Priority based Semantic Web Crawler

Jaytrilok Choudhary
Asst. Professor, Department of CSE
MANIT Bhopal(India)

Devshri Roy
Associate Professor, Department of CSE
MANIT Bhopal(India)

ABSTRACT

The Internet has billions of web pages and these web pages are attached to each other using URL(Uniform Resource Allocation). Web crawler is a main module of Search engine that gathers these documents from WWW. Most of the web pages present on Internet are active and changes periodically. Thus, Crawler is required to update these web pages to update database of search engine. In this paper, priority based semantic web crawling algorithm has been proposed. Ontology is used to get semantics of web page during crawling process. Algorithm starts with initial seed URL. The web page at given URL is downloaded from Internet and semantic score is calculated with given topic. The semantic score of unvisited URL is calculated using its Anchor text semantic similarity score, semantic similarity score of web page of unvisited URL with given topic and semantic score of its parent pages. Priority queue is used to store URL and its semantic score instead of simple queue. So, every time priority queue returns higher priority URL to crawl next. The overall performance gain over simple crawler is 88%, over focused crawling is 28% and priority based focused crawler is 6%.

Keywords

Priority, ontology, Semantic similarity, downloader, search engine.

1. INTRODUCTION

Web search engine is intended to find information which is associated to search query specified by user from Internet. It stores trillions of web documents and their links. These web documents are needed to be updated that makes it more reliable. It uses web crawler for this purpose. Web crawler is a incessant running program which downloads web pages at regular intervals from Internet. The downloaded pages are indexed and stored in a database as shown in Figure 1[1].

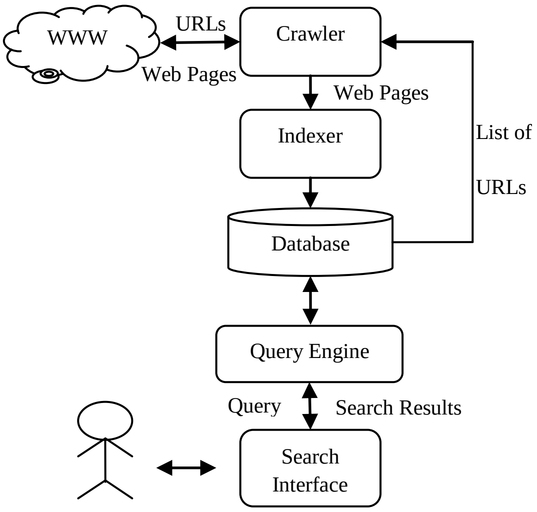

Fig 1: Basic architecture of web search engine

There are two basic web crawling strategy: breadth first crawling and best first crawling [2]. Breadth first crawling is also known as classical web crawler. It starts with initial seed URLs. It downloads web pages for seed links. Then new URLs extracted from the downloaded pages, add them into queue and choose URL one by one and repeat crawling process for specific count.

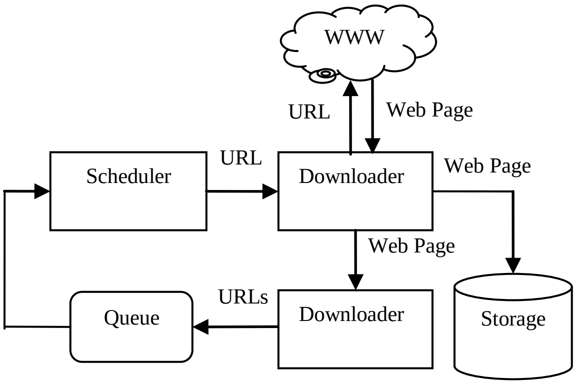

Fig 2: Architecture of Basic Crawler

The best first crawling downloads only relevant web pages of a particular given topic or string. A crawler that uses best first crawling approach is known as a focused crawler [2 and 11]. In other words Focused crawling is an extension of breadth first crawling where web pages related to particular topic are downloaded only.

Semantic web crawling is a special focused crawling where Semantic similarity is calculated between given topic and web pages. A web page and topic can be similar if they share conceptually similar terms but not necessarily lexical terms [3]. e.g. Topic “Computer” is conceptually related to terms “Software, Hardware and Operating System etc.” but not lexically. To find semantic similarity between topic and web page, topic ontology is used. In other words, cosine similarity is calculated between web page and all the concepts and relationship between concepts of topic ontology that is semantic similarity between web page and topic.

There are various ways to define "what ontology is?" from these two important are

From AI point of view, ontology is defined as “explicit specification of conceptualization”. *Conceptualization* is the abstract representation of a real world entity with the help of domain relevant concepts. Ontology should be *formal* so that it becomes machine understandable and it should have to enable *shared* communication across the communities [4].

From knowledge-based systems point of view, it is defined as “a theory (system) of concepts/vocabulary used as building blocks of an information processing system” [5].

In this paper, priority based semantic web crawling approach has been proposed. Priority queue is used to keep URLs with their semantic score. So, priority queue returns maximum

Score URL every time to crawl next. To find semantic score, ontology of given topic and Vector space model is used. The semantic score of unvisited URL is calculated using its Anchor text semantic similarity score, semantic similarity score of web page of unvisited URL with topic ontology and semantic score of its parent pages.

The reminder of this paper is prepared as follows: section 2 contains related work in this area. Section 3 is the architecture of priority based semantic web crawling. Section 4 contains the algorithm of priority based semantic web crawling. Section 5 describes the experimental results and Section 6 is the conclusion and future work.

2. PREVIOUS WORK

Singhal N. et. al.[1] have designed a incremental web crawler. The incremental crawler visits the internet periodically to update its database. Based upon updation of web documents, web documents are categorized and grouped as very frequently, frequently less frequently. The crawler visits a site frequently and the frequency of visits may be adjusted according to the category of the site. This architecture is more suitable for parallel web crawler.

S. Ganesh et. al.[6] have proposed an ontology based web crawler. In this approach, a new metric called association-metric has been proposed. The association-metric analyzes the semantic content of the URL based on the domain dependent ontology. After downloading the page, the association metric estimates the relevancy of the links in that page. Finally, reordering of URL is done based on relevancy of web page.

Mukhopadhyay D. et. al.[7] have proposed a domain specific web crawler which crawls domain specific Web pages from the World Wide Web(WWW). Crawler uses ontology of a domain for which web pages has to be crawl.

Chen X. et. al.[8] gave the methodology for focused crawling. They have focused on content of web page to improve page relevance and also used link structure to improve the coverage of a specific topic. They considered only two factor, content of web page and link structure, to get relevancy of web page.

Hati D. et. al.[9] have proposed an adaptive focused crawling based on link analysis. In this approach, they first calculate the score of unvisited URL based on its anchor text relevancy score, Relevancy score of its parent, its description in Google search engine and calculate the similarity score of description with topic key words. The major issue of this technique is URL queue optimization.

Thenmalar S. et. al.[10] have proposed an algorithm for focused crawling based on ontology. They are preparing topic as an overall conceptual vector that is obtained by combining concept vectors of individual pages associated with seed URLs. Here the role of ontology is to obtain concepts associated with seed page. The next URL to be crawl is based on the conceptual rank of the web page at that level which is obtained by conceptual matching between conceptual vectors of all web pages at each level.

Choudhary J. et. al.[11] have proposed a priority based focused crawler that start with initial seed URLs and focus word. Crawler first download web pages at initial seed URL then extracts all the URLs from downloaded web pages. Again it downloads all the web pages corresponding to extracted URLs. Then it finds relativity score between focus word and new downloaded web pages. It stores web pages and their relativity score in priority queue. The priority queue return maximum score web page from queue and repeat same process until queue empty or specified number of pages. The relativity score is the syntactic similarity not semantic similarity between focus word and web page.

3. PRIORITY BASED FOCUSED WEB CRAWLING

The crawling process begins with initial seed URL and topic.

1. Crawler downloads web page at given seed URL.
2. Now, it finds all the new URLs present in downloaded page.
3. Again, crawler downloads all the web pages corresponding to all new URLs.
4. Now, it calculates semantic score between given topic ontology and all downloaded web pages using its Anchor text semantic similarity score, semantic similarity score of web page of unvisited URL with topic ontology and semantic score of its parent pages.
5. It adds page and its semantic score into the priority queue and every time priority queue return maximum semantic score web page.
6. Now, repeat step 2- 5 for either specified number of pages or until queue is empty.

The overall crawling process is shown by figure 3.

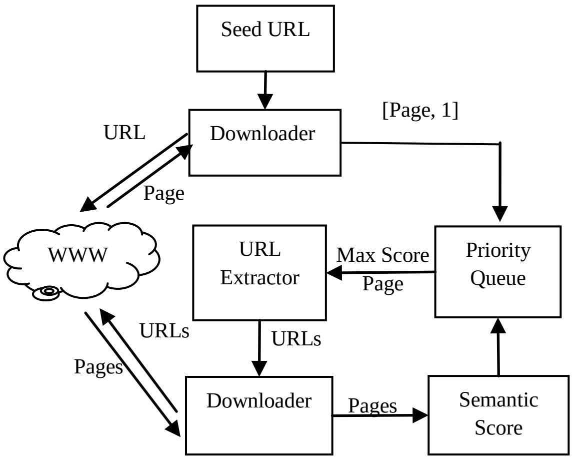

Fig 3: Architecture of Priority based focused

3.1 Similarity Score calculation

The vector space model is used to define the term weight [12]. In general, the weight of each term is given by the following formula:

$$\text{Term Weight } w_t = \mathrm{tf_t} * \mathrm{IDF_t} \quad (1)$$

Where tf_t: Term frequency i.e. number of times a term t occurs in a web page and IDF_t: Inverse Document frequency.

$$\mathrm{IDF_t} = \log\left(\frac{\mathrm{D}}{\mathrm{df_t}}\right) \quad (2)$$

Where D : Total number of web pages under parent web page and df_t: Number of web pages under parent web page in which term t appears.

After calculating term weight each concept term of topic ontology and each word in web page, find the cosine similarity between them [13].

$$\mathrm{Sim}(f,P) = \frac{\sum_j w_{f,j} * w_{P,j}}{\sqrt{\sum_j w_{f,j}^2} * \sqrt{\sum_j w_{P,j}^2}} \qquad (3)$$

Where $w_{f,j}$: weight of term j in topic ontology and $w_{P,j}$: weight of term j in web page. The semantic score of unvisited URL is calculated by the following equation:

$$\text{Semantic score} = SS(f, WP) + SS(f, AT) + \text{Semantic score}(P) \qquad (4)$$

Where SS(f, WP) is cosine similarity score between topic ontology and web page of URL WP, SS(f, AT) is Cosine Similarity between topic ontology and Anchor text AT and Semantic score (P) is semantic score of parent.

3.2 Priority based Focused crawling

Let crawler starts crawling with initial seed URL with semantic score 1. Web page is downloaded from web for seed URL and new URLs are extracted from downloaded page i.e. URL 1, URL 2,... , URL N.

Now, crawler again downloads web pages for every new URL which are Page 1, Page 2,...., Page N. then, Semantic score is calculated between web pages and topic ontology. Let Page 1, Page 2,....., Page N has score 0.7, 0.5,....., 0.8 respectively where 0.8 is maximum score. Downloaded pages and their score are inserted into priority queue. Now a page is deleted from priority queue, page with highest score is deleted. The maximum score URL is URL N, crawler will extract the entire URL from Page N first.

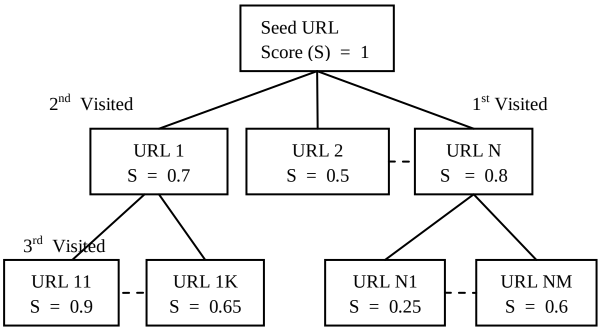

Fig 4: URL Crawling Sequence Tree

Again, it will repeat the same process i.e. extracting URLs (URL N1, URL N2,...., URL NM) and downloading pages (Page N1, Page N2,...., Page NM). Calculate the semantic score between newly downloaded pages and focus word. Let Page N1,....., URL NM has score 0.25, ..., 0.6 respectively where 0.6 is maximum. Downloaded pages and their score is inserted into priority queue. Now a page is deleted from priority queue at this time Page 1 is selected because it has maximum score then remaining other pages present in queue. Every time maximum score page is selected for crawling. This will be the advantages of using priority queue over simple queue. This will definitely improve performance of crawling process over normal crawling process.

4. CRAWLING ALGORITHM

The priority based semantic crawling algorithm works as follows:
Input: Initial seed URL, Topic and PQueue.
Output: Web_Pages related to Focus_String.

```
Page := downloadPage(URL);
addPQueue(Page, 1);
While PQueue is not empty do
        Page := dePqueue();
        newURLs := extractURL(Page);
        for each ith URL in newURLs do
            Page[i] := downloadPage(newURLs[i]);
            Onto_Topic := Ontology(Topic);
            SScore[i] := SimScore(Page[i],  Onto_Topic) +
                    SimScore(Anchor_text,  Onto_Topic) +
                            SScore(Parent(page[i]));
            addPQueue(Page[i],  SScore[i]);
        end for;
end while;
```

Descriptions of various functions used in algorithm are as follows:

1. addPQueue(Page, Score) : add a new downloaded page and its similarity score with Focus_String into Priority Queue.

2. dePQueue( ): returns a page which has maximum score from Priority Queue.

3. downloadPage(URL): downloads web page from WWW corresponding to given URL.

4. extractURL(Page): extracts all URLs which are present in given Page.

6. Ontology(Topic): generates ontology of given topic named Onto_Topic.

5. SimScore(Page, Onto_Topic): calculates similarity score between Page and Onto_Topic.

5. EXPERIMENTAL RESULTS

The efficiency of Focused crawler is calculated by harvest rate. Harvest Rate measures the rate at which relevant pages are crawled and how effectively irrelevant pages are filtered off from the crawl [14].

$$\text{Harvest ratio} = \frac{\text{No.of relevent web pages crawled}}{\text{Total no.of web pages crawled}} \qquad (5)$$

For better crawling performance, harvest ratio should be high. The standard data set is used for experiment that is present on Internet in the form of open directory named “http://www.dmoz.org”. The harvest rate of our crawler, simple crawler, focused crawler [2] and priority based focused crawler [11] are evaluated and compared.

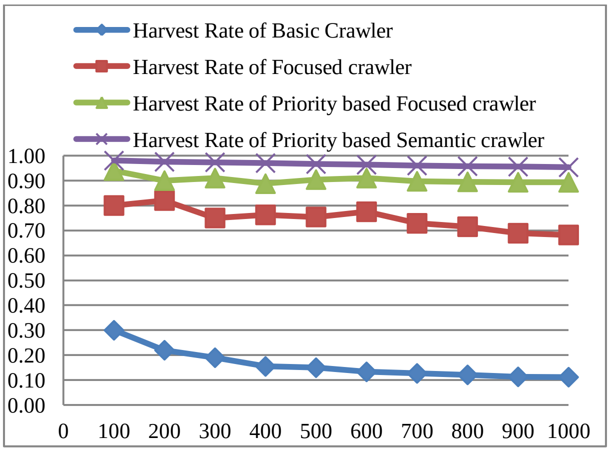

Fig 5: Harvest ratio v/s No. of pages crawled

The 1000 web pages crawled on topic Computer, Science, Arts and Sports, and average of harvest rate is taken. The figure 5 shows the harvest rate of simple crawler, focused crawler, priority based focused crawler and priority based semantic crawler. The over all performance gain over simple crawler is 88%, focused crawling is 28% and priority based focused crawler is 6%.

6. CONCLUSION AND FUTURE WORK

In this paper, a priority based semantic crawler has been proposed that keeps all URLs to be visit in priority queue along with their semantic score. Priority queue always returns maximum score URL every time. The performance of crawler is evaluated on topic Computer, Science, Arts and Sports. The experimental results show that our crawler gives 6% improved results over priority based focused crawler, 28% improved results over simple focused crawler and 88% improved results over simple crawler. In future, we will try to reduce the crawling time by parallel implementation of crawling algorithm.

7. REFERENCES

[1] Singhal, N., Dixit, A. and Sharma, A.K. 2010. Design of a Priority Based Frequency Regulated Incremental Crawler. International Journal of Computer Applications, Volume 1, No. 1, PP. 42-47.

[2] Tsoi, Ah C., Forsali, D., Gori, M., Hagenbuchner, M. and Scarselli, F. 2003. A Simple Focused Crawler. WWW 2003: ACM.

[3] Snasel, V., Moravec, P. and Pokorný, J. 2005. WordNet Ontology Based Model for Web Retrieval. In Proceedings of the International Workshop on Challenges in Web Information Retrieval and Integration (WIRI'05).

[4] Gruber, T. R. 1993. A Translation Approach to Portable Ontology Specifications. *Knowledge Acquisition*, 5, Academic Press Ltd., PP.199–220.

[5] Mizoguchi, R., Vanwelkenhuysen, R. and Iked, M. 1995. Task ontology for reuse of problem solving knowledge. In Proceedings of Towards Very Large Knowledge Bases: Knowledge Building & Knowledge Sharing.

[6] Ganesh, S., Jayaraj, M., Kalyan, V., Murthy, S. and Aghila, G. 2004. Ontology-based Web Crawler. In proceedings of the International Conference on Information Technology: Coding and Computing (ITCC'04), IEEE.

[7] Mukhopadhyay, D., Biswas, A. and Sinha, S. 2010. A New Approach to Design Domain Specific Ontology Based Web Crawler. In proceedings of 10th International Conference on Information Technology, IEEE.

[8] Chen, X. and Zhang, X. 2008. HAWK: A Focused Crawler with Content and Link Analysis. In proceeding of International Conference on e-Business Engineering, IEEE.

[9] Hati, D., Sahoo, B., Kumar, A. 2010. Adaptive Focused Crawling Based on Link Analysis. In proceeding of 2nd International Conference on Education Technology and Computer (ICETC), IEEE.

[10] Thenmalar, S. and Geetha, T. V. 2011. Concept based Focused Crawling using Ontology. International Journal of Computer Applications, Volume 26, No.7, PP. 29-32.

[11] Choudhary, J. and Roy, D. 2013. Priority based Focused Web crawler. International Journal of Computer Engineering and Technology, Vol. 4, No. 4, PP. 163-169.

[12] Salton, B. 1988. Term-Weighting Approaches in Automatic Text Retrieval. Information Processing and Management Elsevier, Vol. 24, No. 5, PP. 513-523.

[13] Lee, D. L., Chuang, H. and Seamons, K. 1997. Document Ranking and the Vector-space Model. IEEE Software, Vol. 14, No. 2, PP. 67-75.

[14] Chakrabarti, S., van den Berg, M. and Dom, B. 1999. Focused crawling: a new approach to topic-specific Web resource discovery. In proceeding of 8th International WWW Conference.