Focused Crawler based on Efficient Page Rank Algorithm

Anand Ratna
Dept. Of CSE
Galgotia College
Gr. Noida, India

Divya
Dept. Of CSE
Galgotia College
Gr. Noida, India

Akshay Sawhney
Dept. Of CSE
Galgotia College
Gr. Noida, India

ABSTRACT
The size of the WWW is increasing rapidly and its nature is dynamic, building an efficient search mechanism is very necessary. A vast number of pages continually being added every day, so fetching information about a special-topic is gaining importance, which poses exceptional scaling challenges for general-purpose crawlers and search engines. This paper describes a web crawling approach based on best first search. Instead of collecting and indexing all available web documents to be able to answer all possible queries, a focused crawler choose the links that are likely to be most relevant for the crawl, and avoids irrelevant links of the document. This leads to significant savings in hardware as well as network resources and also helps keep the crawl more up-to-date. To accomplish such goal-directed crawling, select top most K relevant documents for a given query and then expand the most promising link chosen according to link score, to circumvent irrelevant regions of the web.

General Terms
Web Crawler, search engine, hyperlink.

Keywords
Focused web crawler, TF-IDF, Relevancy calculation, Page Rank.

1. INTRODUCTION
There are basically two types of web crawling strategies used by web search engines breadth first search and best first search. In breadth first search technique, it starts with some set of pages and then explores other pages by following links in breadth first [1] manner. Breadth first technique is mostly used in general purpose web crawler. The best first strategy is used for the retrieval of pages, which are relevant to particular topic or query. The crawler implemented using "best" first strategy is called "focused crawler".

The focused crawler has following components: (a) how to proceed from seed URLs and (b) determining if a particular page is relevant to the particular topic. An early search engine using focused crawling technique proposed in [2] shows a drawback when no relevant page about the topic, is not directly connected it stop.

Focused crawlers[3, 4]aim to search and retrieve onlythe subset of the World Wide Web that is relevant to the given topic. The good focused crawler retrieves the maximal set of relevant pages and traversing the minimal number of irrelevant documents on the web. Hence focused crawlers offer a potential solution for categories where content changes quickly. A previous focused crawling strategy proposed in [5] founded on the perception that relevant pages contain relevant links mostly.The above crawlers show an important drawback when the pages related to a given topic are don't get the advantage of "incoming links". Focused crawlers are furthermore well suited to efficiently generate indices used for niche search engines maintained by portals and user groups [6], where limited bandwidth and storage space are the norm [7]. Lastly, due to the limited resources used by a good focused crawler, users are already using personal computer based implementations [8]. Ultimately simple focused crawlers might become the method of choice for users to perform inclusive searches of web-related materials.

Proposed focused crawler aims at providing an alternative for conquering the issue that pages which are more relevant but having low frequency of topic. It helps in assigning weight on the basis of incoming links using page rank algorithm. By retrieving those pages which are reachable from the initial seeds, a set of candidate pages is obtained. The page which has the highest score with respect to the given topic, from the obtainable set of candidate pages. The crawling process will get continue by the newly added pages.

The rest of the paper is structured as follows. In section 2 an appraisal of the related work is presented. In Section 3 and 4 the design rationale of the system is covered in detail. In Section 5 the system architecture is discussed. In Section 6the focused crawling algorithm is described. At last, the paper is concluded.

2. RELATED WORK
A classic focused crawler takes the query from the user as input that describes the topic, a set of starting seed page URLs and guides the search towards the page of interest. They incorporate a criteria for assigning higher download priorities to the link based on their likelihood to lead to the pages on the topic of query. Pages pointed by the links with higher priorities are downloaded first. The early generation of crawlers which were used for indexing the web in web search engines rely on traditional graph algorithm, such as breadth-first traversal or depth-first traversal. The main aim of the crawl is to cover the whole web.

The crawling process of the topic oriented crawler focuses on the pages pertinent to the topic or query. They try to maintain the Web page downloaded for processing [9, 10] to a least amount, whereas maximizing the number of pertinent pages. The performance of the process is mainly based the selection of the seed pages and seed URLs. In general the user provide input to the crawler as a set of seed pages and these seed pages are selected amongst the top answers returned by a Web search engine [11] related to a query [12] . The pages which are pertinent to the topic or pages from which pertinent pages can be accessed are high quality seed pages.

A focused crawler model the search based on both the link structure and the content of the web [2]. It completely seeks out documents about a specific topic. A focused crawler produce a strategy in which it associates a score with each link in the pages it has downloaded [9, 13]. Those links are

inserted into the queue once it has been sorted according to the scores. Therefore this strategy ensures a promising crawl paths.

A focused crawler is a computer program which is used to find specific information of interest from the World Wide Web. The main aim of the focused crawler is that the crawler selects and retrieves only the pertinent pages and does not select all the web pages. A crawler cannot predict how pertinent a web page is [14] as it is only a computer program. For finding the pages of a specific topic or type, focused crawlers aspire to recognize links that are probably to direct to target documents. Previously Fish Search algorithm and Shark Search algorithm were used for crawling with topic keyword mentioned in query.

The system is driven by query in the Fish Search algorithm. It begins with a set of seed pages and takes the pages which have content similar to the provided query and their relative pages. In Fish-Search technique simple keyword matching is performed by assigning binary priority values (1 for relevant and 0 for not relevant) to pages candidate. Therefore equal value are assigned to all the pertinent pages. The advance algorithm of Fish Search algorithm is Shark Search algorithm where a child link inherits a partial value of the score of its parent link. Vector Space Model (VSM) proposed the Shark-Search method for allocating non binary priority values to candidate pages. The score depends on the anchor text which arises around the link in the Web page [15] and is united.

3. TF-IDF RANKING

TF-IDF, term frequency – inverse document frequency, is a mathematical statistic. It calculates importance of a word to a document in a corpus. Top n pages i.e. seeds for given query is provided by TF-IDF ranking algorithm.

3.1 Term Frequency (TF)

A weight is assigned to each term in a document depending on the number of time that term occurs in the document. This weight is referred to as term frequency.

where,

tf(t,d) : frequency of term t in document d

freq(t,d) : number of *occurrences* of t in d

tf(t,d) is a normalized frequency, to prevent a bias towards bigger documents, e.g. frequency divided by the maximum frequency of any term within the document [16].

$$tf(t,d) = \frac{f(t,d)}{\max\{f(t,d) : t \in d\}} \quad (1)$$

3.2 Inverse Document Frequency (IDF)

Term frequency has a significant problem, all terms are considered equally important when it comes to assessing relevance on a query. In reality, certain terms have little power in characterizing the document. For example, the stop words that is meaningless terms such as articles, prepositions, pronouns etc.

$$idf(t,D) = \log \frac{N}{1+\{d \in D : t \in d\}} \quad (2)$$

Where,

|D| : Total number of documents in the corpus,

| d ∈ D : t ∈ d | : Number of documents in which the term appears i.etf(t,d) ≠ 0.

Suppose, if the term is not present in the collection, it will create a division by zero. As a result change the formula to

$$1 + | d \in D : t \in d |$$

Then tf–idf is calculated as:

$$tfidf(t,d,D) = tf(t,d) \cdot idf(t,d) \quad (3)$$

In other words, tfidf(t,d,D) assigns to term t, a weight in document d.

Relevancy score [5]can be calculated by adding tf-idf weight of every term in, instead of adding the number of occurrences of every query term t in d.

$$score(q,d) = \sum_{t \in q} tfidf(t,d,D) \quad (4)$$

Select top N documents with highest score.

4. PAGE RANK

Page Rank is a link analysis algorithm [17] and it assigns some weight to each element of a hyperlinked set of documents. The page rank measures its relative importance within the set. As the number of incoming links of a document is greater the page rank value of that document will also be greater. The numerical weight that it assigns to any given element *E* is referred to as the *PageRank of E* and denoted by

$$PR(E)$$

In the general case, the PageRank value for any page **u** can be expressed as:

$$PR(u) = \sum_{u \in B_u} \frac{PR(v)}{L(v)} \quad (5)$$

i.e. the Page Rank value for a page **u** is dependent on the Page Rank values for each page **v** out of the set **Bu** (this set contains all pages linking to page **u**),divided[17] by the number $L(v)$ of links from page **v**.

In the proposed idea, initialize the page rank with the score(q,d).

$$PR(d) = Score(q,d) \quad (6)$$

Then PageRank computed iteratively

$$PR(u)_{iterative} = \sum_{u \in B_u} \frac{PR(v)}{L(v)} \quad (7)$$

The rank value indicates an importance of a particular page. The page rank algorithm improves the relativity accuracy of the link because a hyperlink[17] to a page counts as a vote of support. The PageRank of a page is defined recursively and depends on the number of incoming links. A page that is linked to by many pages with high PageRank receives a high rank itself.

This idea can be improved by adding score value with the page rank value.

$$Rel\_Score(d) = PR(d)_{iterative} + Score(q,d) \quad (8)$$

5. SYSTEM ARCHITECTURE

Fig. 1 shows the system architecture where depending on the input keyword or query, related documents get download from the Internet. Then document relevancy is get calculated using TF-IDF and work of focused crawler starts by extracting links, finding out score of link and pass it into the page rank algorithm for calculation of more accurate relevancy scoreFig. 1 shows the system architecture where depending on theinput keyword or query, related documents get download from the internet. Then document relevancy is get calculated usin0g TF-IDF and work of focused crawler starts by extracting links, finding out score of link and pass it into the page rank algorithm for calculation of more accurate relevancy score.

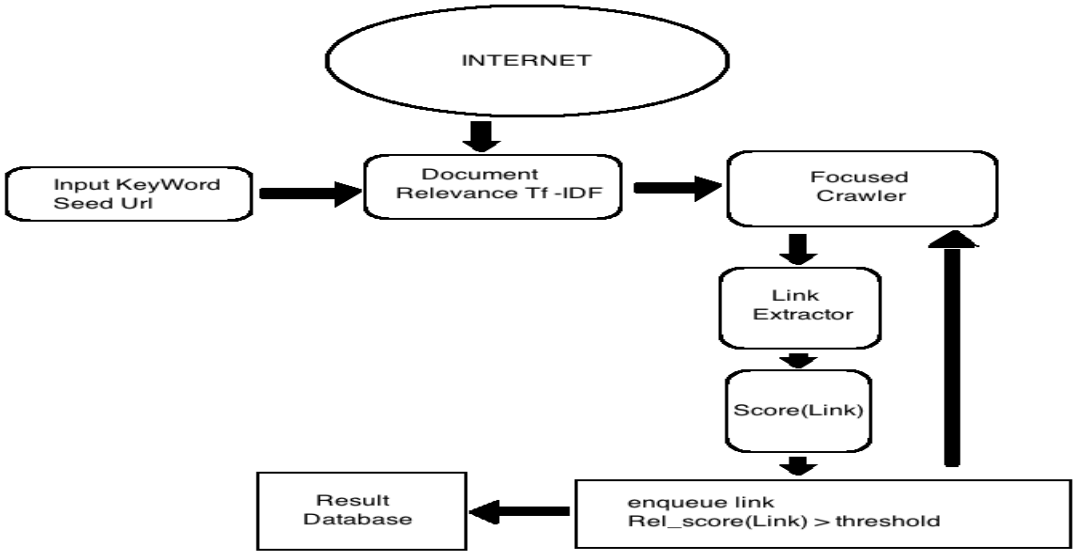

Fig 1: System Architecture

If link Rel_score exceeds threshold value then include into result database. In this way top k documents with highest score will be displayed.

Based on seed pages we can analyze the category of unvisited link. It means, the unvisited URL might relevant to the topics or not. The relevant link can be extracted on the basis of following attributes.

a) Anchor Text Relevancy (ATR)

Anchor text is the characters and words that hyperlink display when linking to another document. The anchor text relevancy is the relevancy between topic keywords and anchor text. The words of anchor text are matched with topic 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. This is possible because anchor text describes the some information about URL.

b) Cohesive Text Relevancy (CTR)

Cohesive text is the text around the hyperlink. Cohesive text also may have some information about the relevancy of topic.

Cohesive relevancy score of URL is the score of URL with respect to topics in sentence. For the extraction of cohesivetext, one sentence or group of meaningful sentences just about the anchor link[5] has to be considered. A sentence can be recognized as starting with a capital letter and ends with a period (dot). The following algorithm [5] describes steps:

1. Identify the anchor link in the page.
2. Extract a sentence in backword 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 backward direction of anchor tag if any

6. SYSTEM ALGORITHM

1. Insert seeds URL in to the ready Queue
2. If more links in ready queue then
3. Extract relevant links
4. Fetch document
5. Calculate score
6. Apply page rank to find rel_score
7. Save (link, Rel_score)
8. Else
9. Sort unprocessed queue

7. CONCLUSION

Generic crawlers and search engines crawl all the link in a document without measuring link relevance, it increases the number of resources required by crawler e.g. storage size, time etc. Best first search crawler calculate link relevancy and using page rank algorithm improve the relativity accuracy of the link because a hyperlink to a page counts as a vote of support, if it relevant to given query then it store link in processing queue and remove irrelevant link. It saves memory space as well as time and also finds more relevant documents. In future the idea can be improved so that the relevant links accessible from irrelevant can also be fetched making the search result more relevant.

8. ACKNOWLEDGMENTS

We would like to thank our Asst. Professor Rajkumar Singh Rathore and Asst. Professor SachinKathuriawho have contributed towards development of this research paper named "Focused crawler based on Efficient Page Rank Algorithm".We would also like to show our gratitude to Professor BhawnaMallick, H.O.D, Galgotias College for sharing their pearls of wisdom with us during the course of this research.

9. REFERENCES

[1] Bing Liu, "Web Content Mining" the 14th international world wide web conference

[2] De Bra, P., Houben, G., Kornatzky, Y., Post, R. ``Information retrieval in distributed hypertexts". Proc. 4th RIAO Conference, 1994.

[3] 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.

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

[5] SunitaRawat, D.R. Patil Department of Computer Science and Engineering, 2013 3rd IEEE International Advance Computing Conference (IACC).

[6] A. McCallum, K. Nigam, J. Rennie, and K. Seymore, "Building domainspecic search engines with machine learning techniques," in Proc. AAAI Spring Symposium on Intelligent Agents in Cyberspace, 1999.

[7] A. K. McCallum, K. Nigam, J. Rennie, and K. Seymore, "Automating the construction of internet por- tals with machine learning," To appear in Information Retrieval.

[8] M. Gori, M. Maggini, and F. Scarselli, "http://nautilus.dii.unisi.it."

[9] Menczer F., Pant G. and Srivasan, P. "Topical Web Crawler: Evaluating Adaptive Algorithms" ACM Transaction on internet Technology (TOIT). Nov. 2014.

[10] S. Chakrabarti, B. Dom, P. Raghavan, S. Rajagopalan, D. Gibson, and J. Kleinberg, "Automatic resource compilation by analyzing hyperlink structure and associated text," in Proc. 7th World Wide Web Conference, Brisbane, Australia, 1998

[11] K. Bharat and M. Henzinger, "Improved algorithms for topic distillation in hyperlinked environments," in Proceedings 21st Int'l ACM SIGIR Conference., 1998.

[12] McCown, F. and Nelson, M. "Agreeing to Disagree: Search Engines and their Public Interfaces". ACM IEEE Joint Conference on Digital Libraries (JCDL 2007). Vancouver, British Columbia, Canada. pp. 309318. June 17-23, 2007.

[13] Bao, S., Li, R., Yu, Y. and Cao, Y. "Competitor Mining with the Web Knowledge". IEEE Transactions on Data Engineering, Volume: 20, Issue: 10, pp. 1297-1310, Oct. 2008.

[14] J. Kleinberg, "Authoritative sources in a hyperlinked environment." Report RJ 10076, IBM, May 1997.

[15] Zhang, T. Zhou, Z.Yu and D.Chen, "URL rule based focusedcrawlers", IEEE International Conference on e-Business Engineering, 2008.

[16] TfIdf weighting from http://nlp.stanford.edu/IRbook/html/htmledition/tf-idf-weighting-1.html

[17] Page Rank form Wikipedia, the free encyclopedia http://en.wikipedia.org/wiki/PageRank/