Mechanism based on Page Relevance." NAZOU project. Bratislava,Slovakia,2008
[17] Duygu Taylan, Mitat Poyraz, Selim Akyokuş and Murat Can Ganiz, "Intelligent Focused Crawler: Learning which Links toCrawl", IEEE 2011
[18] Sotiris Batsakis, Euripides G.M. Petrakis, Evangelos Milios, "Improving the Performance of Focused WebCrawlers", Data & Knowledge Engineering, Vol: 68, No: 10, pp: 1001-1013, October 2009
[19] HTTP://CODE.GOOGLE.COM/P/CRAWLER4J/
[20] http://en.wikipedia.org/wiki/HITS_algorithm#In_journals
[21] Diligenti, M., Coetzee, F., Lawrence, S., Giles, C., and Gori., M. "Focused Crawling Using Context Graphs.". Proc. 26th International Conference on Very Large
[22] Dilip Kumar Sharma , A.K. Sharma," A Comparative Analysis of Web Page Ranking Algorithms", (IJCSE) International Journal on Computer Science and Engineering Vol. 02, No. 08, 2010, 2670-2676
[23] http://en.wikipedia.org/wiki/Support_vector_machine
[24]https://www.princeton.edu/~achaney/tmve/wiki100k/docs/Naive_Bayes_classifier.html
[25] G. R. Dattatreya and V. V. S. Sarma,"Bayesian and decision tree approaches for pattern recognition including feature measurement costs," IEEE Trans. Pattern Anal. Mach. Intell. vol. PAMI-3, 293-298, (1981).
[26] http://www8.org/w8-papers/5a-search-query/crawling/