# RELEVANCE PREDICTION IN FOCUSED CRAWLING: A SURVEY

*DEEPALI DEV*
**Department of Computer Science and Engineering**
**National Institute of Technical Teachers Training and Research, Punjab University, Chandigarh-160019, India**
**ddepali@gmail.com**

***ABSTRACT :*** *A focused crawler is a web crawler that traverses the web to explore information that is related to particular topic of interest only. This paper deals with survey and comparison of various focused crawling techniques based on various parameters to find out the advantages and limitations for relevance prediction of URLs. In this paper we discuss and compare different relevance prediction techniques in focused crawling i.e. relevance prediction based on content i.e. Fish Search and Shark Search, relevance prediction based on content and link analysis i.e. Hawk, relevance prediction based on classifier.*

***KEYWORDS: Focused Crawler, Relevance Prediction***

## 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 800 Gigabytes of information on these pages were updated every month and the size of the Web was doubled every year [14]. 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 [1].
A crawler is a program that automatically collects Web pages to create a local index. Crawling the Web quickly and entirely is an expensive, unrealistic goal because of the required hardware and network resources. 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. A focused crawler tries to identify the most promising links, and ignores off-topic documents [3].
The Key problem of focused crawler is how to determine whether a particular URL is relevant to search topic or not.
The outline of paper is as follows. We discuss basic architecture of focused crawler in section 2. In section 3 we discuss relevance prediction based on content i.e. fish search and Shark search. In Section 4 we discuss Hawk, relevance prediction based on link and content analysis. In section 5 we are discussing relevance prediction of unvisited URL based on classifier. In section 6 we compare all the relevance prediction techniques. In section 7 we have concluded our survey paper.

## 2. THE ARCHITECTURE
Figure 1 shows the architecture of focused crawling System.

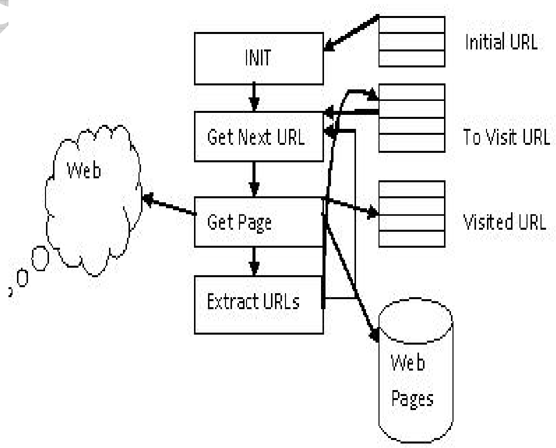


Figure 1 Architecture of focused Crawler

A crawler starts off with an initial set of URLs called seed URLs. It first retrieves the pages identified by the seed URLs, extracts any URLs in the pages, and adds the new URLs to a queue of URLs to be scanned. Then the crawler gets URLs from the queue (in same order), and repeats the process.

## 3. RELEVANCE PREDICTION BASED ON CONTENT
In this crawler, it performs search strategy based on content of web page. We will discuss Fish search and Shark search that performs searching based on content of page.

JIKRCE

### 3.1 FISH SEARCH

In "Fish-search"[12], 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).

Fish-search algorithm treats Internet as a directed graph, webpage as node and hyperlink as edge, so the search operation could be abstracted as a process of traversing directed graph. For every node we judge whether it is relative, 1 for relevant, 0 for irrelevant. Fish-search algorithm maintains a list, which keeps URL of page to be searched. The URLs have different priority, the URL with more superior priority will be located at the front of the list, and will be searched sooner than others. If relative page is found, it stands for that the food has been found

by the fish, and more healthy reproduction.

But it assigns a relevance score in a discrete manner (1 for relevant, 0 or 0.5 for irrelevant) using primitive string- or regular-expression match. More generally, the key problem of the fish-search algorithm is the very low differentiation of the priority of pages in the list.

### 3.2 SHARK SEARCH

Shark search is 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 anchor text that occurs around the link in the web page.

In "Shark Search"[9], One immediate improvement is that instead of the binary (relevant/irrelevant) evaluation of document relevance, it returns a "fuzzy" score, i.e., a score between 0 and 1 (0 for no similarity whatsoever, 1 for perfect "conceptual" match) rather than a binary value.

It uses a Vector space model [10] for calculating the relevance Score. In the vector space model, the similarity between a document and a query is usually based on the distance between the vectors in some metric. The cosine similarity measure is the most common, as

$$Sim(q,d) = \frac{\sum w_{td} w_{tq}}{\sqrt{\sum w_{td}^2 \sum w_{tq}^2}} \qquad (1)$$

In equation (1) above d is the web document fetched and q is query. $W_{tq}$ is weight of words in query and $W_{td}$ is weight of words in document. The relevance score obtained from the equation is between 0 and 1.

But this algorithm neglect information of link structure.

### 3.2.2 TOPIC WORD WEIGHT TABLE

Topic word weight table keeps the weights of topic words. To create table, topic word is sent as a query to a search engine and first n results are retrieved [5].

We use standard tf x idf weighting method [6] to calculate the weight of each term. In this method, tf is number of occurrences of word w in the document and idf varies inversely with the number of documents in the collection that w occurs in. Words are ordered by their weights and first n words are selected as topic keywords.

Then,

$$Weight = \frac{w_i}{w_{max}} \qquad (2)$$

## 4. RELEVANCE PREDICTION BASED ON CONTENT AND LINK

In this crawler, it combines search strategy based on content and link structure. Search Strategy based on Link structure determines the importance of page by analyzing the mutual relations.

### 4.1 HAWK

In HAWK [2] algorithm, it selects and predicts the relevant URL based on content of web page, and then determines the priority of URL in the queue to be crawling.

Advantage of HAWK crawler is that not only it uses the content of web page to improve the page relevance, but also uses the link structure to improve the coverage of a specific topic.

### 4.1.1 RELEVANCE COMPUTATAION

**Definition 1:**

Topic vector T is a topic and denotes the topic vector.

$$T = [(k_1, w_1)(k_2, w_2) \ldots \ldots (k_j, w_j) \ldots\ldots (k_l, w_l)]^T$$

Where $k_j$ denotes $j^{th}$ keyword or phrase of topic T.

$w_j$ is the weight of the $j^{th}$ keyword or phrase, and $\sum w_j = 1$ , $1 \le j \le l$ . $l = \|T\|$, is the amount of Keyword of topic T.

**Definition 2:**

$U_k = \frac{\|UK_k\|}{\|UD\|} * w_k$ denotes the contribution of D for $k^{th}$ keyword of topic T, where $\|UK_k\|$ is the frequency that $k^{th}$ keyword $K_k$ of topic T appears in the web. $\|UD\|$ is amount of effective words in D. $w_k$ is weight of $k^{th}$ keyword in topic T.

**Definition 3:**

Relevance-score: The relevance-score represents the relevance-score of a page. The relevance-score of the page D is defined as follows:

$$Sim(T,D) = \sum_{k=1}^{l} u_k$$

(4)

JIKRCE

Where $l = \|T\|$, it is length of T, $u_k$ is contribution of D for $k^{th}$ keyword of topic T [2].
The relevance score from equation 4 lies between 0 and 1.

## 5. RELEVANCE PREDICTION BASED ON CLASSIFIER

Relevance prediction based classifier is learning based approach to improve the relevancy prediction of unvisited URLs without downloading and visiting many irrelevant pages [11]. In this technique, classification of unvisited URLs is done based on visited URLs attribute score, i.e. Anchor text relevancy, cohesive text relevancy, parent page relevancy, URL relevancy . Relevancy score is calculated based on vector space model and classification [8] is done by supervised or unsupervised classifier.

Classifier with supervised training requires a set of labeled document for its training. Naïve Bayesian, support vector machine, nearest neighbor, Decision tree, and neural network are most popular classifier [7].

Classifier with unsupervised learning use similarity measure when making relevance prediction. Cosine similarity measure is considered as most popular in survey of focused crawling.

### 5.1 RELEVANCY CALCULATION

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

## 6. COMPARISONS

This paper surveyed several relevance prediction techniques in focused crawling classifying them into three categories namely, relevance prediction based on content , relevance prediction based on content and link, relevance prediction based on classifier. These categories are not mutually exclusive and contain certain features common in all [13].

Crawlers based on content analysis make relevance prediction based on content of retrieved web page. I observed it does not utilize methods to identify potential URLs and are greedy to crawl through all URLs found in relevant page. It wastes extra amount of storage and network bandwidth.

Crawler based on content and link analysis considers both content and link when making its relevance judgment. But Link structure analysis has its own problems due to high dynamic nature of web. Studies have shown that within a year, 80 percent of all links in the link structure will have to changed or be new, 50 percent of all contents will be changed, 20 percent of web pages today will disappear [4].

Crawler based on classifier that use training paradigm should be very concern of their training data. The quality of training data is very important because it affects the effectiveness and performance of crawler.

## 7. CONCLUSION

Apart from relevance prediction based on content, both relevance prediction based on content and link and relevance prediction based on classifier are not **domain** specific, giving high productivity and adaptability.

## 8. REFERENCES


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

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

[3] Debashish, Amritesh, Lizashree "Unvisited URL Relevancy Calculation in Focused Crawling based on Naïve Bayesian Classification" International Journal of Computer Applications volume 3, July 2010.

[4] D.Lew, H. Wahlig, and G. Meyer-bautor. The freshness of web search engine Databases, 2006.

[5] D. Taylan, M. Poyraz, S. Akyokus, and M. C. Ganiz, "Intelligent Focused Crawler: Learning Which Links to Crawl", In Proc. Intelligent Systems and Applications Conference (INISTA), pp. 504-508, 2011.

[6] G.Salton and c. Buckley " Term weighting approaches in automatic text reterival" Inf. Process .Manage vol 24, pp. 513-523, 1988.

[7] G.Pant and P. Srinivasan . Learning to Crawl: Comparison classification schemes ACM Trans Inf. Sys. 23:430-462, October 2005.

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

[9] H. Michael, J. Michal, M. Yoelle, P. Dan, S. Menachem, and U. Sigalit, "The Shark-Search Algorithm - An Application: Tailored Web Site Mapping", In Computer Networks and ISDN Systems, vol. 30, no 1-7, pp. 317-326, 1998.

[10] Hong Wei, Cui, Xu-Cheng " An improved Topic Relevance Algorithm for Focused Crawling", IEEE 2011.

[11] Mejdl, Abdullah, Dunren "Improving the Relevance Prediction for Focused Web Crawler" IEEE, 2012

[12] P. De Bra, G-J Houben, Y. Kornatzky, and R. Post, "Information Retrieval in Distributed Hypertexts", In Proc. of RIAO-94, Intelligent Multimedia, Information Retrieval Systems and Management, pp. 481- 492, 1994.

[13] Sameendra, Lakshman "Automatic Text Classification and Focused Crawling" IEEE, 2011

[14] www.google.com


JIKRCE