Concept based Focused Crawling using Ontology

S.Thenmalar
Anna University
Chennai, India

T. V. Geetha
Anna University
Chennai, India

ABSTRACT
The constraint of a web crawler that downloads only relevant pages is still a major challenge in the field of information retrieval systems. Rather than visiting all the web pages, a focused crawler visits only the section of the web that contains relevant pages, and at the same time, tries to skip irrelevant sections. Existing ontology based web crawlers estimate the semantic content of the URL based on a domain dependent ontology, which in turn supports the methods used for prioritizing the URL queue. The crawler maintains a queue of URLs it has seen during the crawl at each level, and then selects from this queue, the next URL to visit based on the conceptual rank of the page at that level obtained from domain ontology. However in this work we represent the topic as an overall conceptual vector, obtained by combining concept vectors of individual pages associated with seed URLs. The conceptual rank is based on comparison between conceptual vectors at each depth, across depths and between the overall topics indicating seed concept vector.

General Terms
Data and Web Mining.

Keywords
Focused crawler, Ontology, Conceptual vector.

1. INTRODUCTION
In an effort to keep up with the incredible growth of the World Wide Web (WWW), many research projects are targeted on how to retrieve and categorize information in a way, that will make it easier for the end users to find the information they want efficiently and accurately. A Web Crawler explores number of Web Servers to find information about a particular topic. However, exploring all the Web Servers and the pages, are not realistic given the growth of the Web and their refresh rates. The focused crawler of a search engine aims to traverse and selectively search for pages that are relevant to a predefined set of topics, rather than to exploit all regions of the Web [4]. It also aims to identify the capable links that lead to target documents, and avoid off-topic searches. Focused Crawlers support decentralization of the crawling process, with a more scalable approach. The existing focused crawlers predict the probability of a document's relevance to the search topic employing probabilistic models, or rules. Due to the fast extension of the Web and the essentially limited resources in a search engine, no single search engine is able to index more than one-third of the entire Web. In order to increase better coverage various approaches have been introduced, such as the development of meta-search engines, special-purpose search engines and so on [2,3]. Special purpose search engines are constructed and optimized in accordance with domain knowledge.

Focused crawling technique enables a search engine to work efficiently within a topically limited document space. The basic process of running a focused crawler is as follows. The crawler begins with a relatively large number of seed pages, which are topic relevant. Whenever it fetches a Web page, the unvisited URLs are extracted from that page and scored by their relevance to the topics. The crawler then picks up the URL with the highest score to crawl. One of the major problems of the focused crawler is the assignment of a proper order to the unvisited pages that the crawler will visit later. Many methods, based on exploiting the structure of linkage information on the Web, have been proposed to predict the importance of the documents. Domain-specific information is used to rank the importance of a web page and to guide the crawlers search through the hyperlinks [1].

In this work we assume that the concepts of all the seed pages together will convey the essence of the topic that is to be crawled. An ontology is used to obtain concepts associated with seed pages and with pages that are linked from seed pages acting as initial starting nodes. In general, the crawler keeps a queue of URLs it has seen during the crawl at each level, and then selects from this queue, the next URL to visit. In this paper, the selection is based on the conceptual rank of the page at that level. This rank is based on ranks obtained by conceptual matching between conceptual vectors of all web pages at each level. The rank of the source documents is based on, - similarity between the page and its source documents, similarity between the page to its source overall concept vector along with the similarity between concept vectors of pages at that level.

The reminder of this paper is organized as follows: section 2 reviews related work in the area. Section 3 is the proposed architecture and the algorithm. Section 4 describes the evaluation and Section 5 describes the conclusion.

2. RELATED WORK
Crawling through the web and adding web pages to the database, which are related to a specific domain and discarding web pages which are not related to the domain [5]. The rank or relevancy score of the URL is calculated based on the division score with respect to topic keywords available in a division i.e., finding out how many topic keywords are present in a division in which this particular URL exists and calculating the total relevancy of parent page of the relevancy score of the URL page [6]. However in our paper, the relevancy score of the URL page is calculated with respect to the similarity of the individual source document and similarity between the corresponding documents at its depth.

In another approach, focused crawler uses link structure of documents as well as keyword based similarity of pages to the

topic in order to crawl the web [7]. Ontologies have been used to improve the effectiveness of focused crawling. Ontology based web crawler [8] estimates the semantic content of the link of the URL in a given set of documents based on the domain dependent ontology, which in turn strengthens the metric that is used for prioritizing the URL queue. The link representing concepts in the ontology knowledge path is given higher priority. However in our work, the content of the page based on the concepts is also used for determining the relevancy of the page. The maximal set of relevant and quality page is to be retrieved [11]. The unvisited URL is classified using Naïve Bayesian classification based on the visited URLs attribute score. However in our work, the page relevance is calculated by the similarity calculation of page at each level as well as to its corresponding source page. Ontology based focused crawling determines concepts from the ontology and generates query [9]. In this work, we assume that the seed documents have already been provided for crawling and we use ontology to build a conceptual vector to represent the pages. In another approach, page relevancy computation is done based on ontology [10]. The page relevancy is computed based on the taxonomic relations. However in our paper, we consider the topic to be represented by a set of concepts present in the seed URLs. The overall combined concept vector of the seed URLs is compared with conceptual vectors at each depth, across depths, to determine relevancy.

3. PROPOSED ARCHITECTURE

The proposed architecture is shown in Fig 1. In the work described in this paper, the seed URLs are used to extract the seed documents. The seed documents are preprocessed and the frequently occurring words are filtered. The frequently occurring word is then mapped with the ontology and the concepts are extracted. Each seed document can be represented by concept vector. The seed documents are ranked based on the overall combined seed concept vector. The seed documents are ranked by dividing the number of individual seed document concept vector to the overall concept vector of all the seed documents. Links are extracted from the seed documents. The page corresponding to the links are downloaded. The page relevancy is calculated as given in the proposed algorithm. Extracting links, downloading the pages, calculating the page relevancy will be processed for various levels of crawling. The page relevancy is based on the similarity between page p at level i and its source document at level i-1, rank of source document at level i-1, similarity between page p at level i and other individual pages linked from same source document at level i, similarity based on page p at level i and overall seed concept vector. Similarity calculation is done by cosine similarity between pages. Relevance calculation is done by adding the rank of the source document, similarity calculation of page at each level and the similarity calculation of page to its corresponding source page. The pages are filtered whose relevancy is below threshold.

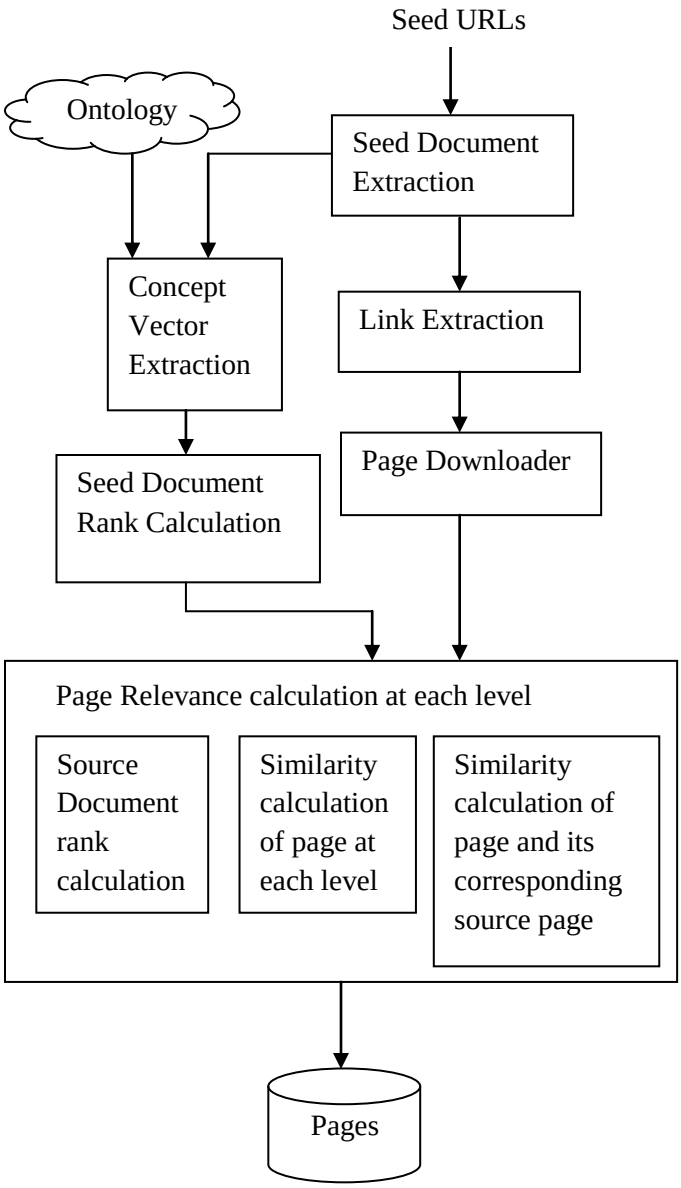

Fig 1: Concept based focused crawler

3.1 Proposed Algorithm

Step 1: Given Seed URLs and the seed documents are extracted

Step 2: Finding the rank of seed documents

- Preprocessing each seed document (root words identification)
- Finding out the most frequent words
- Generating concept vectors for each seed documents using ontology
- Ranking those seed documents based on the combined concept vectors of all the seed documents

The resultant would be rank based on overall combined vectors of all seed documents.

Step 3: For each first level link documents from seed documents

- Similarity between the first level link document and its source document
- Rank of the source seed document
- Similarity between the first level link document and its corresponding other linked documents.
- Similarity between first level link document and its combined vectors of all seed documents.
- Calculating step 3 for all the first level linked documents

Step 4: For each second level link document from the first level link document

- Similarity between the second level link document and its corresponding first level source document
- Rank of the source document at first link level
- Similarity between the second level link document and its corresponding other linked documents.
- Similarity between second level link document and its combined vectors of all seed documents.
- Calculating step 4 for all the second level linked documents

Step 5: Iterating step 4 for each level of crawling process

Step 6: Discarding those pages with minimum page score at its depth.

4. EVALUATION

The proposed algorithm is implemented with the tourism domain ontology and evaluated. One of the metric used to evaluate focused crawling is the harvest ratio, which is rate at which relevant pages are acquired and irrelevant pages are effectively filtered off from the crawl. Precision and recall are the standard indicators of relevancy and coverage. The harvest ratio represents the fraction of web pages crawled that satisfy the crawling target among the crawled pages. Evaluation has been done by comparing baseline focused crawling and the concept based focused crawling. The crawling starts with the given seed URLs, at each depth concept based crawler retrieves concept based documents.

The comparison of baseline focused crawler and the concept based focused crawler is shown in figure 2. This shows that the harvest ratio with respect to the relevancy for concept based focused crawler is better than the baseline focused crawler.

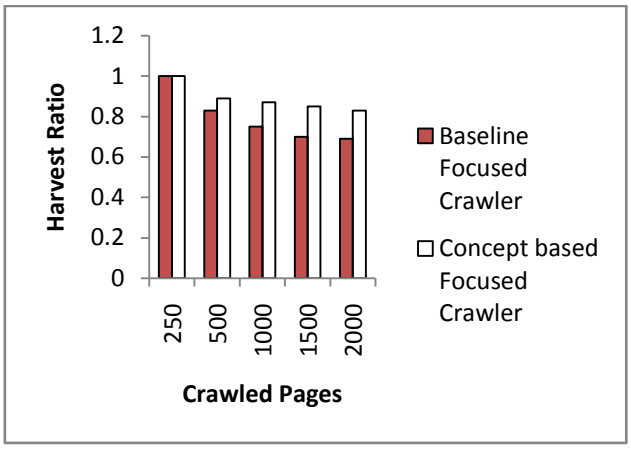

Fig. 1. Comparison between baseline focused crawler and Ontology based focused crawler.

This improvement in harvest ratio is basically because in baseline focused crawling there is a tendency to diverge from the original topic as we traverse the depth of crawling. However in concept based focused crawling, during crawling we select documents by comparing those associated documents at same depth, and in particular comparing them with the original overall conceptual vector obtained from the seed documents. This use of all overall conceptual vector throughout the process of crawling ensures that documents with diverging concepts are less likely to be selected.

5. CONCLUSION

This algorithm has the capability to solve the major problem of crawling relevant pages. This makes use of the combined seed concept vector for its ranking of each document. The rank of each document is based on the similarity between the page and its source document at each level, rank of source document corresponding to the previous level, similarity with its other documents at each level and similarity between the pages with the combined seed concept vectors. Thus the topic is considered as an overall conceptual vector, obtained by combining concept vectors of individual pages associated with seed URLs. The conceptual rank is based on comparison between conceptual vectors at each depth, across depths and between the overall topics indicating concept vector.

6. REFERENCES

[1] X.Zhang, T.Zhou, Z.Yu and D.Chen, 2008 URL Rule based Focused Crawlers, IEEE International Conference on e-Buisness Engineering, pp.147-154.

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

[3] Y. Zhang, C. Yin and F. Yuan, 2007. An Application of Improved PageRank in Focused Crawler, Fourth International Conference on Fuzzy Systems and Knowledge Discovery (FSKD 2007IEEE), Volume 2, pp.331-335.

[4] Q. Cheng, W. Beizhan and W. Pianpian, 2008. Efficient focused crawling strategy using combination of link structure and content similarity, IEEE International Symposium on IT in Medicine and Education, pp. 1045-1048.

[5] Debajyoti Mukhopadhyay, Arup Biswas, Sukanta Sinha, 2007. A new approach to design domain specific ontology based crawler, 10th International Conference on Information Technology, pp. 289-291.

[6] Debashis Hati, Amritesh kumar, 2010. An approach for identifying URLs based on Division score and link score in focused crawler, International journal of computer applications, Volume 2 – No.3.

[7] Mohen Jamali, Hassan Sayyadi, Babak Bagheri, Hariri and Hassan Abolhassani, 2006. A method of focused crawling using combination of link structure and content similarity, Proceedings of the International Conference on Web Intelligence.

[8] S. Ganesh, M. Jayaraj, V. Kalyan and G.Aghila, 2004. Ontology –based Web Crawler, Proceedings of the International Conference on Information Technology: Coding and Computing, Volume 2.

[9] Hiep Phuc Luong, Susan Gauch, Qiang Wang, 2009. Ontology-based Focused Crawling, International Conference on Information, Process, and Knowledge Management, pp. 123-128.

[10] Marc Ehrig, Alexander Maedche, 2003. Ontology-Focused Crawling of Web Documents, Proceedings of the symposium on Applied computing.

[11] Debashis Hati, Amritesh Kumar, Lizashree Mishra, 2010. Unvisited URL Relevancy Calculation in Focused Crawling Based on Naïve Bayesian Classification, International Journal of Computer Applications, Volume 3-No.9.