Available Online at www.ijcsmc.com

International Journal of Computer Science and Mobile Computing

A Monthly Journal of Computer Science and Information Technology

ISSN 2320–088X

IJCSMC, Vol. 2, Issue. 8, August 2013, pg.243 – 247

RESEARCH ARTICLE

Design of Improved Web Crawler By Analysing Irrelevant Result

Prashant Dahiwale[1], Dr. M.M. Raghuwanshi[2], Dr. Latesh Malik[3]

[1]Research Scholar, Dept. of Computer Science & Engineering, G.H.Raisoni College of Engineering, India

[2]Professor, Dept. of Comp Sc. Engg, Rajiv Gandhi College of Engineering & Research, Nagpur, India

[3]Professor, Dept. of Comp Sc. Engg, G.H.Raisoni College of Engineering, Nagpur, India

[1] *prashantdd.india@gmail.com;* [2] *m.raghuwanshi@rediffmail.com;* [3] *latesh.malik@raisoni.net*

Abstract— A key issue in designing a focused Web crawler is how to determine whether an unvisited URL is relevant to the search topic. Effective relevance prediction can help avoid downloading and visiting many irrelevant pages. In this module, we propose a new learning-based approach to improve relevance prediction in focused Web crawlers. For this study, we chose Naïve Bayesian as the base prediction model, which however can be easily switched to a different prediction model. The performance of a focused crawler depends mostly on the richness of links in the specific topic being searched, and focused crawling usually relies on a general web search engine for providing starting points.

Key Terms: - URL; focused crawler; classifier; relevance prediction; links; search engine; ranking

I. Introduction

As the number of Internet users and the number of accessible Web pages grows, it is becoming increasingly difficult for users to find documents that are relevant to their particular needs. Users must either browse through a large hierarchy of archives to find the information for which they are looking or submit a query to a publicly available search engine and wade through hundreds of results, most of them irrelevant.

Typing “Java” as keywords into Google search engine would lead to around 25 million results with quotation marks and 237 million results without quotation marks. With the same keywords, Yahoo search engine leads to around 8 million results with quotation marks and 139 million results without quotation marks, while MSN search engine leads to around 8 million results with quotation marks and around 137 million results without quotation marks. These huge numbers of results are brought to the user, but most of them are barely relevant or uninteresting to the users.

The key issue is the relevance issue of a webpage to a specific topic. Popular search engines depend on indexing databases that rely on running various web crawlers collecting information, thus main aim of a focused crawler is how to classify relevancy of a new, unvisited URL.

II. Literature Survey

In [1], they use “Stock Market” as a sample topic, and extend the learning-based relevance prediction model proposed in “Intelligent Focused Crawler: Learning Which Links to Crawl” from two relevance attributes to four relevance attributes. In “Intelligent Focused Crawler: Learning Which Links to Crawl” for an unvisited URL, only two relevance attributes, i.e., the URL words and the anchor text, are used to build a learning-based focused crawler. We extend to four relevance attributes for each URL and get able to increase the accuracy of relevance prediction for unvisited URLs. The two new relevance attributes they added are regarding the parent pages and the surrounding text of an URL. Meanwhile, they found that using only the URL words and the words of the anchor text are not enough to get an accurate prediction; so they used WordNet (a free online lexical database) to find and add new related keywords to further improve prediction accuracy. They implemented their approach in a prototype crawler and compared the performance of their crawler with two other related crawlers. The results show that their approach is valid and superior in terms of performance due to improved prediction accuracy on unvisited URLs.

The proposed crawling strategy in [2], uses the fundamental that any document is semantically closer to documents hyperlinked with it, than to documents which are not. Thus pages which are one link away are semantically closer to seed pages than pages that are two to three links away. The idea is to rank the documents based on their link distance to the topic pages. This ranking creates an ordering among the documents of the web due to its linked nature. Also the purpose of these experiments in [2] was to study the performance of the crawler on topics which are difficult to crawl, where difficulty is measured in terms of the harvest rate of the baseline crawler. Nalanda iVia focused crawler was used as the baseline crawler; it implements the focused crawler with a logistic regression based binary classifier.

A Focused Crawler which seeks, acquires, indexes, and maintains pages on a specific set of topics that represent a relatively narrow segment of the web [3]. It entails a very small investment in hardware and network resources and yet achieves respectable coverage at a rapid rate, simply because there is relatively little to do. Thus, web content can be managed by a distributed team of focused crawlers, each specializing in one or a few topics. Each focused crawler will be far more nimble in detecting changes to pages within its focus than a crawler that is crawling the entire web. The focused crawler is guided by a classifier which learns to recognize relevance from examples embedded in topic taxonomy, and a distiller which identifies topical vantage points on the web.

The focused crawler aims at providing a simpler alternative for overcoming the issue that immediate pages which are lowly ranked related to the topic at hand. The idea is to recursively execute an exhaustive search up to a given depth d, starting from the ``relatives" of a highly ranked page[4]. Hence, a set of candidate pages is obtained by retrieving pages reachable within a given perimeter from a set of initial seeds. From the set of candidate pages, we look for the page which has the best score with respect to the topic at hand. This page and its ``relatives" are inserted into the set of pages from which to proceed the crawling process. There assumption is that an ``ancestor" with a good reference is likely to have other useful references in its descendants further down the lineage even if immediate scores of web pages closer to the ancestor are low. They define a degree of relatedness τ with respect to the page with the best score. If τ is large, we will include more distant ``cousins" into the set of seeds which are further and further away from the highest scored page.

The architecture of the crawler [5]: Initially a single page is considered as primary seed. An ID is assigned to it and together with its url is stored in the database (i.e. in a table named ’Seed Pages’). On addition of a page to seed pages, following tasks are done (Seed is empty at start, and an initial seed page would be added to it at beginning. Then in each step one of the fetched pages will be added to the seed. Method for selecting such a fetch page comes in following). Each seed page has some links to the other pages. First, such links are downloaded and stored in a special folder, i.e. download folder. Then address of the page and some of its attributes is stored in the database. Attributes include similarity degree of the page to the domain (Sports), number of links from the page to seed pages and number of links from seed pages to it. Thus the paper[5], we have introduced a simple framework for focused crawling using combination of two existing methods, the Link Structure analysis and Content Similarity. Our generic framework is more powerful and flexible than previously known focused crawlers.

The paper [6] discusses intelligent crawling with help of ontology. Ontology is one of the increasingly popular ways to structure information. Ontologies are also called graphs of concepts. Ontology help people and computers to access the information they need, and effectively communicate with each other. They therefore have a crucial role to play in enabling content based access, interoperability, and communication across the Web, providing it with a qualitatively new level of service: the Semantic Web. Evaluation of association metric with the aid of the ontology that may be generic or domain dependent will surely give more relevant Results The knowledge path is devised with the help of ontology in order to map the seed url to most relevant url. The knowledge path is helpful to direct the search in correct direction.

In the paper [7] they address one of these important challenges: How should a crawler select URLs to scan from its queue of known URLs? If a crawler intends to perform a single scan of the entire Web, and the load placed on target sites is not an issue, then any URL order will suffice. That is, eventually every single known URL will be visited, so the order is not critical. However, most crawlers will not be able to visit every possible page for two main reasons: Their client may have limited storage capacity, and may be unable to index or analyze all pages. Currently the Web contains about 1.5TB and is growing rapidly, so it is reasonable to expect that most clients will not want or will not be able to cope with all that data.

Crawling takes time, so at some point the crawler may need to start revisiting previously scanned pages, to check for changes. This means that it may never get to some pages. It is currently estimated that over 600GB of the Web changes every month. In either case, it is important for the crawler to visit “important” pages first, so that the fraction of the Web that is visited (and kept up to date) is more meaningful. In the following sections, they present several different useful definitions of importance, and develop crawling priorities so that important pages have a higher probability of being visited first.

Mining the Web [8]: Discovering Knowledge from Hypertext Data is the first book devoted entirely to techniques for producing knowledge from the vast body of unstructured Web data. Building on an initial survey of infrastructural issues—including Web crawling and indexing—Chakrabarti examines low-level machine learning techniques as they relate specifically to the challenges of Web mining. He then devotes the final part of the book to applications that unite infrastructure and analysis to bring machine learning to bear on systematically acquired and stored data. Here the focus is on results: the strengths and weaknesses of these applications, along with their potential as foundations for further progress. From Chakrabarti's work—painstaking, critical, and forward-looking—readers will gain the theoretical and practical understanding they need to contribute to the Web mining effort.

The focused crawler in [8] is named as Context Focused Crawler (CFC), it uses the limited capability of search engines like AltaVista or Google to allow users to query for pages linking to a specified document. This data can be used to construct a representation of pages that occur within a certain link distance of the target documents. This representation is used to train a set of classifiers, which are optimized to detect and assign documents to different categories based on the expected link distance from the document to the target document. During the crawling stage the classifiers are used to predict how many steps away from a target document the current retrieved document is likely to be. This information is then used to optimize the search.

III. Proposed Work

A family of potential solutions for designing crawler heuristics comes from the area of classifiers. These machine learning tools, built using training sets of examples, have been studied extensively for a variety of purposes including the closely related task of text classification. In the context of topical crawler logic, classifiers offer a natural approach to decide if a given hyperlink is likely or is not likely to lead to a relevant Web page. It may be noted that the task in classifier-guided topical crawling is one of classifying URLs before downloading the pages corresponding to them. The highly cited 1999 paper by Chakrabarti et al. [1999] explores Naive Bayes classifiers in what the authors call a focused crawling framework. Similarly Diligenti et al. [2000] and Johnson et al. [2003] use Naïve Bayes and Support Vector Machine (SVM) classifiers respectively to guide their topical crawlers. Most of these explore a single classification scheme, which typically is Naive Bayes. Additionally, almost all of these investigations are limited to crawls for not more than 10 topics. These studies, although significant in that they draw our attention to the potential of classifiers for crawling, lead us to conclude that a systematic study exploring alternate classification schemes under rigorous experimental conditions is now timely. Our goal is thus to contribute such a systematic study, with experiments exploring multiple versions of the Naive Bayesian and Support Vector Machine (SVM) classification schemes for crawling.

IV. Proposed System

$$P(X|\text{Relevant=Yes})*P(\text{Relevant=Yes}) > P(X|\text{Relevant=No})*P(\text{Relevant=No})$$

Where,

$$P(\text{Relevant=Yes})= \#\text{ of relevant/total}$$

$$\begin{aligned}P(X|\text{Relevant=Yes})=\,&P(\text{URLwordRelevancy}|\text{Relevancy=Yes})*\\&P(\text{AnchorTextRelevancy}|\text{Relevancy=Yes})*\\&P(\text{ParentPage relevancy }|\text{ Relevancy=Yes})*\\&P(\text{SurroundingTextRelevancy}|\text{Relevancy=Yes})\end{aligned}$$

P(Relevant=No)= # of Irrelevant/total
P(X|Relevant=No)=P(URL word relevancy | Relevancy=No)*
P(Anchor Text relevancy | Relevancy=No)*
P(Parent Page relevancy | Relevancy=No)**
P(Surrounding text relevancy | Relevancy=No)

V. PROPOSED ARCHITECTURE

Figure below details high-level layered design for a focused crawling infrastructure. In a crawling loop a URL is picked from the frontier (a list of unvisited URLs), the page corresponding to the URL is fetched from the Web, the fetched page is processed through all the layers (see Figure), and the unvisited URLs from the page are added to the frontier. The networking layer is primarily responsible for fetching and storing a page and making the process transparent to the layers above it. The parsing and extraction layer parses the page and extracts the needed data such as hyperlink URLs and their contexts (words, phrases etc.). The extracted data is then represented in a formal notation (say a vector) by the representation layer before being passed onto the intelligence layer that associates priorities or scores or result with unvisited URLs in the page.

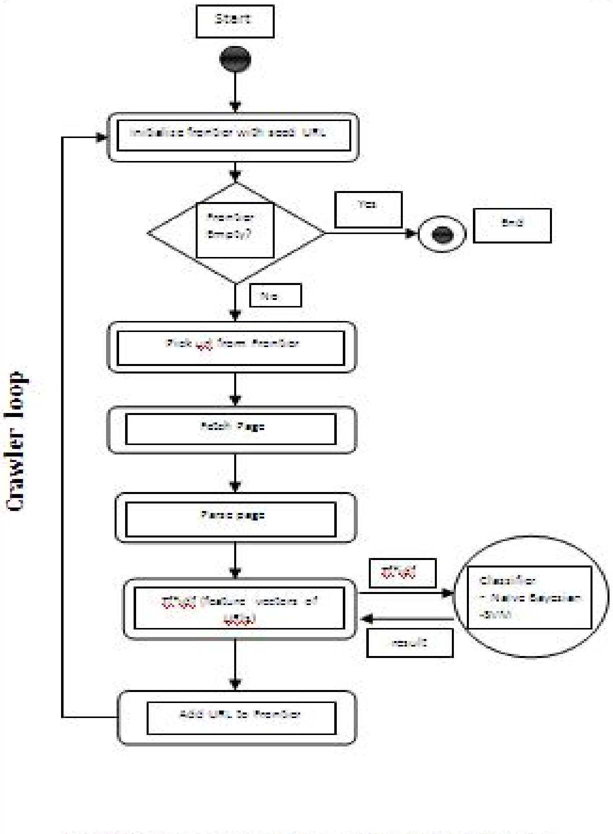

Fig.Crawler model: details of a crawler and its interaction with the guiding classifier.

REFERENCES

[1] Mejdl S. Safran, Abdullah Althagafi and Dunren Che,” Improving Relevance Prediction for Focused Web Crawlers”.In Computer and Information Science (ICIS), 2012 IEEE/ACIS 11th International Conference, Shanghai,2012.

[2] Rashmin Babaria, J. Saketha Nath, Krishnan S, Sivaramakrishnan K R, Chiranjib Bhattacharyya, M.N.Murty,Department of Computer Science and Automation, Indian Institute of Science, Bangalore,” Focused Crawling with Scalable Ordinal Regression Solvers”. In Proceedings of 24th International Conference on Machine Learning,Corvallis,OR,2007.

[3] Soumen Chakrabarti, Martin van den Berg, Byron Dom," Focused Crawling: A New Approach to Topic-Specific Web Resource Discovery". In Proceedings of the 8th International World Wide Web Conference,1999.

[4] Ah Chung Tsoi, Daniele Forsali ,Marco Gori ,Markus Hagenbuchner, Franco Scarselli,"A Simple Focused Web Crawler". In Proceedings of WWW 2003,Budapest,Hugary,2003.

[5] Mohsen Jamali, Hassan Sayyadi, Babak Bagheri Hariri and Hassan Abolhassani," A Method for Focused Crawling Using Combination of Link Structure and Content Similarity". In Proceedings of Web Intelligence IEEE/WIC/ACM International Conference,Hong Kong,2006.

[6] M.M.Raghuwanshi ,Prashant Dahiwale and Anil Mokhade," Intelligent Web Crawler". In Proceedings of the International Conference and Workshop on Emerging Trends in Technology,New York,2010.

[7] Junghoo Cho, Hector Garcia-Molina, and Lawrence Page,"Efficient crawling through URL ordering", In Proceedings of the Seventh International World Wide Web Conference, pages 161--172, April 1998.

[8] Chakrabarti and Ramakrishanan,"Mining the Web".

[9] M. Diligenti, F. M. Coetzee, S. Lawrence, C. L. Giles and M. Gori,"Focused Crawling Using Context Graphs". In Proceedings of the 26th VLDB Conference,Cairo, Egypt, 2000.