Advances in Computer Science and Information Technology (ACSIT)
p-ISSN: 2393-9907; e-ISSN: 2393-9915; Volume 2, Issue 13; July-September, 2015 pp. 76-81

http://www.krishisanskriti.org/Publication.html

Focused Crawling- An Approach for Classification of Web URLs using DTI and Neural Networks

Promila Devi[1] and Ravinder Thakur[2]

[1]Department of Computer Science & Engg. LRIET Solan, HPTU Hamirpur, India
[2]Department of Computer Science & Engg. LRIET Solan, HPTU Hamirpur, India

Abstract—*The size of the www (World Wide Web) is increasing every day and it becomes a challenging task for a search engine to find the relevant information on the web. The focused crawlers are becoming popular in place of search engines as they seek out the pages which are relevant to a search topic and neglecting irrelevant pages. Web crawlers are the software which traverses the internet for retrieving web pages which are of interest to the user. An issue with focused web crawler is to find the maximal set of relevant web pages and to improve the relevancy prediction. In our proposed approach, we will find attributes for seed URLs like parent page, surrounding text etc. and then classify the URLs according to their score and full training dataset is prepared accordingly. Relevancy prediction of unseen URLs is done by data mining classifiers models i.e. Decision tree induction and neural networks and precision rate is calculated and comparative study shows that models we have used gives better performance as compared to the previously used models.*

Keywords*: - Focused Crawler, Web Crawler, Search engine, weight table, Relevancy calculation.*

1. INTRODUCTION

A web search engine is designed to search for information on the World Wide Web. Unlike web directories, which are maintained only by human editors, search engines also maintain real-time information by running an algorithm on a web crawler. Search engine optimization (SEO) is the process which affects the visibility of a website or a web page in a search engine's search results. SEO may target different kinds of search, including image search, text search, audio/video search, academic search, news search and industry-specific vertical search engines. It considers how search engines work, what people search for, the actual search terms or keywords typed into search engines and which search engines are preferred by their targeted audience [7].

A Web crawler systematically browses the World Wide Web basically for the purpose of Web indexing. Web search engines and some other sites use Web crawling to update their web content or indexes of others sites web content. Web crawlers can copy all the pages they visit for later processing by a search engine that indexes the downloaded pages so that users can search them much more quickly[7].

The Focused crawling technique enables a search engine to focus only on specific topic. The crawler starts with several seed pages, which are topic relevant and whenever a web page is fetched, the unvisited URLs are extracted and then they are scored by relevancy like. The crawler then picks up the URL with highest score. Domain specific knowledge is used to rank the importance of a web page and to guide the crawler's search through hyperlinks [4].

With the vast amount of information available on the web and about millions of pages available on the web, it becomes challenging task to search all the pages on the web by Search engines. Whenever a user type a word on search engines like Google, Yahoo etc., it gives millions of results to the user. But most of the results are of no use to them. Therefore, focused crawlers were introduced to select URLs which are relevant to a topic. But here issue is how to predict whether the content of the URLs are relevant or not. The aim of a focused crawler is to seek out a subset of the Web to only gather documents on a specified topic which are relevant and quality documents and avoid irrelevant documents [11]. So, in this paper we are applying data mining techniques to improve relevancy prediction & precision rate.

2. RELATED WORK

It was Chakrabati who has firstly introduced Focused crawling in 1999[3]. One of the first web crawlers was proposed by Choj et. al.[4] and they introduced a best first strategy. The Fish-Search Algorithm is one of the algorithms used for crawling of pages with keywords specified in the query and it was proposed by P. Debra et. al. [10]. In Fish –Search, the system is query driven. Starting from a set of seed pages, it considers only those pages that have content matching a given query & their neighbours. Shark- Search is a modification of Fish-Search algorithm proposed by M. Hersovici et. al. [9]. In this algorithm, vector space model (VSM) is used & priority

values are computed based on the priority values of parent pages, page content & anchor text. Many researchers have given different approaches based on link analysis. The Link-Structure based method is analysing the reference information among the pages to evaluate the page value. Effective focused crawling based on content and link analysis has been proposed for link analysis based on URL score, anchor score and relevance score, HAWK a Focused Crawler with content and link analysis [2]. Tao Meng et. al. [12] has presented a model for incremental crawling and all link-attributes of web pages in a website as well as their evolution, and reveals the correlation between their changing frequencies and link-attributes. Crawlers often start from some vertices of the web graph, and reach others following the edges. This work is based on Tianwang System which has been gathering and indexing web pages from the Chinese web for seven years. We used its crawlers to gather pages in our research and also used our conclusions to enhance their performance. In the paper author can find a way to skip the scanning. First, since pages with high link-attribute values change more often than the others, and these pages are important, we can only scan the pages with big Page Rank values. This research can be used to enhance the performance of an incremental crawler greatly.

According to D. Saraswathi et. al. [7], the Web is both an excellent medium for sharing information as well as attractive platform for delivering products and services. This platform is, to some extent, mediated by search engines in order to meet the needs of users seeking information. Search engine has enormous amount of information and return a customary answer for given query with only small set of results. Most of the web sites are interested to display the web pages within ten search results. In this paper author has designed a spam detector the spam detector detects whether the links are spam or not spam, if it is spam remove from the search engine results, otherwise display the results to the user. Due to the similarities between spam and non-spam the existing link spam identifiers are not an effective method to classify the web links. Since spam and non-spam links are so similar, it is sometimes very difficult for a human to differentiate between the two. The ideas and the concepts identified this research would benefit the users who search information in the search engine. Since the users will get quality web links and no need to spend much amount of time to search the information in the web. Bireshwar Ganguley et. al.[1] has designed a crawler, which is a system that learns the specialisation from examples, and then explores the web, guided by a relevance and popularity rating mechanism. It filters at the data-acquisition level, rather than as a post-processing step. The author has also discussed about Internet, Search Engines, Web Crawlers, Focused Crawlers and Block Partitioning of Web pages. In this paper, approach was to partition the web pages into content blocks. Using this approach we can partition the pages on the basis of headings and preserve the relevant content blocks. The information can be used to collect more on related data by intelligently and efficiently choosing what links to follow and what pages to discard. This process is called Focused Crawling.

3. PROPOSED ARCHITECTURE

Fig. 1 shows the architecture of proposed system.

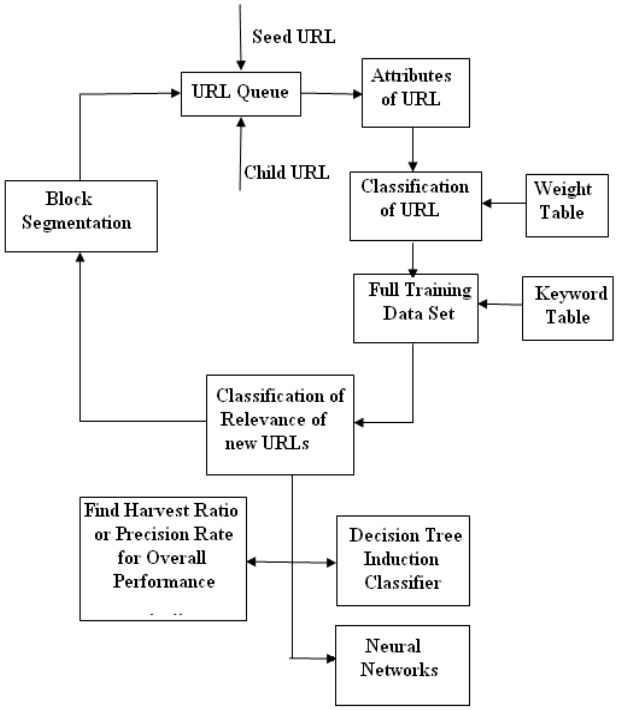

Fig. 1: Architecture of Proposed System

URL Queue stores the list of unvisited URLs and is initialized by seed URLs. The attributes for various URLs will have been calculated to prepare the full training dataset by using keyword and weight table. URLs are classified according to relevancy calculation and then Decision Tree Induction classifier and Neural Network models are used to calculate the overall performance and to compute precision rate.

4. PROPOSED APPROACH

4.1 Problem Formulation

Search engines today form the backbone of World Wide Web. Any information that a user has to access is searched on the internet using popular search engines like Google, Bing, and Yahoo! etc. These general purpose search engines are characterized by low to medium level performance due to the use of generic web-crawlers. As a result crawlers were designed that would learn the specialisation from examples, and then explores the web, guided by a relevance and popularity rating mechanism. The web pages are partitioned into content blocks on the basis of headings. The information can be used to collect more on related data by intelligently and efficiently choosing what links to follow and what pages to

discard. This process is called Focused Crawling. In our proposed work we propose a method for this efficient block segmentation as is required for the above described web crawler. Our work is similar to as proposed in [5][9].

5. OBJECTIVES

- To understand the working of content Block segmentation form previous research.
- To find attributes for seed URLs and their child URLs.
- To classify the URLs according to their weight or score in the weight table.
- To prepare the full training data set and maintain the keyword table.
- To classify the relevancy of the new unseen URLs using Decision tree induction classifier and Neural Networks**.**
- To find out the Harvest ratio or Precision rate for overall performance evaluation.

4.2 Seed URL Extraction

URL queue contains a list of unvisited URLs. The topic name is sent as query to different search engines like Google, Yahoo and MSN, then first K results are retrieved which are common to all the search engines, as they can be assumed as the most relevant for this query and the URLs which are common among two search engines are considered as relevant. The common URLs among various search engines are considered as seed URLs. For seed URL extraction we have taken four search topics and these are "Stock Market", "Computer books", "Iron ore" and "Genetic DNA" and corresponding seed URLs have been extracted. Based on results, the seed URLs for the topic "Iron Ore" is shown in the Table 1 below and following the same procedure, seed URLs for other keywords are also extracted and also listed in tabular form.

Table 1: Seed URLs Table for Keyword "Iron

Seed URLs
http://www.technologystudent.com/joints/iron2.htm
http://en.wikipedia.org/wiki/iron_ore
http://www.metalprices.com/metal/iron-ore
http://miningartifacts.homestead.com/IronOres.html
http://facebook.com

4.3 Topic Specific Weight Table Construction

Seed URLs
http://www.technologystudent.com/joints/iron2.htm
http://en.wikipedia.org/wiki/iron_ore
http://www.metalprices.com/metal/iron-ore
http://miningartifacts.homestead.com/IronOres.html
http://facebook.com

When a search topic is sent as query and corresponding URL pages are fetched and parsed by eliminating stop words like is, are etc, the keyword table is constructed and weight to that keyword is assigned in the table as in .The weight of a keyword is computed as w= tf * df, where tf is the term frequency and df is the document frequency of the word. Then the words with the highest weight are ordered in the keyword table and then the weights are normalized as

$$W = Wi / Wmax \quad (1)$$

where Wi is the weight of keyword i and Wmax is the weight of keyword with highest weight[9]. By following this procedure the keyword weight table construction for various search topics is shown in tables 2, 3, 4 and 5 respectively.

Table 2: Keyword Weight table for Genetic DNA

S. No.	Keyword	Weight
1	DNA	1
2	Genetic	0.9167
3	Family	0.7083
4	Ancestor	0.6667
5	History	0.3333
6	Micro	0.1042
7	Bio	0.4167

Table 3 Keyword Weight table for iron ore

S. No.	Keyword	Weight
1	Iron	1
2	Ore	0.86
3	Mine	0.51
4	Mining	0.52
5	Metal	0.31
6	Hema	0.21
7	Magne	0.12

Table 4: Keyword weight table for stock market

S. No.	Keyword	Weight
1	Stock	1
2	Market	0.78
3	Quote	0.42
4	Finance	0.35
5	Trade	0.26
6	Invest	0.24
7	Exchange	0.13

Table 5: Keyword weight table for Computer books

S. No.	Keyword	Weight
1	Book	1
2	Free	0.89
3	program	0.45
4	Computer	0.38
5	Web	0.251
6	Ebook	0.256
7	Site	0.262

4.4 Relevancy calculation

There are different types of attributes for measuring that a particular link is relevant or not. The relevancy calculation is same as done in [9]

1. Parent page relevancy

Parent pages are the pages where the links to the seed URL are available. The contents of parent pages of each seed URL are extracted and then the top keywords are taken and weight is assigned to the keywords to construct parent page keyword table. These keywords are matched with those in the keyword weight table and if keyword is available in both the table then, parent page weight will be assigned to it otherwise zero will be assigned.

2. URL words Relevancy

URL words are the words given in the seed URL and we have to assign weight to each URL word. The relevancy of URL words can be calculated by firstly copying all the topic keywords in the table and then if the keyword is also the URL word then the weight as that given in the keyword table is assigned to it, otherwise zero is assigned.

3. Anchor Text Relevancy

Anchor text is the clickable text on web pages. It can be hypertext or hyperlink that navigates us to other pages. From the link it is not possible to know anchor text, so we need to find parent pages that hold the references. To calculate the relevancy of anchor text, the anchor texts of each seed URL are extracted and top keywords are fetched to construct anchor text keyword table and then match them with keyword weight table. If exist in both tables, then corresponding weight is assigned to the keyword and if not matched assign the keyword zero weight.

4. Surrounding Text Relevancy

Surrounding text is the text that surrounds the seed URL i.e. before or after it in the parent pages. To compute the relevancy of surrounding text, some non-stop words are taken from parent pages of each seed URL and top keywords are extracted to construct surrounding text keyword table and then these keywords are matched with our earlier made weight table and if the keyword exist is both the tables then weight is assigned to it as that of surrounding text keyword table otherwise zero weight is assigned.

The relevancy calculation for the keyword "Iron Ore" is shown in the Fig. 2 below. In the same way attributes for various search word is calculated.

Seed URL	URL words relevancy	Anchor text relevancy	Parent page relevancy	Surrounding text relevancy
http://www.technologystudent.com/joints/iron2.htm	0.24	0.35	0.15	0.42
http://en.wikipedia.org/wiki/iron_ore	0.26	0.45	0.16	0.51
http://www.metalprices.com/metal/iron-ore	0.28	0.21	0.38	0.58
http://miningartifacts.homestead.com/IronOres.htm	0.40	0.31	0.71	0.51
http://www.facebook.com	0.0	0.0	0.08	0.16

Fig. 2 Transition data set for "Iron Ore"

6. EXPERIMENTAL RESULTS

The experiments are conducted on search topics i.e. Genetic DNA, Iron Ore, Stock Market and Computer Books by using fixed training dataset. We also implement Naive Bayesian (NB) classifier for the purpose of comparison along with decision tree induction (DTI) classifier and Neural Networks. The experiment compares the relevance of visited URLs to the search topic among three crawlers. Precision metric is used to evaluate crawler performance.

$$\text{Precision rate} = \frac{\text{Number of relevant pages}}{\text{Number of downloaded pages}} \quad (2)$$

Table 6 shows the precision rates of various search topics for different classifiers i.e. NB, DTI and NN.

Table 6: Precision Rate of various Search topics

Search Topic	NB Classifier	DTI Classifier	Neural Networks
Genetic DNA	0.23325	0.30388	0.50125
Iron Ore	0.26513	0.44763	0.5075
Stock Market	0.24013	0.37262	0.46375
Computer Books	0.22513	0.53875	0.39763

We have illustrated the results of different classifiers using two-dimensional graph.

Topic: Genetic Data

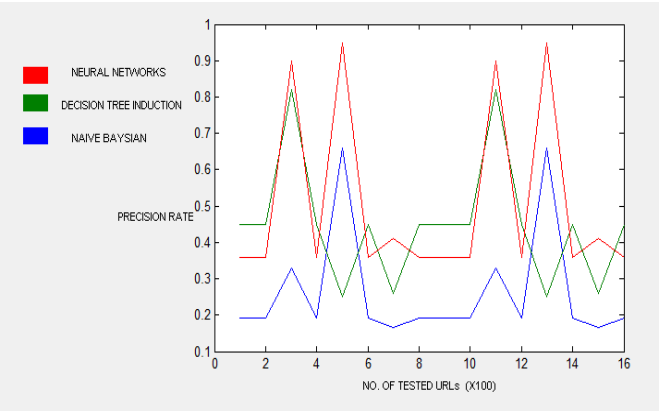

Fig. 3: Comparison Graph (NB, DTI, NN) for Genetic Data

Topic: Iron Ore

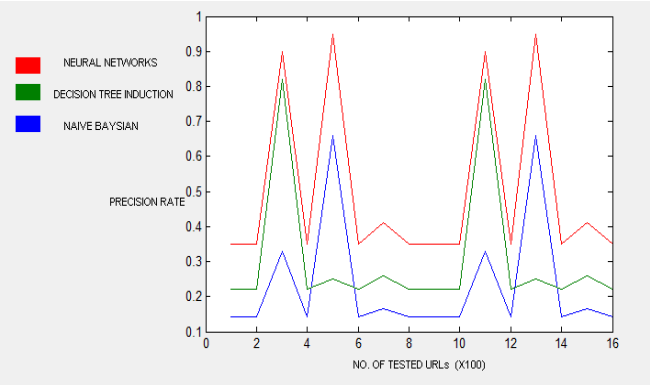

Fig. 4: Comparison Graph (NB, DTI, NN) for Iron Ore

The Figs. 3, 4, 5 and 6 shows comparison graph based on three models viz. Naïve Baysian (NB), Decision Tree Induction (DTI) and Neural Networks (NN) for search topics Genetic DNA, Iron Ore, Stock Market, and Computer Books respectively. Graph shows that Neural Networks and DTI has higher precision rate as compared to Neural Network as the number of tested URLs increased.

Topic: Stock Market

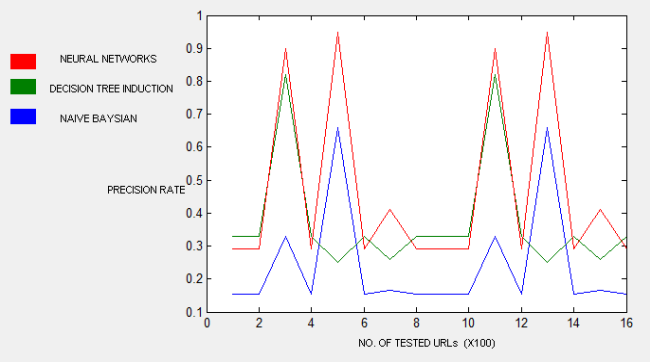

Fig. 5. Comparison Graph (NB, DTI, NN for Stock Market

Topic: Computer Books

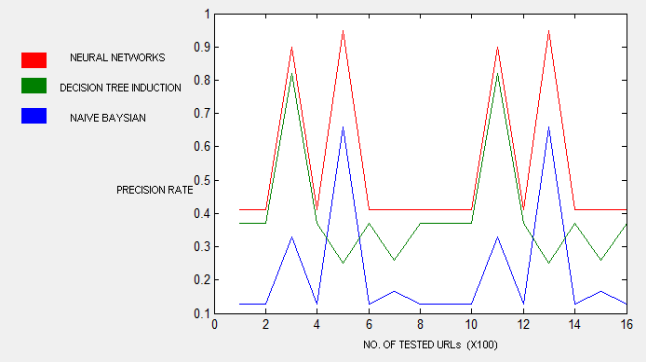

Fig. 6: Comparison Graph (NB, DTI, NN) for Computer Books

7. CONCLUSION AND FUTURE WORK

In this paper, we elaborate the focused Web Crawler and various method on which crawling is based. This paper also gives brief summary of Web Crawling Schemes present in the literature. A focused crawler that makes use of the classifier based crawler of the web similar to is presented in the paper. An ordinal regression formulation is proposed to rank the URLs that need to be crawled by the crawler. A Neural Network based ordinal regression formulation is proposed to handle large seed sets. The web crawlers can be efficiently implemented to index the search results and to make them more relevant to the search request made by the user. Block segmentation as proposed can enable the web pages to be sorted in a more relevant pattern and enable the web crawler to search for relevant topics only from a web page and display it to the search results. Above all the segregation would make the results to be displayed faster and more efficiently. In this paper, we review the learning-based focused crawling approach that uses four relevance attributes to predict the relevance of unvisited URLs. The four attributes are the URL words, its anchor text, the parent pages, and the surrounding text. We conducted the experiment, and confirmed that the algorithm functions correctly. During the last few years, the research activities of web crawling have attracted tremendous attention with a large number of academic publications, even with the lack of large scale and long term applications. We hope this article would further motivate the research interest in Web Crawling.

We are working further on the implementation to improve the algorithm. Our further investigations include experiments with large web pages and multimedia data. There are different ways to manipulate parameters which influence the performance of the algorithm. Our approach adopted DTI and NN classification model, which can be extended to other more sophisticated models. As future work, more extensive tests can be done with larger volumes of web pages, to incorporate and test with other classification and clustering models.

REFERENCES

[1] Bireshwar Ganguly, Devashri Raich, “Performance Optimization of Focused Web Crawling Using Content Block Segmentation”,ICESC '14 Proceedings of the2014 International Conference on Electronic systems, Signal Processing and Computing Technologies, Pages 365-370.

[2] Chain,X and Zhang, X. 2008, HAWK: A Focused Crawler with Content and Link Analysis. IEEE International Conference on e-Business Engineering.

[3] Chakrabarti S., Van den berg M, Dom B, “Focused Crawling: A new approach to Topic Specific Web Resource discovery”, Proceedings WWW 1999.

[4] Choj, Garcia Molina H, Page L., “Efficient Crawling Through URL ordering ”, Proceedings WWW 1998.

[5] Debashis Hati, Amritesh Kumar, Lizashree Mishra,“Unvisited URL Relevancy Calculation in Focused Crawling Based on

Naïve Bayesian Classification", International Journal of Computer Applications (0975-8887) Volume 3-No.9, July 2010.

[6] D. Saraswathi, A.Vijaya, "A Generic Tool for Link Spam Detection in Search Engine Results using Graph Mining", Proceedings of the 2013 International Conference on Pattern recognition, Informatics, and Mobile Engineering (PRIME), February 21-22.

[7] https://en.m.wikipedia.org.

[8] Mejdl S. Safran, Abdullah Althagafi and Dunren Che, "Improving Relevance Prediction for Focused Web Crawlers ", IEEE/ACIS 11th International Conference on Computer and information Science.

[9] M. Hersovici, A. Heydon, M. Mitzenmacher, D. Peeleg, "The Shark-Search Algorithm: An application Tailored website mapping", Proceedings of WWW conference, Brisbane, Australia, 1998, 317-326.

[10] P.M.E. De Bra, R.D.J. Post, "Information Retrieval in the World Wide Web making client based searching feasible ", Computer Networks & ISDN systems, 27(2) 183-192.

[11] Sun, Y.,Jin,P., and Yue, L., 2008. "A framework of a Hybrid Focused Web crawler", 2nd international conference on Future Generation Communication & Networking Symposia.

[12] Tao Meng, Hongfai Yan, Jimin Wang, Xiaoming Li,"The Evolution of Link-Attributes for Pages and Its Implications on Web Crawling", Proceedings of the 2004 IEEE/WIC/ACM International conference on Web