Study of Different Focused Web Crawler to Search Domain Specific Information

Nisha N. Pawar
Department of Computer Engineering
Pimpri Chinchwad College of Engineering
Pune-44

K. Rajeswari, PhD
Department of Computer Engineering
Pimpri Chinchwad College of Engineering
Pune-44

ABSTRACT
In human life, use of correct medicinal plant for treating diseases is very important. Medicinal plant is the important part of the Indian Ayurvedic system. So there is a need of heuristic search of medicinal plants over India. The huge amount of information related to medicinal plants is available on World Wide Web. For collecting such domain specific information from the internet, focused web crawler is useful. This paper proposed an efficient focused web crawler using hybrid model of Naïve Bayes classifier and Decision tree classifier. The result of classifier defines that web page is relevant or not relevant. The proposed hybrid classification will improve the accuracy of web crawling.

Keywords
Medicinal plants, Focused web crawler, Naïve Bayes classifier, Decision Tree.

1. INTRODUCTION
In the Ayurvedic system, use of correct medicinal plant for treating diseases gives better results. Different medicinal plant species are available in India. So there is a need of heuristic search of medicinal plant information over India. Many web sites give information related to medicinal plants. This information will be helpful for doctors and others who are interested in doing research on medicinal plants.

To extract those specific web pages, focused web crawler can be used. Typical web crawler collects some relevant web pages and some irrelevant web pages also. To improve the accuracy of data collection focused web crawling is used. Focused web crawler is used to search domain specific as well as most relevant web pages. The best first strategy is mostly helpful in focused web crawler to search web pages for domain specific queries. Hence the best first heuristic algorithm can be used in focused web crawler.

The goal of this paper is to design an efficient focused web crawler to search different medicinal plant information. Many classification techniques such as Naïve Bayes algorithm, decision tree, K nearest neighbor (KNN), Support vector machine, etc. were used in focused web crawler. The classification technique defines whether the current web page is relevant or not relevant for a given topic. Those relevant web pages then retrieved and stored in the repository. In this paper, the hybrid classification approach of Naïve Bayes classifier and Decision tree classifier is proposed for classifying the web pages. As decision tree classifier has potential to handle noisy data and Naïve Bayes is particularly suited for high dimensional inputs. This hybrid approach will improve the accuracy of classification methods used in other focused web crawlers for narrow segment of web.

The rest of the paper is organized into the following sections: Section 1 gives an introduction about the importance of plants in our life and heuristic search of medicinal plants. Section 2 describes a brief literature review on the various different papers of focused web crawler. Section 3 gives a description of focused web crawler. Section 4 describes proposed model for efficient web crawler. Finally, Section 6 concludes this paper**.**

2. LITERATURE REVIEW
In the paper [1], Madjid Khalilian, Hassan Abolhassani, Ali Alijamaat proposed system in which constructing a specific domain directory which is included high quality web pages has challenged the development of web portals. This paper proposed a semiautomatic framework that combines knowledge of human with automatic techniques. We also proposed methods for improving automatic component. Dynamic threshold with decay value can improve accuracy. Manifold ranking in this study couldn't help us with accuracy because of data nature.

In the paper [2], Anshika Pal, Deepak Singh Tomar, S.C. Shrivastava proposed a method for focused crawling that allows the crawler to go through several irrelevant pages to get to the next relevant one when the current page is irrelevant. This method for traversing the irrelevant pages that met during crawling to improve the coverage of a specific topic. This paper's approach has better performance than the BFS crawler.

In the paper [3], Nandar Win Min, Aye Nandar Hlaing introduced a new effective focused web crawler. It uses smart methods to speed up crawling of relevant pages and then follow the promising links first in order to find the most important, topically related pages. In this paper, particular emphasis is given to learning focused crawlers capable of learning not only the content of target pages but also paths leading to target pages. In fact, learning crawlers perform a very difficult task: they attempt to learn web crawling patterns leading to relevant pages possibly through other not relevant pages thus increasing the probability of failure. But, with a good crawling strategy, it seems to be possible to build crawlers that can rather quickly obtain a significant portion of the hot pages.

Table 1 Comparative Study of Realated Work

Paper Title	Techniques Used	Observations
PCI: 'Plants Classification & Identification' Classification of Web pages for constructing plants web-	•Ontology • Cosine similarity • Manifold	• Implements semiautomatic framework for creating plant web directory • Less accurate due to use manifold ranking for this data nature.

directory (2009)	ranking algorithm	
Effective Focused Crawling Based on Content and Link Structure Analysis (2009)	• Cosine similarity • Links ranking algorithm	• Gives better performance than BFS crawler • Better coverage of relevant web pages • Required more improvement on crawling efficiency.
An effective focused web crawler for web resource discovery (2013)	• Naïve Bayes Classifier • Cosine similarity	• Quickly obtain relevant pages • Capable of learning content and links of target web pages
Ranking Hyperlinks Approach for Focused Web Crawler (2014)	• KNN Classifier • Cosine similarity	• Decay concept is used to enhance the accuracy of crawling. • Improves crawling efficiency.

In the paper[6], Sameendra Samarawickrama1, Lakshman Jayaratne have discussed a new approach to focus crawling based on named entities for narrow domains. This paper has conducted experiments in focused web crawling in three narrow domains: baseball, football and American politics. A classifier based on the centroid algorithm is used to guide the crawler which is trained on web pages collected manually from online news articles for each domain. The collection built with the proposed crawler is better than the traditional focused crawler based on lexical terms, in terms of the harvest ratio.

In this paper [7], Nandar Win Min, and Aye Nandar Hlaing introduced an effective focused web crawler containing smart methods. In text analysis, similarity measurement applies to different parts of the Web pages including title, body, anchor text and URL tokens. It can increase the relevance and quality of the Web pages pointed to by target URLs. To enhance the accuracy of crawling, decay concept is used to determine the optimal order in which the targeted URLs are visited. In this measurement, two kinds of threshold are used to limit the crawler to the effective web pages. Finally, to provide sorting URLs, priority equation is used.

Table 1 shows a study of papers related to focused web crawler.

3. FOCUSED WEB CRAWLER

A focused web crawler is a web crawler that attempts to search and retrieve web pages that relevant to a specific domain. Figure 1 shows the system Architecture of focused web crawler. The perfect focused crawler retrieves the maximal set of relevant pages while concurrently traversing the minimal number of irrelevant documents on the web [5]. In focused web crawling, seed URLs are used to initiate the crawling process. Those seed URLs are called as 'seed set'. Each URL is visited by the web crawler. It identifies the different hyperlinks in the page Different pages from the internet are downloaded by the parser and the generator and stored in the database system of the search engine. The URLs are then placed in the queue. During each crawling iteration, the top link is selected and then that web page is classified using classification method. The web page is classified as relevant or irrelevant. The relevant URL is added to crawler frontier. This process is continued until the URL queue is empty or the crawl limit has been met.

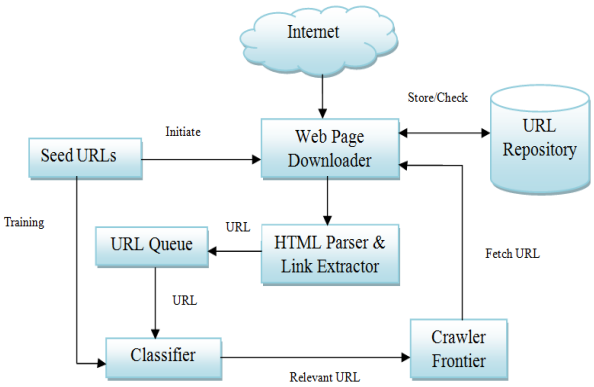

Figure 1 System Architecture of focused web crawler

4. PROPOSED ARCHITECTURE

In this paper, we have proposed efficient web crawler to search different medicinal plant information. In the first module of the proposed model, query related to medicinal plant is given by the user. The input query is tokenized using tokenization process. Then stopwords are removed from the tokenized input query. Stemming process is done on these keywords where stemming is a process of converting grammatical variations of words into base form. This preprocessed query is given to the next module of the proposed model.

Efficient Focused Web Crawler :

Figure 2 shows the proposed model of efficient web crawler. In this model, depending on the input keyword or query, related web documents downloaded from the Internet.

a. Classification:

In the classification module, the downloaded web document is classified using a hybrid model of Naïve Bayes classification algorithm and Decision tree algorithm with NEs terms and following four lexical features:

The notation for feature k and observed web page O is fk(O).

1. Title Text Feature:

Maximal similarity value between the title of the content of a given candidate page and the set of targets [4].

f1(O) = similarity of topic and keywords of page O in title

2. Body Text Feature:
Maximal similarity value between the body of the content of a given candidate page and the set of targets [4].

f2(O) = similarity of topic and keywords of page O in body

3. Anchor Text Feature:
The anchor text around the link pointing to an observed page O often is closely related to the topic of the page[4].

f3(O) = similarity of topic and keywords of page O in anchor text.

4. URL Tokens Feature:
The tokens in the URL of an observed page may contain valuable information about predicting whether or not a page is a target page or potentially leading to a target[4]. We first parse the URL into tokens, then compute the similarity between tokens and topic keywords.

f4(O) = similarity of topic and keywords of page O in URL tokens.

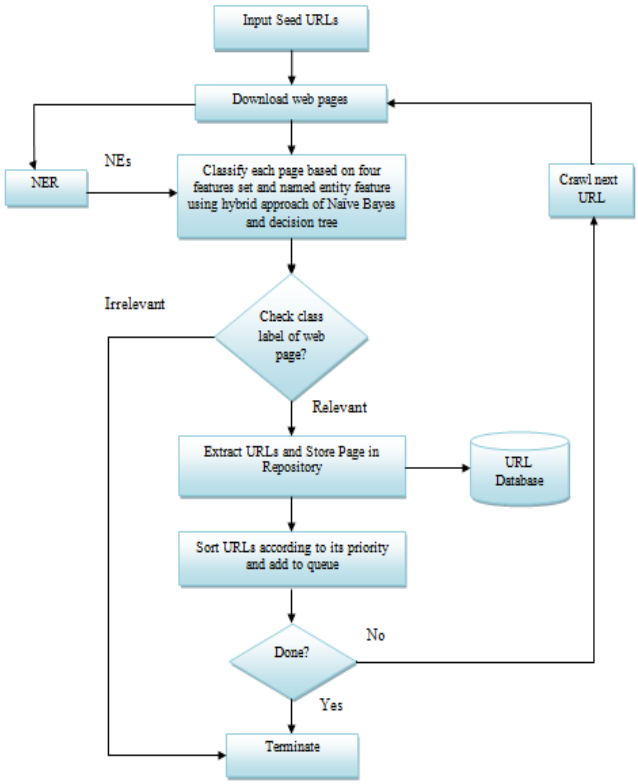

Figure 2 Proposed model of focused web crawler

Firstly, training set is built with those four relevance attributes. This training set is used to train the system. Then classification method is trained by using training set. In our proposed model, hybrid approach of the Naïve bayes algorithm and the decision tree algorithm is used to define whether current web page is relevant or not related to the medicinal plant. In hybrid approach, decision tree will be constructed by using lexical features of web pages. One of the features will be selected as a root of the tree and continue constructing branches and leaf. The Naïve Bayes classifiers are added to the leaf nodes of the tree. If web page is related to medicinal plant information then it is classified as "*Yes*", if web page is not related to medicinal plant then classified as "*No*".

b. Sorting priority queue

The web pages, classified as a relevant, are stored and sorted in priority queue. To sort the priority queue, average of values of above four features is calculated.

c. Crawling

In crawling phase, the next top link is fetched from the priority queue. The best first search crawling will be used to fetch new web pages.

This complete process continues until crawl limit reaches.

5. CONCLUSION

We have proposed a framework of an efficient web crawler to organize large amount of different medicinal plant's information in a web directory. In focused web crawler, higher the accuracy of classification method higher the relevant web pages will be collected. In this paper, the hybrid classification approach of combining the Naïve Bayes classification and Decision Tree classification is used in proposed focused web crawler. This focused web crawler will provide more accurate information for a given query.

6. REFERENCES

[1] Madjid Khalilian, Hassan Abolhassani, Ali Alijamaat, "*PCI: 'Plants Classification & Identification' Classification of Web pages for constructing plants web-directory*", Sixth International Conference on Information Technology: New Generations,2009

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

[3] Nandar Win Min, Aye Nandar Hlaing, "*An effective focused web crawler for web resource discovery*", International Journal of Advanced Research in Computer Engineering & Technoogy(IJARCET) Volume 2, Issue 11, November 2013

[4] Yajun Du,∗, Wenjun Liu, Xianjing Lv, Guoli Peng, "*An improved focused crawler based on Semantic Similarity Vector Space Model* ", Applied Soft Computing 36 392–407, 2015.

[5] Sunita Rawat, D. R. Patil, "*Efficient Focused Crawling based on Best First Search*", 978-1-4673-4529-3/12/$31.00_c 2012 IEEE.

[6] Sameendra Samarawickrama1, Lakshman Jayaratne, "*Focused web crawling using named entity recognition for narrow domains*", IJRET: International Journal of Research in Engineering and Technology, Volume: 02 Issue: 03, Mar-2013.

[7] Nandar Win Min, and Aye Nandar Hlaing, "*Ranking Hyperlinks Approach for Focused Web Crawler*", International Conference on Advances in Engineering and Technology (ICAET'2014) March 29-30, 2014 Singapore.

[8] S. Chakrabarti, M. van der Berg, and B. Dom, "*Focused crawling: a new approach to topic-specific web resource discovery*," in Proc. of the 8th International World-Wide Web Conference (WWW8), 1999.

[9] J. Cho, H. Garcia-Molina, and L. Page, "Efficient crawling through URL ordering," in Proceedings of the Seventh World-Wide Web Conference, 1998.

[10] Wenxian Wang, Xingshu Chen, Yongbin Zou,Haizhou Wang, Zongkun Dai, "A Focused Crawler Based on Naive Bayes Classifier", Third International Symposium on Intelligent Information Technology and Security Informatics,2010.

[11] Gunjan H. Agre, Nikita V.Mahajan, "Keyword Focused Web Crawler", IEEE sponsored 2nd international conference on electronics and communication systems(icecs) 2015.

[12] H. Isozaki and H. Kazawa, "Efficient support vector classifiers for named entity recognition", in Proceedings of the 19th international conference on Computational linguistics-Volume 1. Association for Computational Linguistics, 2002.

[13] Anish Gupta, Priya Anand, "Focused web crawlers and its approaches", 1st International Conference on Futuristic trend in Computational Analysis and Knowledge Management (ABLAZE),2015.

[14] Mukesh Kumar, Renu Vig, "Learnable Focused Meta Crawling Through Web", 2nd International Conference on Communication, Computing and Security (ICCCS),2012.

[15] Li, Jun, Furuse, K. and Yamaguchi, K.,"Focused Crawling by Exploiting Anchor Text Using Decision Tree", Proceedings of the 14th International World Wide Web Conference. 2005, pp. 1190-1191.