Unvisited URL Relevancy Calculation in Focused Crawling Based on Naïve Bayesian Classification

Debashis Hati
Assistant Professor, School of Computer Engineering
KIIT University
Bhubaneswar, India

Amritesh Kumar
Research Associate, School of Computer Engineering
KIIT University
Bhubaneswar, India

Lizashree Mishra
Research Associate, School of Computer Engineering
KIIT University
Bhubaneswar, India

ABSTRACT
Vertical search engines use focused crawler as their key component and develop some specific algorithms to select web pages relevant to some pre-defined set of topics. Crawlers are software which can traverse the internet and retrieve web pages by hyperlinks. The focused crawler of a special-purpose search engine aims to selectively seek out pages that are relevant to a pre-defined set of topics, rather than to exploit all regions of the Web. Maintaining currency of search engine indices by exhaustive crawling is rapidly becoming impossible due to the increasing size of the web. Focused crawler aims to search only the subset of the web related to a specific topic, and offer a potential solution to the problem. A focused crawler is an agent that targets a particular topic and visits and gathers only a relevant, narrow web segment while trying not to waste resources on irrelevant material. As the crawler is only a computer program, it cannot determine how relevant a web page is. The major problem is how to retrieve the maximal set of relevant and quality page. In our proposed approach, we classify the unvisited URL based on visited URLs attribute score, i.e., unvisited URLs are relevant to topics or not, and then decide based on seed page attribute score. Based on score, we put "Yes" or "No" values in the table. URLs attributes are: it's Anchor text relevancy, its description in Google search engine and calculates the similarity score of description with topic keywords, cohesive text similarity with topic keywords and Relevancy score of its parent pages. Relevancy score is calculated based on vector space model. Classification is done by Naïve Bayesian classification methods.

General Terms
Crawling technology, Focused crawling

Keywords
Crawler; Focused crawler; Vector space model; Naïve Bayesian classification methods

1. INTRODUCTION
Recent developments on the computer and networking technologies have made the Internet to be the most popular and the largest information source over the world. It was found that about a decade ago, the Web contained more than 350 million pages such that 600 Gigabytes of information on these pages were updated every month and the size of the Web was doubled every year. Due to the growth and flux of the information on the Web, it may not possible for a general purpose crawler and search engine to index and search all the pages on the Web. To overcome this problem, focused crawling of the Web was proposed. The aim of a focused crawler is to traverse a subset of the Web to only gather documents on a specific topic and to identify the promising links that lead to on-topic documents, and avoid off-topic branches [6].

A Web Crawler searches through all the Web Servers to find information about a particular topic. However, searching all the Web Servers and the pages, are not realistic, given the growth of the Web and their refresh rates. Crawling the Web quickly and entirely is an expensive, unrealistic goal because of the required hardware and network resources [8]. Focused Crawling is designed to traverse a subset of the Web to gather documents on a specific topic. It also aims to identify the promising links that lead to target documents, and avoid off-topic searches. In the large area of websites, traditional web crawlers cannot function well to solve this problem. The focused crawler of a special-purpose search engine aims to selectively seek out web pages that are relevant to a pre-defined set of topics, rather than to exploit all regions of the Web. Focused crawlers aim to search only the subset of the web related to a specific topic, and offer a potential solution to the problem. The general-purpose search engines, such as Google, have provided us with a lot of facilities, and become very popular. However, they have some disadvantages because a general-purpose search engine aims to cover the network as enough as possible. So, it usually returns many web pages users are not interested in. Therefore, it is extremely important for a search engine how to effectively build up a semantic pattern for specific topics [2, 7]. The traditional process of focused web crawler is to harvest a collection of web documents that are focused on the topical subspaces. They traverse the web collecting only relevant data to a predefined topic while neglecting on the same time off-topic pages. The crawler is kept focused through a crawling strategy which determines the relevancy degree of the web page to the predefined topic and depending on this degree a decision is made whether to download the web page or not. In our proposed approach, we classify the unvisited URL based on visited URLs attribute score, i.e., unvisited URLs are relevant to topics or not, and then decide based on seed page attribute score [5, 9].

The outline of the paper is as follows. We discuss the existing works of focused crawling in section 2. In section 3, we introduce the architecture of our proposed approach. In section 4, we present the key algorithm of our proposed approach. In section 5, we have presented our experimental analysis and in section 6, we have concluded our research paper.

2. PRIOR WORK

Maintaining currency of search engine indices by exhaustive crawling is rapidly becoming impossible due to the increasing size and dynamic content of the web. Focused crawlers aim to search only the subset of the web related to a specific category, and offer a potential solution to the currency problem. The major problem in focused crawling is performing appropriate credit assignment to different documents along a crawl path, such that short-term gains are not pursued at the expense of less-obvious crawl paths that ultimately yield larger sets of valuable pages [3].

A focused crawler is a program used for searching information related to some interested topics from the Internet. The main property of focused crawling is that the crawler does not need to collect all web pages, but selects and retrieves relevant pages only. As the crawler is only a computer program, it cannot determine how relevant a web page is. In order to find pages of a particular type or on a particular topic, focused crawlers aim to identify links that are likely to lead to target documents, and avoid links to off topic branches. However, the concept of prioritizing unvisited URLs on the crawl frontier for specific searching goals is not new, and Fish-Search and Shark-Search were some of the earliest algorithms for crawling for pages with keywords specified in the query. 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 (expressed as a keyword query or a regular expression) and their neighborhoods (pages pointed to by these matched pages).

Shark-Search is a modification of Fish-search which differs in two ways: a child inherits a discounted value of the score of its parent, and this score is combined with a value based on the anchor text that occurs around the link in the Web page. Many researchers have written their approaches based on link analysis. For example, Effective Focused Crawling based on content and link structure analysis has been proposed for link analysis based on URL score, anchor score and relevance score and HAWK: A Focused Crawler with Content and Link Analysis [1]. Some have written their approaches based on page rank value. For example, An Application of Improved Page Rank in Focused Crawler based on To-page rank value and an Improvement of Page Rank for Focused Crawler based on T page rank. Some have written based on ontology. For example, A Survey in Semantic Web Technologies-Inspired Focused Crawlers and A Transport Service Ontology-based Focused Crawler based on ontology. Some have developed based on meta search and content block partition "A Framework of a Hybrid Focused Web Crawler."

Some have developed rule based focused crawler. For example, Design of an Enhanced Rule based Focused Crawler and URL rule based focused crawler. A working process of a focused crawler is composed of two main steps. The first step is to determine the starting URLs and specify user interest. The crawler is unable to traverse the Internet without starting URLs. The second step in a focused crawling process is the crawling method. In theoretical point of view, a focused crawler smartly selects a direction to traverse the Internet. A clever route selection method of the crawler is to arrange URLs so that the most relevant ones can be located in the first part of the queue. The queue will then be sorted by relevancy in descending order. The performance and efficiency of a focused crawler is mainly determined by the ordering strategy that determines the order of page retrieval.

3. THE PROPOSED ARCHITECTURE

The proposed architecture is shown below in Figure 1.

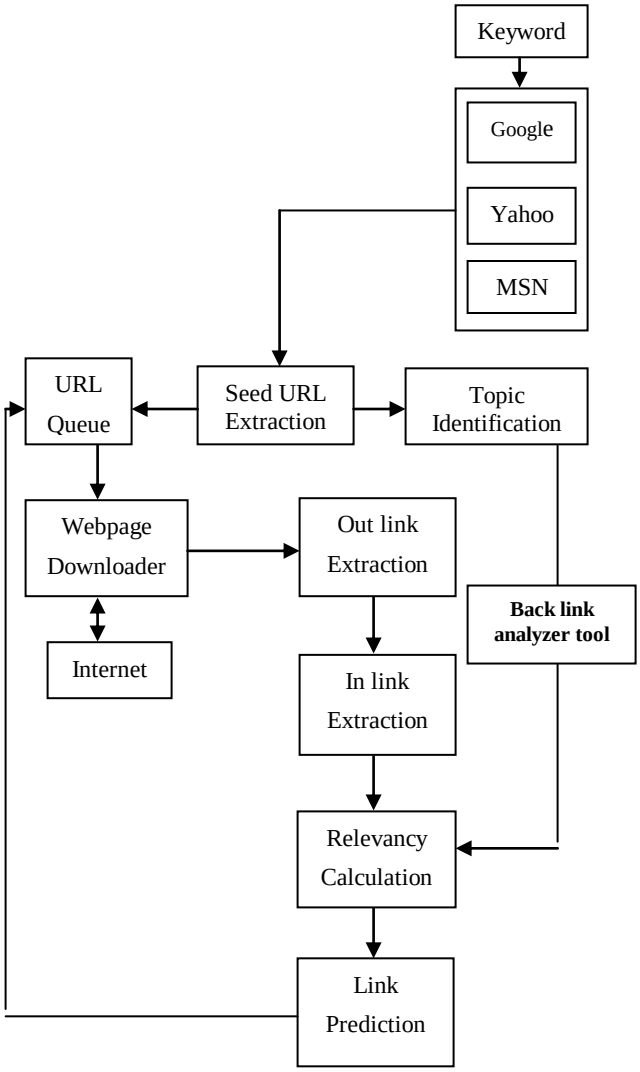

Figure 1. The Proposed Architecture

4. PROPOSED APPROACH

4.1 Seed URL Extraction

In our proposed approach, seed URLs are extracted by one search engine known as threesearches.com. We put a query in this search engine and it shows the result of three most popular search engines like Google, Yahoo, and MSN search. We extract resulting URLs based on certain condition. We first extract URLs which are common in all three search engine results. We assume that this common search result URLs are most relevant for this query and thus these URLs are grouped into most relevant group seed URLs. Now, we extract those URLs which are common in

any two search engine results, such as in "Google and Yahoo" or "Yahoo and MSN search" or "MSN search and Google". We put these extracted URLs in relevant group. Our proposed approach is based on Naïve Bayesian classification approach. We know that in Naïve Bayesian classification approach, to classify the unvisited URLs, we require some pretrained data about visited URLs. Based on visited URLs property, we classify the unvisited URLs. There are positive and negative results in pretrained data. So, for negative results, we put the query in threesearches.com with negative (-) sign, and here we put a positive query "computer books" and negative query –computer books (use negative sign (-) before query) in threesearches.com and find out resulting URLs. URLs www.freecomputerbooks.com and www.computer-books.us belong to all three search engine results. So, it is categorized as most relevant seed pages. URLs www.onlinecomputerbooks.com, www.freetechbook.com, www.intelligentedu.com/free_computer_books.html, and www.computer-books.us belong to two search engine results. So, it is categorized as relevant seed pages and URLs www.facebook.com, www.books.google.com/books belong to two search engine results but those are found out by negative query so it is categorized as irrelevant seed pages. Based on the results, we categorize seed URLs into groups in Table 1.

Table 1. Seed URLs Table

Seed URLs	Categories
www.freecomputerbooks.com	Most relevant
www.computer-books.us	Most relevant
www.onlinecomputerbook.com	Relevant
www.freetechbook.com	Relevant
www.intelligentedu.com/free_computer_books.html	Relevant
www.facebook.com	Irrelevant
www.books.google.com/books	Irrelevant

The negative query result seed URLs are categorized into irrelevant categories.

4.2 Topic Specific Weight Table Construction

Weight table defines the crawling target. The topic name is sent as a query to the Google Web search engine and the first k results are retrieved. The retrieved pages are parsed. To avoid indexing useless words, a text retrieval system often associates a stop list with a set of documents. A stop list is a set of words that are deemed "irrelevant." Stop lists may vary per document set. For example, database systems could be an important keyword in a newspaper. However, it may be considered as a stop word in a set of research papers presented in a database systems conference. A group of different words may share the same word stem. A text retrieval system needs to identify groups of words where the words in a group are small syntactic variants of one another and collect only the common word stem per group. Starting with a set of d documents and a set of "t" terms, we can model each document as a vector "v" in the "t" dimensional space "R^t", which is why this method is called the vector-space model. Let the term frequency be the number of occurrences of term "t" in the document d, i.e., freq(d, t). The (weighted) term-frequency matrix TF(d, t) measures the association of a term "t" with respect to the given document "d." It is generally defined as 0 if the document does not contain the term, and nonzero otherwise.

Order the word by their weights and extract a certain number of words with high weight as the topic keywords. After that weights are normalized as:

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

where "Wi" is the weight of keyword "i", and "Wmax" is weight of keyword with highest weight.

For example, we have taken a topic keywords "computer books." For Topic Specific Weight Table construction, we put the "computer books" is as a query to the Google Web search engine and the first 7 results are retrieved. After removing stop words except word computer (as we know that computer is a stop word, our query is "computer book"; so we take a word computer as an important word) and stemming the words, for calculating the weights the term frequency (tf) and inverse document frequency of each word is calculated. Here, we have taken top 10 most occurrences words in Table 2.

Table 2. Topic Specific Weight Table

Terms	Weight
Book	1
Free	0.894259882
Program	0.459214501
Computer	0.380664665
Web	0.25679758
Ebook	0.25679758
Site	0.250755287
Linux	0.223564954
Java	0.208459214
Post	0.187311178

4.3 Relevancy Calculation

The weight of words in page corresponding to the keyword in the Topic Specific Weight Table is calculated. The weight calculation of words in page uses same approach which is used by Topic Specific Weight Table weight calculation. In our proposed approach, it uses a cosine similarity measure to calculate the relevance of the page on a particular topic.

$$\text{Relevance}(t, p) = \frac{\sum wkt \times wkp}{\sqrt{((\sum wkt \times wkt) \times (\sum wkp \times wkp))}}$$

where "t" is the topic specific weight table, "p" is the web page under investigation, "wkt" and "wkp" are the weights of keyword "k" in the weight table and in the web page respectively.

5. EXPERIMENTAL ANALYSIS

There are different types of URL attributes for measuring that a particular link is relevant for the topics or not.

1. Average Parent Page Relevancy
2. Anchor Text Relevancy
3. URL Description Relevancy
4. Cohesive Text Relevancy

5.1 Average Parent Page Relevancy

Based on seed pages, we can analyze the category of unvisited link. It means that the unvisited URL is relevant to the topics or not. In our proposed approach, first we extract all parent pages of unvisited link and then we measure the relevancy of parent pages with topic keywords.

5.1.1 Seed URLs

1.1. www.freecomputerbooks.com

Parent page of this seed URL are:

http://librarykvpattom.wordpress.com/2007/11/

http://www.vocescuola.it/tag/noun/

http://freecomputerbooks.com/store.html

http://liberdadegrafica.blogspot.com/2008_12_01_archive.html

http://www.leren.nl/rubriek/computers_en_internet/software_ontwikkelen/oo/

There are number of parent pages but for analysis we have taken only five parent pages.

1.2. www.computer-books.us

http://itdiscover.com/links/99_free_and_best_books_tutorials_sites_all_programming_languages

http://toniocastro.wordpress.com/category/virutalizacion/

http://aaacomputer.com/HintsandTips/tabid/99/ctl/ArticleView/mid/433/articleId/246/100-Plus-Sites-to-check-out.aspx

http://smallvoid.com/links/developer/

http://madsyair.wordpress.com/

1.3. www.onlinecomputerbooks.com

http://www.onlinecomputerbooks.com/site-map.php

http://www.vocescuola.it/tag/genoma/

http://www.scmad.com/j2me-tutorials1.php

http://liberdadegrafica.blogspot.com/2008_12_01_archive.html

http://www.gayanb.com/articles_javaboutique.php

1.4. www.freetechbooks.com

http://zaidlearn.blogspot.com/2008/06/university-learning-ocw-oer-free.html

http://www.kiet.edu/Library/e-ref.htm

http://librarykvpattom.wordpress.com/2007/11/

http://kindlehomepage.blogspot.com/2010/03/kindle-nation-daily-free-book-alert-for_25.html

http://www.vocescuola.it/2010/03/16/800-000-e-book-gratuiti-possono-bastare/

1.5. www.intelligentedu.com/free_computer_books.html

http://intelligentedu.tradepub.com/?pt=cat&page=Info

http://www.ebookslab.info/download-free-e-books-html

http://www.degreetutor.com/library/career-starter/115-secrets

http://www.intelligentedu.com/newsletter46.html

http://www.intelligentedu.com/newsletter69.html

1.6. www.facebook.com

http://watch.discoverychannel.ca/

http://www.universiag10.org/tag/iberoamerica/

http://www.alumni.ucdavis.edu/s/787/start.aspx?sid=787&gid=1&pgid=336

http://www.martinoticias.com/MoreStories.aspx

http://www.forumblog.org/blog/religion/

1.7. www.books.google.com/books

http://googleblog.blogspot.com/2006/10/scary-stories.html

http://www.googlelabs.com/?tags=apps&sort_by=popularity

http://www.wired.com/epicenter/2010/01/google-apologises-to-chinese-writers-overbook-flap-yahoo-news/

http://notes.kateva.org/2010_01_01_archive.html

http://scienceblogs.com/bookclub/

Now, we calculate the average parent page relevancy score of each of 7 URLs. Relevancy score of each parent page is calculated by the vector space model. The weight of words in page corresponding to the keyword in the table is calculated, and we find out which URLs have relevant average parent page relevancy score and which URLs have irrelevant average parent

page relevancy score. If the average parent page relevancy score is greater than some threshold value, then it is identified as relevant. Otherwise, it is irrelevant (see Table 3).

Table 3. Average Parent Page Relevancy Score Table

URLs	Average Parent Page Relevancy Score
www.freecomputerbooks.com	0.694802718
www.computer-books.us	0.797621566
www.onlinecomputerbook.com	0.648112149
www.freetechbook.com	0.828339066
www.intelligentedu.com/free_computer_books.html	0.712560398
www.facebook.com	0.1854898548
www.books.google.com/books	0.532127555

From experiment, it has been seen that if threshold value is 0.7, i.e., if average parent page relevancy score is greater than 0.7, then this value is relevant. Depending upon the parameter threshold, value is changed. Now, if any URL's average parent page relevancy score is greater than 0.7, then its value is in table is "yes." Otherwise, its value is "no."

5.2 Anchor Text Relevancy

It is the relevancy between topic keywords and anchor text. We find out the related word of anchor text with the help of tool, and find out how much percentage of topic keywords are there in set of related words of topic keywords. The more topic keywords are in set of related words of anchor text, the anchor text is more relevant to topics. Our proposed approach mentions this attribute anchor text relevancy because anchor text describes the some information about URL. It is the textual information about URL. For example, in "http://www.freecomputerbooks.com" page there are numbers of URLs exist. We have taken one URL "http://www.freecomputerbooks.com/dbCategory.html" whose anchor text in this seed page is "databases and Storage." There are number of related words of the anchor text "databases and Storage".

"Alexa, amazon elastic compute cloud, amazon mechanical turk, amazon payments, amazon simple storage service, amazon web services, apis, archives and records management, article, articles, aws, aws user group, books, cloud computing, cultural property, data mining, database, database storage, databases, developer forum, developer tools, devpay, dynamic, ec2, file, flexible payments service, freedom of act, grid computing, image, information, information age, information literacy, information technology, journalism, journals, library science, management, media, modern, movie, newspapers in america, online library, photo, research, retrieval, simple queue service, simpledb, storage, text utility computing, web application development, web hosting, website."

Now, we find out weights of all related words in seed page. Weight calculation is done by "tf - idf" basis. Here, we have taken only one seed page. So, "idf" is calculated based on only one seed page. So, the anchor relevancy scores of "databases and Storage" anchor text is 0.2 (see Table 4).

Table 4. Anchor Tag Table

URLs	Anchor Tag	Keywords Present
www.freecomputerbooks.com	Free computer book	7
www.computer-books.us	Computer-book	7
www.onlinecomputerbooks.com	Online computer book	7
www.freetechbook.com	Free tech book	5
www.intelligentedu.com/free_computer_books.html	Books	3
www.facebook.com	Face book	7
www.books.google.com/books	Books	3

Here, keywords present 7 means 7 keywords out of 10 present in the set of related word of anchor text, and the threshold value is 0.7. As the score depends on page author, some author give matching name with URL in anchor text and some page author give unmatched name. Whether web page is relevant or irrelevant, it also depends upon some threshold value.

5.3 URL Description Relevancy

It is the relevancy score of URL description with respect to topics. We put the URL as a query to Google Search Engine with the name description of URL and find out top 10 results, and then find out top 10 weighed words after calculating the Term Frequency. Our proposed approach calculates relevancy score URL description because it gives detailed information about URL. In relevancy score URL description, we put the threshold value is 0.8 because here Google provides more description about topics.

For example, we put the URL as a query "description of http://www.freecomputerbooks.com" in Google search engine and find out weight of topic keywords in this description (see Table 5 and Table 6).

Table 5. Description of URLs Table

Terms (Description of www.freecomputerbooks.com)	Weight
Books	0.806486486
Free	1
Program	0.479892761
Computer	0.67131351
Web	0.128648648
E-Book	0.258378378
Site	0.064864864
Linux	0

Java	0
Post	0

Table 6. Description Score Table

URLs	Description Score w.r.t. Topic Keywords
www.freecomputerbooks.com	0.93615100
www.computer-books.us	0.81948239
www.onlinecomputerbooks.com	0.791265451
www.freetechbook.com	0.953638167
www.intelligentedu.com/free_computer_books.html	0.954884907
www.facebook.com	0.685672724
www.books.google.com/books	0.817865479

5.4 Cohesive Text Relevancy

Cohesive relevancy score of URL is the score of URL with respect to topics in sentence. For the extraction of cohesive-text, one sentence or group of meaningful sentences around the anchor link has to be considered. A sentence can be identified as starting with a capital letter and ends with a period (dot). The following algorithm describes the steps for extracting cohesive-text:

1. Identify the anchor link in the page.

2. Extract a sentence in the backward direction of the anchor link if any.

3. If this sentence starts with the words 'It', 'This', 'And', then extract one more sentence in the backward direction, if any.

4. Repeat steps 2 & 3 until the sentence starts with a word excluding the words mentioned in step 3.

5. Extract a sentence in the forward direction of anchor tag, if any.

After calculating relevancy score URL description, our proposed approach calculates cohesive relevancy score of URL because it gives information about schematic similarity of URL with respect to topic keywords. Cohesive relevancy score of URL is also calculated because it gives information about how many topic keywords surrounding the unvisited URLs are to be fetched. Cohesive relevancy totally depends on author because some authors give detailed information in cohesive text and some authors give less information of surrounding URLs. It also depends on some threshold value.

For example, in www.freecomputerbooks.com web page one anchor link is www.freecomputerbooks.com/specialWebServicesBooks.html and its surrounding text is "This book provides a pragmatic introduction to RESTFUL web services, and covers the key principles: Identifiable resources, links and hypermedia, standard methods etc". Here, we can see that the topic keywords "book", "web" exist in cohesive text of this link. So, the cohesive relevancy score of URL is 0.2 because out of 10 topic keywords 2 topic keywords surround the anchor link (see Table 7).

Table 7. Cohesive Relevancy Score Table

URLs	Cohesive Relevancy Score of URL
www.freecomputerbooks.com	0.9
www.computer-books.us	0.9
www.onlinecomputerbooks.com	0.7
www.freetechbook.com	0.5
www.intelligentedu.com/free_computer_books.html	0.7
www.facebook.com	0.4
www.books.google.com/books	0.7

Now, our proposed approach use Naïve Bayesian Classification to classify the unvisited URL that it is relevant or irrelevant with respect to topic keywords. Here, we use class label training tuple of seed pages because based on seed pages, we can predict that unvisited URL will be topic relevant or not. In Table 8, the class attribute is "relevant."

By experiment in URL www.books.google.com/books and www.freetechbook.com result is unpredictable because we have taken only 10 topic key words and these topic keywords are common in all sites which give information about books. But we think that if we take more topic key words, then this table RESULT may change.

By using Naïve Bayesian Classification [4] methods, we can find out that particular URL is relevant for the topic or not. For example, one unvisited URL is http://www.freecomputerbooks.com/javaCategory.html. Now, we have to find out this URL is relevant URL for topic or not.

The attribute value of this URL is

X = (average parent page relevancy = No, URL description relevancy = yes, Anchor text relevancy = no, cohesive text relevancy = yes)

If this URL is relevant for topic, then this URL will belong to relevant class. For this, we require $P(X|C_i)P(C_i) > P(X|C_j)P(C_j)$ for $i <= j <= m$, $j <> i$. In other words, the predicted class label is the class Ci for which P(X|Ci)P(Ci) is the maximum.

Here, class c1 is "yes" for relevant value and class c2 is "no" for relevant value. Now, we need to maximize $P(X|C_i)P(C_i)$, for i = 1, 2 only. The prior probability of each class can be computed based on training tuples.

P(Relevant = Yes) = 5/7

P(Relevant = No) = 2/7

P(average parent page relevancy = No|Relevant = Yes) = 2/5

P(average parent page relevancy = No|Relevant = No) = 2/2

P(URL description relevancy = Yes|Relevant = Yes) = 4/5

P(URL description relevancy = Yes|Relevant = No) = 1/2

P(Anchor text relevancy = Yes|Relevant = Yes) = 3/5

P(Anchor text relevancy = Yes|Relevant = No) = 1/2

P(cohesive text relevancy = Yes|Relevant = Yes) = 4/5

P(cohesive text relevancy = Yes|Relevant = No) = 1/2

Using the above probabilities, we obtain

P(X|Relevant = Yes) = P(average parent page relevancy|Relevant = Yes) * P(URL description relevancy|Relevant = Yes) * P(Anchor text relevancy|Relevant = Yes) * P(cohesive text relevancy|Relevant = Yes) = 0.512

P(X|Relevant = No) = P(average parent page relevancy|Relevant = No) * P(URL description relevancy|Relevant = No) * P(Anchor text relevancy|Relevant = No) * P(cohesive text relevancy|Relevant = No) = 0.0625

To find the class Ci that maximizes P(X|Ci)P(Ci), we compute

P(X|Relevant = Yes) * P(Relevant = Yes) = 0.1536 * 0.714285714 = 0.109714285

P(X|Relevant = No) * P(Relevant = No) = 0.125 * 0.285714285 = 0.017857142

Therefore, the Naïve Bayesian classifier predicts Relevant = yes for tuple X. From this calculation, it is showing that the link http://www.freecomputerbooks.com/dbCategory.html is relevant for topics.

6. CONCLUSION AND FUTURE WORK

Focused crawlers are becoming a more and more important topic, and focused crawling methods are important members in the search engine family.

One of the key problems of vertical search engines is to develop an effective algorithm for the topic-specific searching and similarity measurement. In our proposed approach, we calculate the relevancy of unvisited URLs based on Naïve Bayesian classification method.

Naïve Bayesian classifiers assume that the effect of an attribute value on a given class is independent of the values of the other attributes. They predict class membership probabilities, such as the probability that a given tuple belong to a particular class. Here, we have taken four attributes of URL and one attribute is class variable. Here, we can predict the unvisited URL relevancy based on existing URL attributes value and its class attributes value. In our future work, we'll avoid the problem of zero probability and find out the relevancy of unvisited URLs based on clustering approach.

7. REFERENCES

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

[2] Chakrabarti, S., Berg, M. V. D. and Dom, B. 1999. Focused Crawling: A New Approach to Topic-Specific Web Resource Discovery. Proc. Eighth Int'l World Wide Web Conf.

[3] Diligenti M., Coetzee F. M., Lawrence S., Giles, C. S., and Gori M. 2000. Focused Crawling using Context Graphs. 26th International Conference on Very Large Databases (VLDB), Cairo, Egypt, pp. 527-534.

[4] Han, J. and Kamber, M. 2003. Data Mining: Concepts and Techniques. San Francisco: Morgan Kaufman.

[5] Pal, A., Tomar, D. S., and Shrivastava, S. C. 2009. Effective Focused Crawling Based on Content and Link Structure Analysis. International Journal of Computer Science and Information Security (IJCSIS), Vol. 2, No. 1, June 2009.

[6] Sun, Y., Jin, P., and Yue, L. 2008. A Framework of a Hybrid Focused Web Crawler. 2nd International Conference on Future Generation Communication and Networking Symposia.

[7] Yuan, F., Yin, C. and Liu, J. Improvement of Page Rank for Focused Crawler. 8th ACIS International Conference on Software Engineering, Artificial Intelligence, Networking, and Parallel/Distributed Computing.

[8] Zhang, X., Zhou, T., Yu, Z., and Chen, D. 2008. URL Rule Based Focused Crawlers. IEEE International Conference on e-Business Engineering.

[9] Zhang, Y., Yin, C., and Yuan, F. 2007. An Application of Improved PageRank in Focused Crawler. 4th International Conference on Fuzzy Systems and Knowledge Discovery (FSKD).

Table 8. Naïve Bayesian Classification Table

S. No.	URL Name	Average Parent Page Relevancy	URL Description Relevancy	Anchor Text Relevancy	Cohesive Text Relevancy	Relevant
1	www.freecomputerbooks.com	No	Yes	Yes	Yes	Yes
2	www.computer-books.us	Yes	Yes	Yes	Yes	Yes
3	www.onlinecomputerbook.com	No	No	Yes	Yes	Yes
4	www.freetechbook.com	Yes	Yes	No	No	Yes
5	www.intelligentedu.com/free_computer_books.html	Yes	Yes	No	Yes	Yes
6	www.facebook.com	No	No	Yes	No	No
7	www.books.google.com/books	No	Yes	No	Yes	No