VOL. 11, NO. 15, AUGUST 2016 ISSN 1819-6608
ARPN Journal of Engineering and Applied Sciences

www.arpnjournals.com

FOCUSED CRAWLING OF ONLINE BUSINESS WEB PAGES USING LATENT SEMANTIC INDEXING APPROACH

Thamer Salah[1] and Sabrina Tiun[2]
[1]Faculty of Information Science and Technology, Universiti Kebangsaan Malaysia, Selangor, Malaysia
[2]Center for Artificial Intelligence Technology, Faculty of Information Science and Technology, Universiti Kebangsaan Malaysia, Selangor, Malaysia
E-Mail: thameralj1@gamil.com

ABSTRACT

With the exponential growth of textual information available from the Internet, there has been an emergent need to find relevant, in-time and in-depth knowledge about business topic. The huge size of such data makes the process of retrieving and analyzing and use of the valuable information in such texts manually a very difficult task. In this paper, we attempt to address a challenging task i.e. a crawling business-specific knowledge on the Web. To do that, the main goal of this paper is to describe a new method of focused crawling with latent semantic indexing for online business web pages. We describe a new model for online business text crawling which seeks, acquires, maintains and filter business pages. This model consists mainly from two main modules: a crawling system and a text filtering system. The crawler is used to collect as many web pages as possible from the news websites. This focused crawler is guided by a latent semantic index and information from Word Net (business filter) which learns to recognize the relevance of a web page with respect to the business topic and it is also utilized a set of domain specific keywords. The obtained results also on online real word data show that the focused crawler is very effective for building high-quality collections of business Web documents.

Keywords: business data mining, web mining, focused crawling, classification.

1. INTRODUCTION

As of today the indexed web contains more than 50 billion web pages [1] and continues to grow at a rapid pace. With the rapid growth of the World Wide Web, the volume of the Business information that is available on the web is growing exponentially. Since there has been an explosion of media reports for different kind of business news, this makes the process of analyzing and processing them manually is a very difficult task. It is also widely known that general purpose search engines are not tailored at providing topic specific information [1]. Therefore, the huge amounts of business news need to be organized in an effective way. One ways of organizing this overwhelming amount of data is to gather these business web pages from the Internet and classify them into their appropriate categories. This organized and classified data is essential to many information retrieval tasks such as constructing or expanding web directories (web hierarchies), improving search results, helping question answering systems and building domain-specific search engines. To gather such domain-specific web pages, domain-specific web crawler has to be developed to collect web pages from the Internet by choosing to gather only pages related to this domain. This type of web crawler does not need to gather every web page from the Internet. In fact, during the focused crawling process of a search engine, the crawler uses an automatic classification mechanism to determine whether the Web page being considered is ‘‘on the specific topic” or not [2, 3].

Our contribution: In this paper, we describe a new model for online business focused-crawling which seeks, acquires, maintains and classifies pages on business topic. This model consists mainly from two main modules: a crawling system and a text filtering system. The crawler is used to collect as many web pages as possible from the news websites. This focused crawler is guided by a latent semantic index and information from WordNet (Business filtered) which learns to recognize the relevance of a web page with respect to topic on the Business and it is also utilized a set of domain specific keywords.

This paper is organized as follows: In Section 2, we give a summary of related works in focused crawling. Section 3 describes our model for focused crawling of online business web pages. Section 4 describes the evaluation methods. The experimental results and discussion on the results are presented in Section 5. Finally, Section 6 concludes the study and gives some future works.

2. RELATED WORK

Focused crawling is a promising approach to improving the recall of expert search on the Web. A variety of methods for focused crawling have been proposed [1,4-12]. The term focused crawler was first coined by Chakrabarti in 1999. Chakrabarti [13] uses a canonical topic taxonomy and seed documents to build a model for classification of retrieved pages into categories. Earliest work on focused crawling dealt with simple keyword matching or regular expression matching. Related research in focused web crawling algorithms is presented in [14, 15]. Topic specific crawlers attempt to focus the crawling process on pages relevant to the topic. They try to keep the overall number of downloaded web pages for processing as minimum as possible and maximizing the percentage of relevant pages [8]. Angkawattanawit (2002) deal with improving re-crawling performance by utilizing several databases (seed URLs, topic keywords and URL relevance predictors) that are built from previous crawl process and used to improve

harvest rate. Several works [6, 16] utilize search engines as a source of seed URLs and back–references.

3. RESEARCH METHOD

As the size of the Web grows, topic-specific Web Crawlers are becoming more important due to their ability to acquire, and maintain a collection of Web pages relevant to a certain topic .The objective of our proposed task is to classify business-specific web pages to help the user to find relevant, in time and in-depth knowledge about business topic. In this section, we describe the architecture of our business-specific crawler. (Figure-1) shows the architecture of our business-topic crawler. As shown in Figure-1, there are two main modules in the architecture: (i) module A: webpages downloader phase and (ii) module B: webpages filtering phase:

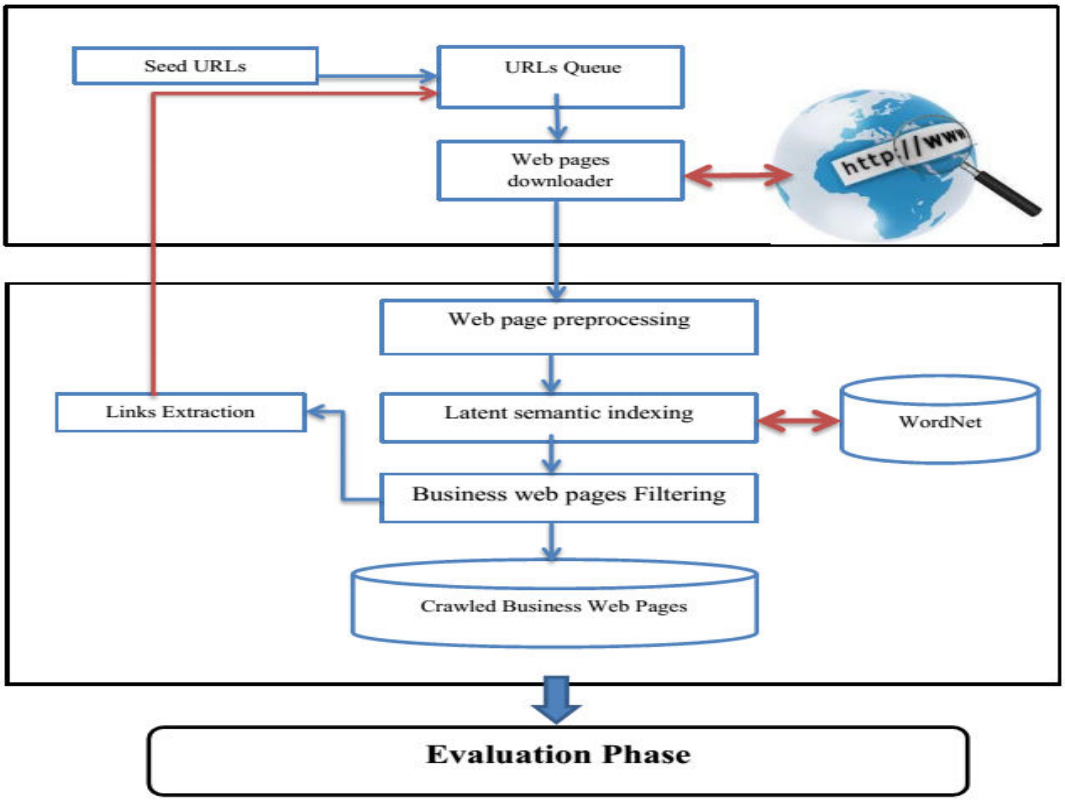

Figure-1. The architecture for focused crawling with Latent Semantic Indexing of online business Web Page.

3.1 Module A: Webpages downloader

Crawlers are given a starting set of web pages (seed pages) as their input, extract outgoing links appearing in the seed pages and determine what links to visit next based on certain criteria. Web pages pointed to by these links are downloaded, and those satisfying certain relevance criteria are stored in a local repository. Crawlers continue visiting Web pages until a desired number of pages have been downloaded or until local resources (such as storage) are exhausted. The Crawling Phase has the following steps:

a. **Input:** Crawlers take as input a number of starting (seed) URLs. Crawlers are given a starting set of web pages (seed pages) as their input, extract outgoing links appearing in the seed pages and determine what links to visit next based on certain criteria. These are stored in URLs queues.

Page downloading: In this step, the crawler selects the highest score URL from the URL Queue. A page downloader downloads the page associated with this URL. Then, a preprocessing module is used for parsing the web page, performing stop word removal, and removing all the HTML tags. After that, a latent semantic filtering is used determine whether the downloaded web page is business page or not. New extracted URLs will be ordered by their scores and then be put into a URL queue. The links in downloaded pages are extracted and placed in a queue. A non-focused crawler uses these links and proceeds with downloading new pages in a first in, first out manner. A focused crawler reorders queue entries by applying content relevance or importance criteria or may decide to exclude a link from further expansion (generic crawlers may also apply importance criteria to determine pages that are worth crawling and indexing).

3.2 Module B: Webpages filtering phase

In this phase the crawled pages are first preprocessed to remove noise from their content. After that, these pages are filtered using latent semantic indexing.

3.2.1 Web Page Preprocessing

The pre-processing phase is an essential phase in the design of focused crawling especially in Latent Semantic filtering of online business Web Pages. The downloaded web pages are never 100% pure, so it may have some noise. HTML, links, or the programming language code have been removed from the text. Web page normalization includes several parts, as following:

a) Eliminating script code which influence analysis, such as the content of the script node and the action script in normal web page nodes;
b) Remove style code in the page, such as the content of the style node and style attribute of normal web page nodes;
c) To eliminate commented code, that is comments in your web page, the general format is "<! -- comment content -- - >";
d) Rejecting other irrelevant code, such as banner ads, copyright statement and so on.

3.2.2 Latent Semantic Indexing (LSI) filtering phase

Generally information retrieval is based on exact matching, that is, the terms in the query are matched to those in the document. Latent Semantic Indexing (LSI) is used to filter or retrieve web pages based on the basis of conceptual meaning of the query and web page. For this, LSI is a technique that enables us to analyses relationships between terms and concepts occurring in a text. LSI uses a mathematical technique called Singular Value Decomposition (SVD) by (Brand, M. (2006)).

In SVD a $t \times d$ matrix X of terms and documents is formed, where each matrix element is the term frequency or other weight, t is the number of rows of X, d is the number of columns of X, m is the rank of X. X can be decomposed into the product of three other matrices as in equation (1):

$$X = T_0 \, S \, D_0 \tag{1}$$

This is called the singular value decomposition of X. T_0 And D_0 are the matrices of left and right singular vectors, and S_0 is the diagonal matrix of singular values. Singular value decomposition (SVD) is unique up to certain row, column and sign permutations and by convention the diagonal elements of S_0 are constructed to be all positive and ordered in decreasing magnitude.

The singular values in S_0 are ordered by size, the first k largest may be kept and the remaining smaller ones set to zero. The product of the resulting matrices is a matrix $\hat{X}$ which is only approximately equal to$X$. Deleting the zero rows and columns of S_0 to obtain a new diagonal matrix S_0, and then deleting the corresponding columns of T_0 and D_0 to obtain T and S respectively. The result is a reduced model, as in equation (2):

$$X \approx \hat{X} = T * S * D' \tag{2}$$

Which is the rank-k model with the best possible least-squares fit to X. The dot product between two row vectors of $\hat{X}$ reflects the extent to which two terms have a similar pattern of occurrence across the set of documents. It is easy to verify that equation (3):

$$\hat{X} * \hat{X} = TS^2T' \tag{3}$$

Where the i,j cell of $\hat{X}$ is the similarity between term i and term$j$. In order to compare a query to other documents, we need to start with its term vector x_q and derive a representation q in reduced dimensional vector space, as in equation (4):

$$q = X_q' T S^{-1} \tag{4}$$

And then q can be made between or within comparisons, respectively. The cosine similarity between query q and document d can be equation (5):

$$sim(q,d) = \frac{\sum_{i=1}^{t} q_i d_i}{\sqrt{\sum_{i=1}^{t} q_i^2 \sum_{i=1}^{t} d_i^2}} \tag{5}$$

However, the general document representation does not able to handle the two classic problems arising in natural languages: synonymy and polysemy. Synonymy refers to the problem where two different words have the same meaning. Because the vector space representation fails to capture the relationship between synonymous terms such as car and automobile. In this work, WordNet is utilized to enrich of the body of the crawled web pages. The WordNet is used in two ways. First, the WordNet is used to extract the business domain vocabularies. Second the WordNet is used utilized to enrich of the body of the crawled web pages by replacing a word in the crawled web pages by its synonyms.

Below is the example how WordNet is used in the filtering phase. The WordNet is used utilized to enrich of the body of the crawled web pages by replacing a word in the crawled web pages by its synonyms without changing the true meaning to crawling web page like changing word (lorry) to (truck) and switch zero to one as shown the WordNet process in Figure-1. Therefore, an original sentence (without WordNet intervention) as " Gold silver lorry", and all the documents that will be crawled containing all or some of the terms - "gold", "silver" and "lorry" (see document d1, d2, and d3). With the intervention of WordNet the original sentence changed into this:

"gold silver truck", and the collection of crawled documents will be changed as well.

A "collection" consists of the following "documents":

d1: Shipment of gold damaged in a fire.

d2: Delivery of silver arrived in a silver truck.

d3: Shipment of gold arrived in a truck.
Crawling webpage:" gold silver lorry"
Before WordNet: "Gold silver lorry
Set term weights and construct the term-document matrix A and crawling webpage matrix:

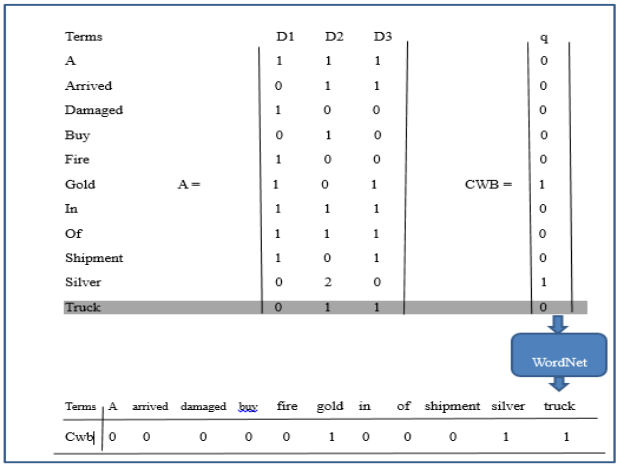

Figure-2. Sample example of WordNet process.

After WordNet: Crawling webpage "gold silver truck".

4. EVALUATION METHOD

We empirically evaluate the effectiveness of our model through evaluating the components of the model on online real world data. Harvest ratio is a significant IR performance metric, and is broadly used in the evaluation of focused Web crawlers [17, 18], see in equation (6). The crawler performance is typically measured by the harvest rate i.e. the percentage of downloaded pages that are relevant to the business topic. Equation (6):

$$Harvest\_rate = \frac{relevant\_pages}{pages\_downloaded} \qquad (6)$$

4.1 Experimental Result

In this section, we present the evaluation of our business specific crawler. At the initial stage, the crawler is initialized using a set of seed URLs. These URLs are selected automatically from the fetched results of Bing search engine. All crawling and searching processes are limited to Malaysian news web pages. Then, the URLs are put into a priority queue according to the cosine similarity of their description and the business corpus. After that, the crawler begins to download these web pages according to their URLs priority. Hyperlinks are extracted from the downloaded web pages and put into the priority queue. The results for each method are represented by a plot showing the number of relevant pages returned by the method as a function of the total number of downloaded pages. Table-1 shows the Harvest Rate percentage of business pages to the number of crawled pages. The results of the crawler also shown in (Figure-2). The x-axis shows the total number of pages crawled. The y-axis shows the number of crawled business web pages. In general, it can be observed that all used model is highly accurate business filters. The business-focused crawling can crawl relevant business web pages more effectively as it has high harvest rates.

Table-1. Harvest rate (Percentage of business pages to the number of crawled pages) of the business-Topic Crawler.

Crawled pages	Relevant pages	Harvest rate
100	80	0.8
300	272	0.91
400	373	0.93
500	465	0.93
800	756	0.95
1000	883	0.88
1500	1185	0.79
2000	1518	0.76
2500	1823	0.73
3000	2128	0.71
3500	2393	0.68
4000	2632	0.66
4500	2898	0.64

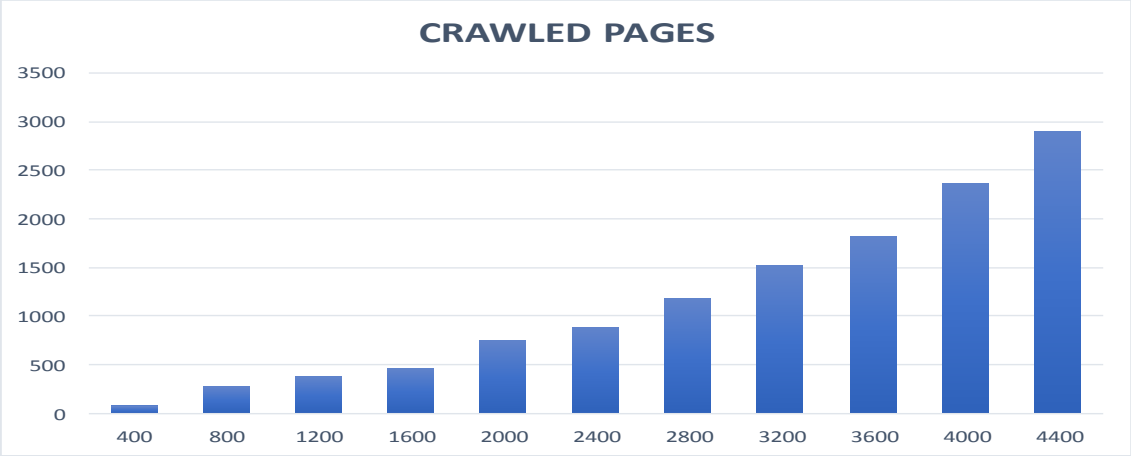

Figure-3. Harvest rate (Number of business pages to the number of overall crawled pages) rate of the business-topic.

5. CONCLUSIONS

This paper describes our work in crawling and filtering the business Web pages. In particular, we have proposed a new model for online business text crawling and filtering. This model consists mainly from two main modules: a crawling system and a text filtering system. The crawler is used to collect as many web pages as possible from the news websites. This focused crawler is guided by a latent semantic index and information from WordNet (business filter) which learns to recognize the relevance of a web page with respect to the business topic and it is also utilized a set of domain specific keywords. Our results also on online real word data show that the focused crawler is very effective for building high-quality collections of business Web documents.

In the future, we have identified several important directions for future research. We plan to will expand the multi-class classification methods into multi-label and multi-classes classification models in which a business document can be assigned to more than one class. We also plan to integrate utilizes natural language processing technology and machine learning algorithms to analyze content, extracting useful information and to provide clear insight into the content of business Web pages.

ACKNOWLEDGEMENT

This research project is partially funded by Malaysia Government under research Grant ERGS/1/2013/ICT07/UKM/03/1.

REFERENCES

[1] Samarawickrama S. and L. Jayaratne. 2011. Automatic text classification and focused crawling. In Digital Information Management (ICDIM), 2011 Sixth International Conference on. IEEE.

[2] Qi X. and B.D. Davison. 2009. Web page classification: Features and algorithms. ACM Computing Surveys (CSUR). 41(2): 12.

[3] Özel S.A. 2011. A Web page classification system based on a genetic algorithm using tagged-terms as features. Expert Systems with Applications. 38(4): 3407-3415.

[4] Can A.B. and N. Baykal. 2007. MedicoPort: A medical search engine for all. Computer methods and programs in biomedicine. 86(1): 73-86.

[5] Yang S.-Y.A. 2010. Focused crawler with ontology-supported website models for information agents, in Advances in Grid and Pervasive Computing. Springer. pp. 522-532.

[6] Yang S.-Y. 2010. OntoCrawler: A focused crawler with ontology-supported website models for information agents. Expert Systems with Applications. 37(7): 5381-5389.

[7] Wang W., et al. 2010. A focused crawler based on naive Bayes classifier. in Intelligent Information Technology and Security Informatics (IITSI), 2010 Third International Symposium on. IEEE.

[8] Batsakis S., E.G. Petrakis and E. Milios. 2009. Improving the performance of focused web crawlers. Data and Knowledge Engineering. 68(10): 1001-1013.

[9] Zheng H.-T., B.-Y. Kang and H.-G. Kim. 2008. An ontology-based approach to learnable focused crawling. Information Sciences. 178(23): 4512-4522.

[10] Ehrig M. and A. Maedche. 2003. Ontology-focused crawling of Web documents. in Proceedings of the 2003 ACM symposium on Applied computing. ACM.

[11] Hsu C.-C. and F. Wu. 2006. Topic-specific crawling on the web with the measurements of the relevancy context graph. Information Systems. 31(4): 232-246.

[12] Liu H., J. Janssen and E. Milios. 2006. Using HMM to learn user browsing patterns for focused web crawling. Data and Knowledge Engineering. 59(2): 270-291.

[13] Chakrabarti S., B. Dom and P. Indyk. 1998. Enhanced hypertext categorization using hyperlinks. in ACM SIGMOD Record. ACM.

[14] Novak B. 2004. A survey of focused web crawling algorithms.

[15] Liu J.-H. and Y.-L. Lu. 2007. Survey on topic-focused Web crawler. Application Research of Computers. 10: 006.

[16] Martin N. and K. Khelif. 2011. Focused crawling using name disambiguation on search engine results. in Intelligence and Security Informatics Conference (EISIC), 2011 European. IEEE.

[17] Seyfi A., A. Patel and J.C. Júnior. 2016. Empirical evaluation of the link and content-based focused Treasure-Crawler. Computer Standards and Interfaces. 44: 54-62.

[18] Patel R. and P. Bhatt. 2015. Semantic Focused Web Crawler for Service Discovery Using Data Mining Technique. Compusoft. 4(7): 1923.

[19] Brand M. 2006. Fast low-rank modifications of the thin singular value decomposition. Linear algebra and its applications. 415(1): 20-30.