Journal of Web Engineering, Vol. 13, No.5&6 (2014) 525-538

THE MODIFIED CONCEPT BASED FOCUSED CRAWLING USING ONTOLOGY

S. THENMALAR
Anna University, Chennai
tsthensubu@gmail.com

T. V. GEETHA
Anna University, Chennai
tvgeedir@cs.annauniv.edu

Received December 12, 2013
Revised August 10, 2014

The major goal of focused crawlers is to crawl web pages that are relevant to a specific topic One of the important issues of focuses crawlers is the difficulty in determining which web pages are relevant to the desired topic. The ontology based web crawler uses domain ontology to estimate the semantic content of the URL and the relevancy of the URL is determined by the association metric. In concept based focused crawling a topic is represented by an overall concept vector, determined by combining concept vectors of individual pages associated with the seed URLs. The pages are ranked in comparison between concept vectors at each depth, across depths and between the overall topics indicating concept vector. However in this work, we determine and rank the seed page set from the seed URLs. We rank and filter the page sets at the succeeding depths of crawl. We propose a method to include relevant concepts from the ontology that have been missed out by the initial set of seed URLs. The performance of the proposed work is evaluated based on the two new evaluation metrics – convergence and density contour. The modified concept based focused crawling process produces the convergence value of 0.82 and with the inclusion of missing concepts produces the density contour value of 0.58.

Key words: Concept Vector, Focused Crawling, Information Retrieval, Ontology
Communicated by: G.J. Houben & E.-P. Lim

1 Introduction

The web crawler also called as robot or spider is the information gathering component of the search engine. Gathering useful web pages along with the associated interconnecting link structure in an automated methodical manner is called crawling. The pages associated with the seed URLs are retrieved; links in the pages are extracted and the process of page extraction is continued after prioritizing the pages until the required depth of crawl is reached. However, searching all the servers and the associated pages is unreasonable given the growth of the web and the frequency of their refresh rates and therefore no single search engine is able to index more than one-third of the entire web [19]. In general, crawler searches and collects pages on a specific set of topics that represent a relatively narrow segment of the web. In order to achieve better coverage, various approaches have been

introduced, such as the development of meta-search engines that take the results of search and optimize them in accordance with domain knowledge [16]. On the other hand, - a focused crawler of a search engine aims to traverse and selectively search for only pages that are relevant to a predefined set of topics, rather than consider all regions of the web [6]. Focused crawlers also aim to identify the appropriate links that lead to target documents avoiding off-topic pages [13].

Topic based focused crawler allow users to download and classify relevant pages from seed pages based on their topic similarity. The identification of the topic of the web pages is done automatically, where users provide a set of seed pages and at the end of the crawling process obtain topically similar pages. The challenge of identifying specific and relevant topic sub-spaces of the web is usually carried out by means of appropriate heuristics that direct the crawling process. Such a strategy basically involves determining how relevant a certain page is to a specific topic of interest. Most of the current strategies rely on text classifiers to determine such relevance [2][3], with the additional cost of having to train the classifiers. The focused crawler identifies the pages and their topical content before the web pages are fully downloaded and processed. Existing focused crawlers predict the probability of a document's relevance to the search topic employing probabilistic models or rules [12][1][29].

The major challenges in focused crawling are the scalability - catering up to billions of web pages; the coverage – new pages getting added all the time; the freshness – web pages getting updated over time and the detection of the off-topic pages. Another challenge lies in selecting links that are to be followed in the crawling process. In addition, the focused crawler has to decide the criterion of selecting the crawled pages that are to be downloaded. One criterion uses semantics to find the relevancy of pages to be downloaded. Such semantics can be provided by Ontologies [11]. In this paper, we assume that the concepts contained in the seed documents will together convey the topic of interest. This overall concept representation is used to rank the seed page set and also rank and filter page sets obtained by following hyperlinks at succeeding depths of crawl. We also include the concepts to the crawl that are left out by identifying the missed concepts from the ontology. We use specially designed evaluation metrics that convey the convergence of the focus of the documents and coverage of topic obtained from our crawling procedure.

The outline of this paper is as follows: Section 2 reviews the related work providing an overview of focused crawling. Section 3 describes the framework of modified concept based focused crawling with page set filtering. Section 4 describes the determination of including missing concepts and the process of modified concept based focused crawling with the inclusion of missing concepts. Section 5 provides an evaluation that makes a comparison with the baseline system. Section 6 discusses the conclusion.

2 Related Work

Focused crawler uses link structure of documents as well as keyword based similarity of pages to the topic in order to crawl the web. Each page is ranked according to the number of links to and from the seed, along with the content similarity between the page and the domain of interest. The page with the highest rank is crawled first and necessary link adjustment is made for the remaining pages [14]. In another approach, the relevancy score of the URL is calculated based on the weight of the topic keywords in the topic table obtained from pages corresponding to seed URLs and the weight of the

topic keywords in the web page table obtained from web documents corresponding to the hyperlinks. The division score with respect to topic keywords are those available in a division of HTML web page i.e., finding out how many topic keywords are present in a division in which this particular URL exists [12]. Both these scores are used to calculate the link score based on which the crawler decides the order in which links are to be crawled. In another approach of effective focused crawling based on content and link structure analysis, the relevance of a page is calculated with respect to the topic and uses metadata information (URL score, Anchor score, links from the relevant pages) to prioritize the extracted out links[26]. The context driven focused crawler [5] is based on the augmented hypertext document wherein the context of the keywords is stored in the form of table of contents. The table of contents and the category tree (predefined canonical topic taxonomy) provides the context of the keywords. The user agent, matcher agent, Dbase agent are responsible for the selection of keywords, retrieving the context and storing the downloaded web pages. The priority based focused web crawler [20] downloads relevant pages related to a particular topic and uses the priority queue denoting the similarity score of the URL. Every time when performing a delete operation, the queue will return maximum score web page to crawl.

The relevance of a document to the specific topic is determined by locality based document segmentation where a document is segmented into a set of sub-documents using contextual features around the hyperlinks [28]. The features are selected by the χ^2 measure and anchor text of a hyperlink along with the hyperlink descriptors in a parent node. Here Naive Bayes classifier is used to predict the most-likely class with the maximum posterior value for the features extracted from a new document when compared to the training examples. In another approach, focused crawling exploits not only content-related information but also genre information present in Web pages to guide the crawling process [2][3]. Genre-aware approach to focused crawling relies on the fact that some specific topics of interest can be represented by considering two distinct aspects: its genre (text style) and content-related information. The content is analysed for each of the visited page by determining the degree of similarity between the page and the predefined set of terms which describes the required topic of interest. The genre, content and the URL string based specific similarity scores are calculated for the current page and the sets of terms. These similarity scores are combined to a single score and compared to the given threshold. The page relevancy is judged for those single score which is greater than the threshold value. The priority of the non-visited pages is updated by changing new score to the URLs in Frontier that correspond to sibling pages of a given URL. This is done by checking the URL Type as the non-seed and comparing for the single score greater than the threshold. The crawlers are capable of learning the content of pages and also paths leading to relevant pages [4]. The Hidden Markov Model describes a novel learning crawler and provided an unbiased evaluation framework for a comparative analysis of their performance.

Ontology based focused crawling [22] adds the web pages to the database, which are related to specific domain and discard web pages which are not related to the domain. The relevancy of the page is calculated by the weight assigned to the concepts present in the ontology. Ontology based web crawler [9] 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 for prioritizing the URL queue. The relevancy of the links in that page is determined by the association metric. Ontology based focused crawling [21] determines concepts from the ontology and generate queries. These queries are given to different search engines and digital libraries. Support Vector Machine classifier is used to filter retrieve

documents that matches the query. The information extracted from the crawling process is used to enhance the domain ontology. In another approach, page relevancy computation is done based on ontology taxonomic relations [7]. The entropy based analysis mechanism for analysing the entropy of anchor texts and links to eliminate the redundancy of the hyperlinked structure [15]. The InfoDiscoverer is capable to crawl all pages, analyses the structure and extracts the informative structures and content blocks of the site. The CORE crawler [18] enhances the crawling process by making use of Ontology based Relevancy Score (based on the ontology terms that occurred in the context link of the web page) and Look Ahead Relevancy Score (based on the adaptive rules derived from the ontology terms and the link access)

An Ontology-Based Focused Crawler [17] identifies web pages that relate to a set of pre-defined topics and download them regardless of their web topology or connectivity with other popular pages on the web. The ontology-based focused crawler requires the training examples for its subsequent web visit. The topical content of the web page is identified through topical ontology and then the relevance of the page is calculated by the probability with which every page belongs to an ontology topic. The topical content of the page is processed to identify the text nugget that is most semantically close. With the topic relevance values and the topic similar extracts for a large number of web pages the training examples are built so that the crawler can judge the page to be downloaded or not. In general, ontology based focused crawling utilizes domain specific concepts to evaluate page relevancy and also to filter the retrieved pages.

The concepts from a domain specific ontology have been used to construct an overall concept representation that conveys the topic of interest for the focused crawling [27]. This paper discusses the use of the overall concept representation at various stages of focused crawling [27]. The paper also discusses a methodology to improve the coverage of concepts in the ontology by including relevant concepts from the ontology that have been missed out by the initial set of seed URLs. In addition, this paper discusses two new evaluation measures – convergence and density contour for evaluating the focus cohesiveness of the crawling process.

3 Representation of the Focus of the Crawling Process

One of the assumptions made in [27] is that the ontological concepts associated with the terms present in the seed documents together will convey the essence of the topic to be crawled. In this paper, the set of seed documents is called seed page set. Each seed document is represented by a concept vector consisting of concepts along with its frequency within the document. Then the topic of the crawl is represented by an overall concept vector obtained by combining concept vector of individual documents associated with seed URLs based on number of documents in which each concept occurs along with the concept weight in the document. This overall concept vector representing the topic of crawl is used throughout in all depths of crawling for ranking and selecting appropriate documents.

The modified concept based focused crawling approach describes the page set filtering during the crawling stages and the inclusion of missing concepts. The major function of the crawling process includes the ranking of the overall concept vector representation and the rank of the page set.

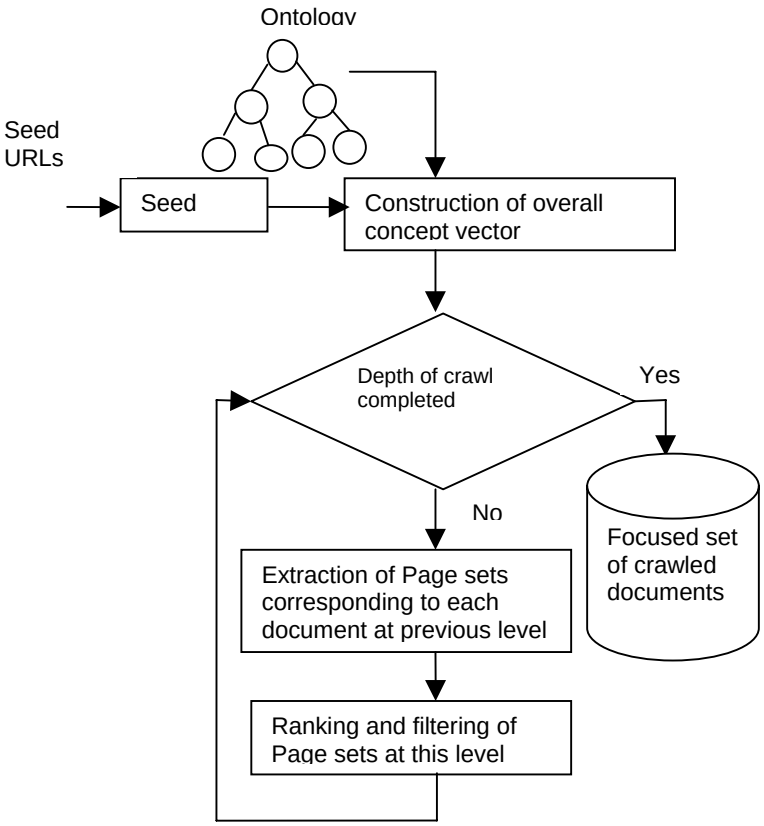

Figure 1. Process of Concept based Focused Crawling with page set filtering

3.1 Process of Modified Concept based Focused Crawling with Page set Filtering

Figure 1 depicts the overview of concept based focused crawling with page set filtering process. Initially, the crawling process begins with the set of seed URLs. The seed page set is obtained from the seed URLs. The seed page set is used to construct the overall seed page set concept vector [27]. The seed page set is constructed from the seed URLs, where the terms present in the seed page set is mapped to the ontology and the overall concept vector is constructed. The links from the seed page set is extracted and the associated documents are downloaded. The crawling process is continued through the extraction of appropriate page set that corresponds to the links available at the previous level. The rank of each page set is determined and selecting appropriate page sets is done at that level. The crawling process is continued for succeeding depths until the final set of crawled documents is obtained.

3.2 Ranking of Overall Concept Vector Representation

In general, in the context of focused crawling, the rank of a page is determined by the similarity between the terms associated with the topic of crawling and the frequencies of occurrence of these terms within the page [2][3]. However, in this work we carry out ranking at two stages – ranking of the documents in the seed page set and ranking of page sets at subsequent depths of crawling.

The ranking of the documents in the seed page set according [27] is based on the number of documents in which each concept occurs along with the concept weight in each document. The ranking of seed page set is given in Eq. (1)

$$\textit{Ranking} \text{ of seed page set} = \frac{\sum_{k=1}^{n} \text{Concept weight in the document}}{\sum_{j=1}^{m} \textit{Concept} \text{ weight of all the documents}} \tag{1}$$

Where, n denotes the number of concepts in the document and m denotes the number of concepts of all the documents. The concept weight is calculated by the concept frequency and its inverse document frequency.

The rank of a page set takes into account the importance of each seed page set where the page set is associated with documents obtaining through hyperlinks at each level originating from each seed document. As already described seed page set consists of all seed documents. Each page set at level 1 is obtained by considering all documents obtained through hyperlinks of each seed document.

Figure 2 shows that at subsequent levels, there exists a page set for every page set at level 1; consisting of all documents obtained through hyperlinks of all pages of preceding level page set. Page $set_{1,1}$ . . . Page $set_{m,1}$ are all obtained by crawling from source document d_1 of seed page set. In this context, the rank of a page set at level i where i varies from 1 to m, is indicative of how close the documents of that page set is to the seed page set, indicated by the overall seed concept vector. Therefore, the rank of the page set at level i is based on the following:

1. The rank of the originating page set at the preceding crawl depth. This factor indicates the importance of the seed document i from which page $set_{1,i}$, page $set_{2,i}$, . . . page $set_{m,i}$ all initially originate from, and the subsequent hyperlinks of that seed document.
2. The number of interlinks from the originating page set at level i-1. This factor takes into account the number of all pages of the originating page set at level i-1, where the number of out links is an indication of the connectivity of the documents.
3. The conceptual similarity between the overall concept vector of all pages in the page set and the overall seed page set concept vector. As we consider that the overall seed page set concept vector indicates the focus of crawl, the similarity between this overall concept vector and the overall concept vector of the page set is an importance of how close to the focus of crawl this page set is.
4. The number of intra-links between the documents in the page set indicates whether the documents within the page set are connected by hyperlinks which in turn indicate the connectivity between documents as decided by the authors of the documents. The rank of the page set starting from depth 1 is given in Eq. (2).

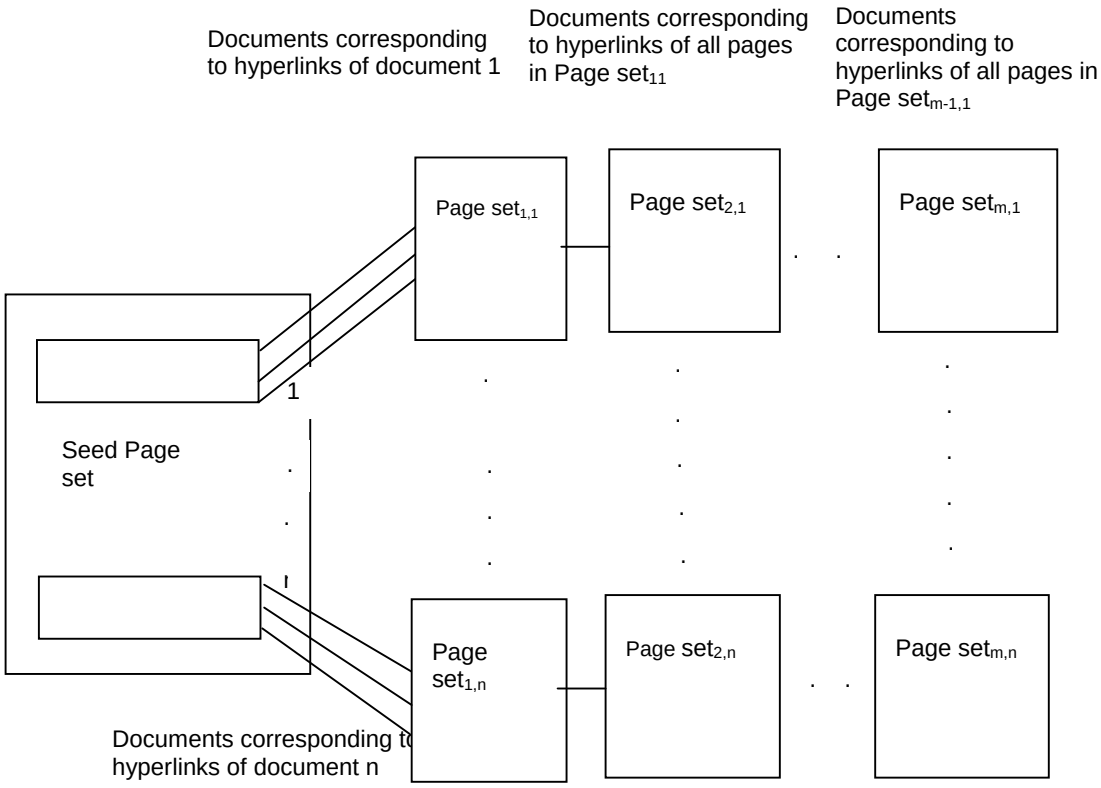

Figure 2. Page set at subsequent levels

$$\begin{aligned} \textit{Rank of Page set} = {} & \text{Rank of originating Page set} + \text{Number of Inter links from originating page set} \\ & + \text{conceptual similarity between overall concept vector of the page set and the overall} \\ & \text{seed page set concept vector} + \text{Number of Intra links within page} \end{aligned} \quad (2)$$

Conceptual Similarity is calculated by the similarity between the overall seed page set concept vector and the overall concept vector of the page set at its associated depth. The similarity here represents the cosine similarity between the vectors and is shown in Eq. (3)

$$\cos ine_{sim(Cs,\,Cd)} = \frac{C_s * C_d}{|C_s||C_d|} \quad (3)$$

where C_s denotes the overall seed page set concept vector and C_d denotes the overall concept vector of the page set at its depth.

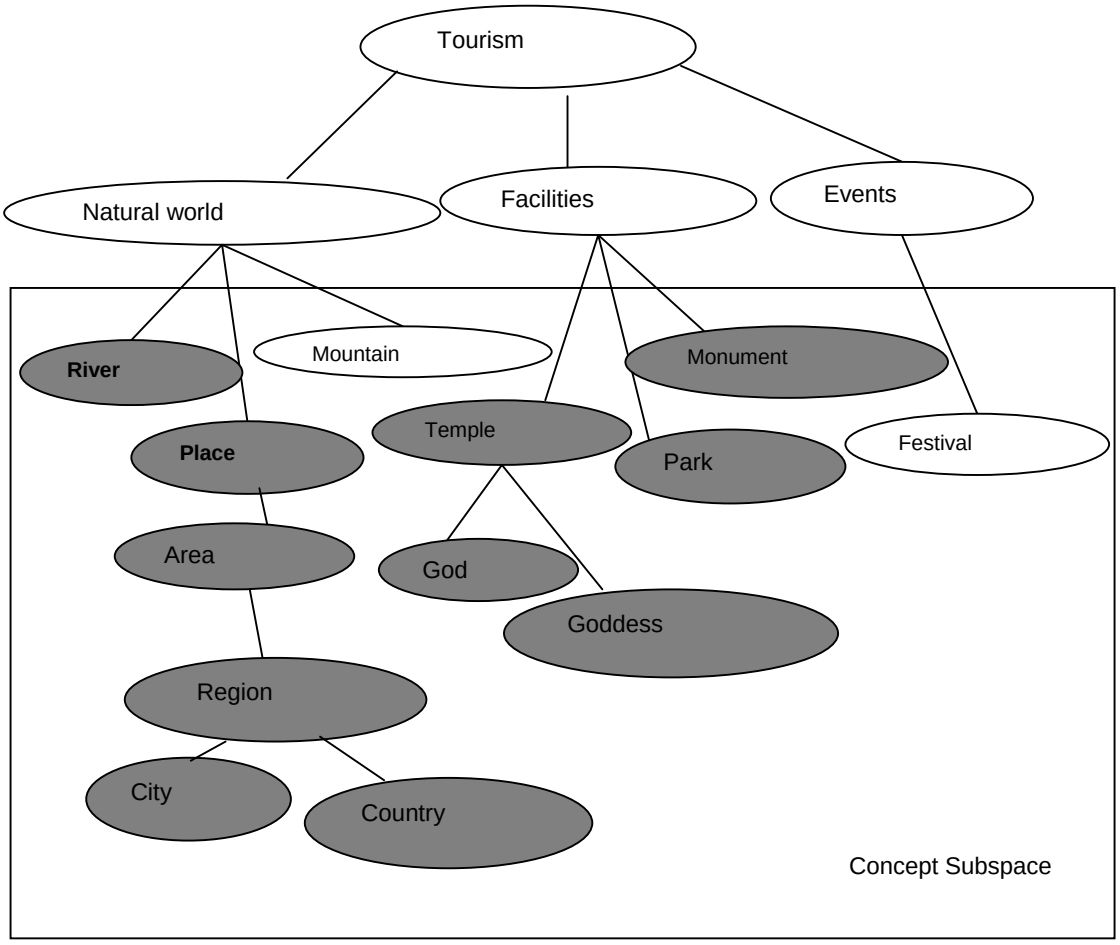

Figure 3. Concept Subspace defined in the structure of an ontology

Based on the association between terms in the URL and concepts in the ontology, corresponding pages have been included in the crawl [9]. However, in this paper, we filter out page sets whose rank is below the average page set rank of all page sets at that level, based on conceptual similarity with overall seed page set concept vector.

4 Use of Ontology in Inclusion of Missing Concepts

In the absence of an ontology, the academic papers belonging to a topic area that are missing from a collection of digital libraries are identified by crawling with document belonging to author's home pages as seed [30]. Identification of seed URLs is carried out by first taking a keyword conveying the topic. The keyword is given to three different search engine, and the common result pages (seed pages) obtained are used to extract other keywords belonging to the topic and build a topic keyword table.

This topic keyword table is used for determining relevancy during the crawling process [12]. Identifying concepts of a topic of interest is carried out in a number of ways.

From the above two approaches, the importance of obtaining as many keywords as possible related to the topic of crawl. When we consider documents belonging to a topic, where a topic is represented by concepts in an ontology, it is quite possible that not all concepts within a subtree or subarea of the ontology are included. The overall seed page set concept vector representing the topic defines a concept subspace of the domain ontology of the topic as shown in Figure 3.

However, from the focused crawling viewpoint there may be some concepts in the subspace that are nearby and closely related to the concepts in the overall seed page set concept vector but are not present in the concept subspace. Using these conceptual missing concepts, to obtain new web pages which can be used as additional seed URLs will increase the density of documents obtained from the concept subspace of the focus crawl.

In this paper, we discuss how the domain ontological structure is used to identify these missing concepts. The missing concepts are identified based on:

1. The number of concepts in the overall seed page set concept vector and / or overall concept vector of page set in subsequent depths of crawling to which the missing has a direct parent / sibling / child relation. Here the relation to concepts in the overall seed page set concept vector is given a higher priority than other overall concept vector of page set.
2. From the concepts of (1) as shown in Figure 4-these concepts which have maximum frequency in the final crawled document set and those contained in maximum number of documents are given higher priority.

Based on the above two factors the missing concepts are selected and used for obtaining new seed URLs and the whole crawling process is repeated with the new overall seed page set concept vector.

4.1 Process of Modified Concept based Focused Crawling with the Inclusion of Missing Concepts

From the seed URLs the seed page set is constructed, where the terms present in the seed page set is mapped to the ontology and the overall seed page set concept vector is constructed. The overall seed page set concept vector, the domain ontology and the final crawled page set is used to discover the missing concepts. The identification of the missed concept is given as the query to the existing search engine. The top ten results obtained from the search engine is compared with the overall seed page set concept vector and the best pages are selected. The selected pages are added to the seed page set and yields the modified seed page set. The overall seed page set concept vector is constructed for the modified seed page set. The links from the modified seed page set is extracted and the associated documents are downloaded. The crawling process is continued until all the concepts are identified in the concept subspace with respect to the modified seed page set concept vector.

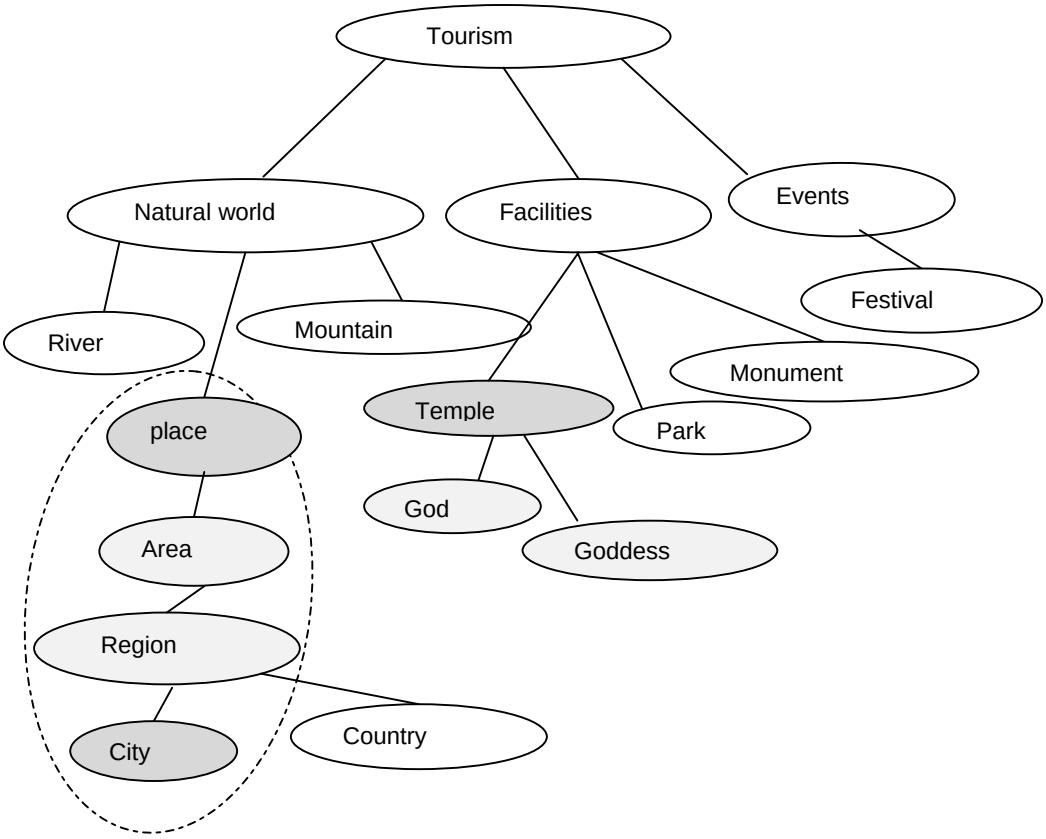

Figure 4. Inclusion of missing concepts

5 Evaluation

The crawling process is implemented with the tourism domain ontology and the crawling process is initially carried out with 100 seed URLs and crawled at the depth of five, finally we obtain 10,500 crawled documents. The existing performance measure to evaluate focused crawling is the harvest ratio which represents the fraction of web pages that are relevant to the topic crawl among the crawled pages. However, in this paper, we define two new evaluation parameters – convergence and density contour.

5.1 Convergence

Convergence has been defined in the context of clustering to evaluate the purity of cluster. The purity of clusters is defined as the ratio of number of samples in the document category to the total number of samples [10]. In this context, in focused crawling however we define convergence as the similarity between the final crawled document set and the initial set of seed documents. Convergence is calculated as shown in Eq. (4)

$$Convergence = \frac{\text{No. of concepts that are present at the end of the crawl and is present in the initial seed page set}}{\text{No. of concepts that are present at the end of the crawl data set}} \quad (4)$$

In another words convergence measures the number of concepts in final set of crawled documents that are also present in the initial seed page set. In order to evaluate the performance of our system we compare our system with a baseline system – Apache Nutch [24]. Nutch [25] is a web crawler software product that can be used to aggregate data from the web. Apache Nutch crawls a page in multiple step processes such as inject, generate, fetch, parse, update and index. Each process runs through MapReduce algorithm. MapReduce uses some data as input to run each of these processes and generates different data. For example crawldb, linkdb are the generated data by inject and update process. The crawling strategy involves the breadth-first strategy and configure via online page importance scoring [23]. We have downloaded and installed Apache Nutch 1.2 and given the same set of seed URLs. Concept based focused crawling with page set filtering and inclusion of missing concepts is compared with the baseline system and the concept based focused crawling without page set filtering [27].

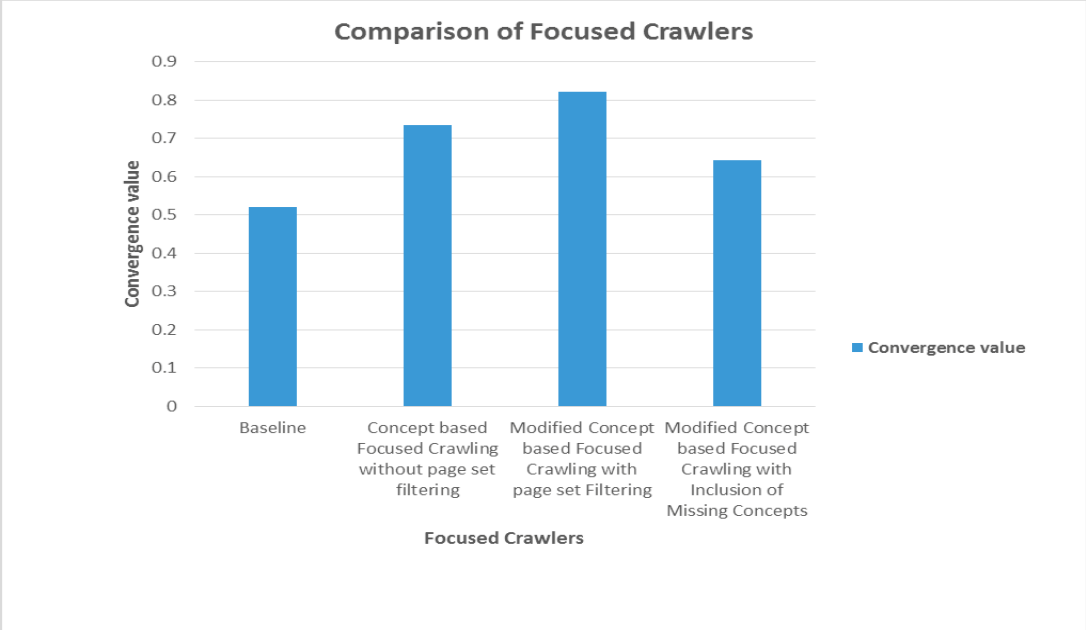

Figure 5. Comparison of Focused Crawlers in terms of convergence

From the result Figure 5, we see that the concept based focused crawling with page set filtering gives the best convergence, since at every stage of the crawling only the best page sets (conceptual similar to the overall seed page set concept vector) are considered. It is also seen from

figure 5, that concept based focused crawling with page set filtering but with the inclusion of missing concepts gives a lower value of convergence since in this case we include concepts not obtained in the overall seed page set concept vector.

5.2 Density Contour

Another parameter that we have defined to evaluate focused crawling is the density contour which essentially evaluates the coverage of the concepts within the ontological concept subspace specified by the topic of crawl. Metrics to evaluate the structural dimension of ontologies was proposed [8]. These measures include absolute depth, which measures the cardinality of each path from a given set of paths, average depth, which gives the average cardinality of all paths and the similarly maximal depth. Similar measures have also been defined for the measuring cardinality of the breadth of an ontology. They also define measure of density which indicates the presence of clusters of classes of non-taxonomic relations [8]. It is from the definition of depth, breadth and the density that we have derived our evaluation measure – density contour.

In our work density contour is based on the number of concepts covered in the defined concepts subspace and the number of documents associated with these concepts. The density contour is calculated by Eq. (5)

$$Density\ \text{Contour} = \text{maximum depth} * \text{maximum breadth of the connected concepts} * \frac{2e}{|v||v-1|} \quad (5)$$

Where e denotes the number of connected edges in the ontology structure, v denotes the number of concepts in the ontology structure. This density contour gives the density of concepts along with the number of documents associated with these concepts in the concept subspace.

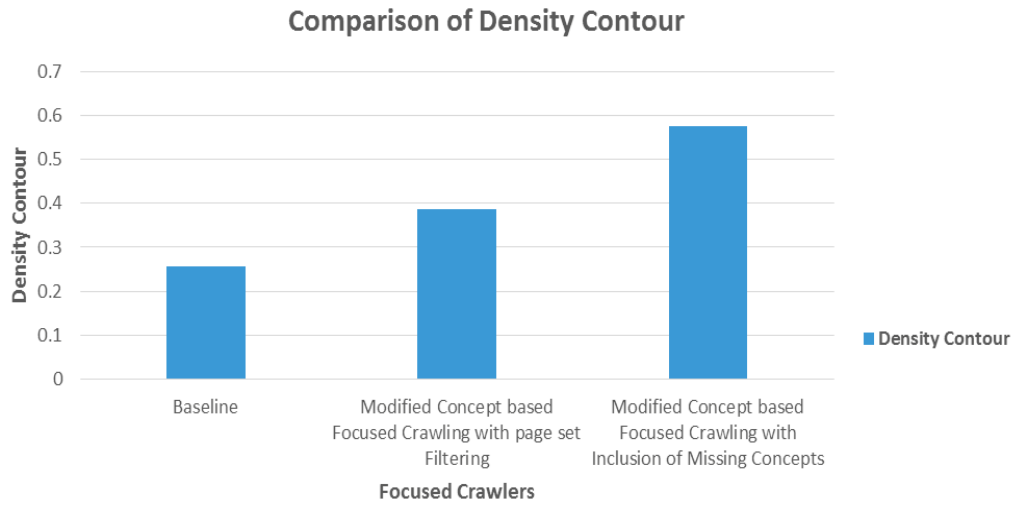

Figure 6. Comparison of Focused Crawlers in terms of density contour

Figure 6 shows the comparison of density contour for baseline system, concept based focused crawling with page set filtering and the concept based focused crawling with inclusion of missing concepts. In baseline system, the crawling was based on hyperlinks and concepts associated with the ontology which was not considered for page selection. Therefore, the performance of baseline method shows poor value of density contour. When considering the performance of concept based focused crawling with page set filtering, the method tries to select pages for crawling based on the concept similarity with overall seed page set concept vector. Therefore, if the overall seed page set concept vector had a good density contour associated with it, then the final set of crawled pages would also exhibit good density contour. In case of the concept based focused crawling with inclusion of missing concepts, the method tries to not only find pages with respect to the overall seed page set concept vector, but also includes missing concepts in the concept subspace associated with the overall seed page set concept vector. Therefore this method shows good coverage of concepts in the concept subspace.

6 Conclusions

The major criterion of selecting the crawled pages has been discussed in this paper. The overall seed page set concept vector describes the topic of crawl. This overall seed page set concept vector is used to rank the seed documents present in the seed page set, and also rank and filter the page sets at the succeeding depths of crawl. The methodology to include relevant concepts from the ontology that have been missed out by the initial set of seed URLs is described in this paper. The cohesiveness of the crawling process is evaluated by the convergence and the coverage metrics. The concept based focused crawling with page set filtering yields best page sets and concept based focused crawling with the inclusion of missing concepts produces a better coverage of concepts.

References

1. Altingoyde, I. S., and Ozgur U., Exploiting Interclass Rules for Focused Crawling. Journal of IEEE Intelligent Systems, 19, 2004, 66-73.
2. Assis G. T. D., Laender A.H. F, Goncalves M. A. and Silva A. S. D., Exploiting Genre in Focused Crawling, Proceedings of 14th International Conference on String processing and information retrieval, 2007, 62-73.
3. Assis G. T. D., Laender A.H. F, Goncalves M. A., and Silva A. S. D., A Genre-Aware Approach to Focused Crawling, Journal of World Wide Web, 12, 2009, 285-319.
4. Batsakis S., Petrakis E. G.M., and Milios E. E., Improving the performance of focused web crawlers, Journal of Data Knowledge Engineering, 68, 2009, 1001-1013.
5. Chauhan, Naresh. and Sharma, A. K ., "Design of an agent based context driven focused crawler", International journal of Information Technology, 2008, 61-66.
6. Cheng Q., Beizhan W. and Pianpian W., Efficient focused crawling strategy using combination of link structure and content similarity, IEEE International Symposium on IT in Medicine and Education, pp. 1045-1048, 2008.
7. Ehrig M. and Maedche A., Ontology-Focused Crawling of Web Documents, Proceedings of ACM Symposium on Applied computing, 2003, 1174-1178.
8. Felix A. A., Taofiki A. A., and Adetokunbo S., On Algebraic Spectrum of Ontology Evaluation, International Journal of Advanced Computer Science and Applications, 2, 2011, 159-168.
9. Ganesh S., Jayaraj M., Kalyan V., Murthy, S. and Aghila, G., Ontology–based Web Crawler, Proceedings of International Conference on Information Technology: Coding and Computing, 2004, 337-341.
10. Ghosh J. and Strehl A., Similarity-Based Text Clustering: A Comparative Study, In Grouping

Multidimensional data, Berlin-Heidelberg:Springer, 2006, 73-97.
11. Goyal R.K., Gupta V., Sharma V. and Mittal P., Ontology based web retrieval. Proceedings of International Symposium of Computer Science and Technology, 2008, 141-144.
12. Hati D. and Kumar A., An approach for identifying URLs based on Division score and link score in focused crawler, International Journal of Computer Applications, 2, 2010, 48-53.
13. Hati D., Mishra L. and Kumar A., Unvisited URL Relevancy Calculation in Focused Crawling based on Naive Bayesian Classification, International Journal of Computer Applications, 3, 2010, 23-30.
14. Jamali M., Sayyadi H., Hariri B. B and Abolhassani H., A method of focused crawling using combination of link structure and content similarity, Proceedings of International Conference on Web Intelligence, 2006, 753-756.
15. Kao H. Y., Lin S. H., Ho J. M. and Chen M. S., Mining web Informative Structures and Contents based on Entropy Analysis, Journal of IEEE Transactions on Knowledge and Data Engineering, 16, 2004, 41-55.
16. Ke Y., Deng L., Ng W. and Lee D.L., Web dynamics and their ramifications for the development of web search engines, International Journal of Computer and Telecommunications Networking-Web dynamics, 50, 2006, 1430-1447.
17. Kozanidis, Lefteris, "An ontology based focused crawler", Proceedings of the 13th International Conference on Natural Language and Information Systems: Applications of Natural Language to Information Systems, NLDB '08,2008, 376—379.
18. Kumar, Muhesh. and Vig, Renu., "Design of CORE: Context Ontology Rule Enhanced Focused Web Crawler", International conference on Advances in Computing, Communication and Control, 2009, 494-497.
19. Lawrence S. and Giles C. L., "Searching the World Wide Web", Science Journal, 280, 1998, 98-100.
20. Lokhande, Kiran. P., Honale, Sonal. S. and Gangavane, H. N., "Web Crawler Using Priority Queue", International Journal of Research in Advent Technology, 2014.
21. Luong H. P., Gauch S. and Wang Q., Ontology-based Focused Crawling, International Conference on Information, Process, and Knowledge Management, 2009, 123-128.
22. Mukhopadhyay D., Biswas A. and Sinha S., A new approach to design domain specific ontology based crawler, 10th International Conference on Information Technology, 2007, 289-291.
23. Nioche, Julien., "Large Scale Crawling with Apache Nutch", ApacheCon Europe 2012.
24. Nutch, http://nutch.apache.org/.
25. Nutch Crawler, http://nutch.apache.org/downloads.html.
26. Pal, Anshika., Tomar, Deepak. Singh. and Shrivastava S.C., "Effective Focused Crawling Based on Content and Link Structure Analysis", International Journal of Computer Science and Information Security (IJCSIS), 2009.
27. Thenmalar S. and Geetha T. V., Concept based Focused crawling using Ontology, International Journal of Computer Applications, 26, 2011, 29-32.
28. Yang J., Kang J. and Choi J., A Focused Crawler with Document Segmentation, Proceedings of International Conference on Intelligent Data Engineering and Automated Learning, 2005, 94-101.
29. Yuvarani M., Iyengar N.Ch. S. N. and Kannan A., LSCrawler: A Framework for an Enhanced Focused Web Crawler based on Link Semantics, Proceedings of IEEE/WIC/ACM International Conference on Web Intelligence, 2006, 794-800.
30. Zhuang Z., Wagle R. and Giles C. L., What's there and what's not? Focused Crawling for Missing Documents in Digital Libraries, Proceedings of Joint Conference on digital libraries, 2005, 301-310.