ISSN:2320-0790

An International Journal of Advanced Computer Technology

Semantic Focused Web Crawler for Service Discovery Using Data Mining Technique

[1]Ruchika Patel, [2]Pooja Bhatt

[1,2]Department of Computer Engineering, Ipcowala Institute of Engineering & Technology, Dharmaj, Anand, Gujarat, India – 388430

Abstract: Data mining is the process of extraction of hidden predictive information from the huge databases. It is a new technology with great latent to help companies focus on the most important information in their data warehouses. Web mining is a data mining techniques which automatically discover information from web documents. The amount of data and its dynamicity makes it impossible to crawl the World Wide Web (WWW) completely. It's a challenge in front of crawlers to crawl only the relevant pages from this information explosion. Thus a focused crawler solves this issue of relevancy by focusing on web pages for some given topic or a set of topics. Nowadays finding meaningful information among the billions of information resources on the World Wide Web is a difficult task due to growing popularity of the Internet. This paper basically focuses on study of the various techniques of data mining for finding the relevant information from World Wide Web using web crawler.

Keywords: Web Mining, Web Crawler, Focused Crawler, World Wide Web (WWW)

I. Introduction

The internet has becoming the largest unstructured database for accessing information over the documents. [8] It is well recognized that the information technology has a profound effect on the conduct of the business, and the Internet has become the largest marketplace in the world. Innovative business professionals have realized the commercial applications of the internet for for their customers and strategic partners. [2] With the rapid growth of electronic text from the complex the WWW, more and more knowledge you need is included. But, the massive amount of text also takes so much trouble to people to find useful information. For example, the standard Web search engines have low precision, since typically some relevant Web pages are returned mixed with a large number of irrelevant pages, which is mainly due to the situation that the topic-specific features may occur in different contexts. So, one appropriate way of organizing this overwhelming amount of documents is necessary.[1] The World Wide Web is an architectural framework for accessing linked documents spread out over millions of machines all over the Internet. The popularity of www is largely dependent on the search engines. Search engines are the gateways to the huge information repository at the internet. Search engine consist of four discrete components: Crawling, Indexing, Ranking and query processing. The earliest Web search engines relied for retrieving information from Web pages on the bases of matching with the words in the search query. As the Web continues to grow, and as the diversity of users increases search engines must utilize semantic clues to satisfy user's information needs. Semantic search, which takes into account the interests of the user as well as the specific context in which the search is issued, is the next step in providing users with the most relevant information possible.

Currently the general purpose search engines strive as entry points for the web pages perform the coverage of information that is as broad as possible. They use Web crawlers to maintain their index databases. These crawlers are blind and exhaustive in their approach, with comprehensiveness as their major goal. A URL (Uniform Resource Locator) that is a URI (Uniform Resource Identifier) specifies where an identified resource is available. In order to search most relevant information, crawlers can be more selective about the URL they fetch and refer as to be

crawled this mechanism.[14] The outline of this paper is organized as follows. We
Review the introduction of semantic focused web crawler in section II. Section III illustrates the techniques used in Web crawling for finding relevant information among the World Wide Web. Section IV draws the conclusions.

II. Introduction to Semantic Focused Web Crawler

A crawler is an agent which can automatically search and download WebPages. Focused (topical) crawlers are a group of distributed crawlers that specialize in certain specific topics. Each crawler will analyze its topical boundary when fetching WebPages. [7]
A semantic focused crawler is a software agent that is able to traverse the Web, and retrieve as well as download related Web information for specific topics, by means of semantic Web technologies. The goal of semantic focused crawlers is to precisely and efficiently retrieve and download relevant Web information by understanding the semantics underlying the Web information and the semantics underlying the predefined topics. [2]
After the World Wide Web emerged, researchers attempted to enhance its quality by various semantic technologies. Currently there are three new forms of recognized webs enhanced by various semantic technologies, which are semantic web, semantic grid, and knowledge grid. Semantic web is "a web of data", which is used to express the meaning of web data by means of diverse ontological mark up languages, such as XML, RDF, OWL and so forth. [14] It provides the machine understandable information for computers to retrieve, share and merge knowledge on the internet.[8]

III. Techniques used in Web crawling

Lu LIU et al presents a novel clustering-based topical Web Crawling for domain-specific information retrieval guided by link-context. The method this paper proposed achieves a topical crawling quality comparable to previous Best First, Anchor text, Link context, and DTC crawling methods. It provides a series quantitative analysis for the effectiveness of the approach. One of the most prominent advantages of clustering-based topical Web crawling is that it does not require any labeled training data, which is quite costly, exhausted and difficult to obtain. Next, a new kind of data structure CFu tree is adopted to speed up the process of hierarchical clustering. Finally, from the view of cluster quality, it defines three metrics CIP, Cprecison and
Crecall to evaluate the eroor rate, precision and recall respectively. The results show that this hierarchical clustering method can achieve a relatively high precision and recall. [1]
Bireshwar Ganguly et al used Focused Crawlers and Block Partitioning of Web pages. In this paper this approach is to partition the web pages into content blocks. The method of partitioning the web pages into blocks on the basis of headings gives an advantage over conventional block partitioning is that it divides the blocks which include a complete topic. The heading, hyperlinks and the body of a particular topic is included in one complete block. After calculating the relevancy of different regions it calculates the relevancy score of web page based on its block relevancy score with respect to topics and calculates the URL score based on its parent pages blocks in which this link does exist. [6]
Rodolfo Zunino et al presents the architecture of the semantic Focused Crawler features static flexibility in the definition of desired concepts, used metrics, and crawling strategy; in addition, the method is capable to learn the analyst's expectations at runtime. The user may instruct the crawler with a binary feedback (yes/no) about the current performance of the surfing process, and the crawling engine progressively refines the expected targets accordingly. The method implementation is based on an existing text mining environment, integrated with semantic networks and ontologies. [3]
Farookh Khadeer Hussain et al presents conceptual model of an ontology based focused crawler serving in the domain of transport services. The whole crawling system consists of three parts – a focused crawler, a transport service ontology base, and a transport service metadata base. The focused crawler takes the responsibility of downloading web pages, analyzing and parsing web documents, extracting meaningful information from the documents and forming metadata based on the information, and logically linking the metadata and ontological concepts; the ontology base is for storing a transport service ontology, with the purpose of limiting the crawling scope; the metadata base is for storing the transport service metadata and the links of metadata to concepts. The ECBR algorithm for the logical links is derived from the CBR algorithm in the data mining field, by rebuilding it on the basis of index terms. [7]
Hai Dong et al propose an ontology learning-based focused crawling approach integrates an
Ontology based focused crawling framework, a vocabulary based ontology learning framework, and a hybrid mathematical model for service advertising information similarity computation. The proposed ontology learning based focused crawler primarily consists of three components based on the functionalities, i.e., a storage component - the service

knowledge base, a processing component - the crawling and processing module, and a computing component - the service advertising information classification and ontology learning module. [2]

Hai Dong et al present the framework of a novel self-adaptive semantic focused crawler – SASF crawler, with the purpose of precisely and efficiently discovering, formatting, and indexing mining service information over the Internet, by taking into account the three major issues namely heterogeneity, ubiquity, and ambiguity. This approach involved an innovative unsupervised ontology learning framework for vocabulary-based ontology learning, and a novel concept-metadata matching algorithm, which combines a semantic similarity based SeSM algorithm and a probability based StSM algorithm for associating semantically relevant mining service concepts and mining service metadata. This approach enables the crawler to work in an uncontrolled environment where the numerous new terms and ontologies used by the crawler have a limited range of vocabulary. [4]

The proposed flow chart is as shown in figure 1. The TSVM classifier for each concept is designed to best aggregate the results of the semantic similarity based algorithm and the probability based algorithm in order to decide on the semantic relatedness between a concept description and a page description, through a supervised training paradigm. This classifier provides a binary classification function (relevant/non relevant) which is characterized by a hyperplane in a given feature space. The result of an TSVM is a maximum margin hyperplane which separates training tuples in the feature space as precise as possible while the distance of the closest members on each side is maximized.

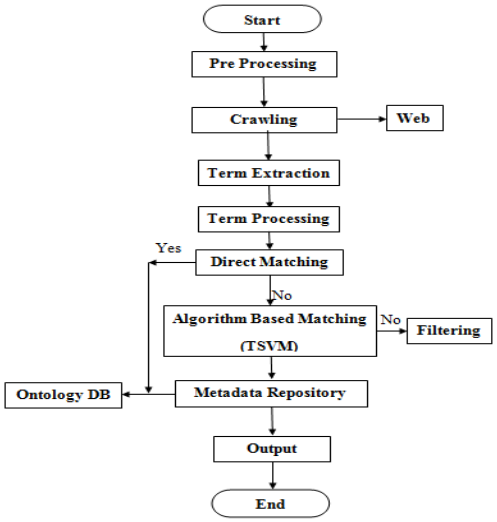

Fig. 1. The flow chart of proposed system

IV. System Evaluation

Five performance indicators of information retrieval to evaluate ontology based focused crawler are used which are harvest rate, precision, recall, harmonic mean, fallout rate and crawling time.

Harvest rate in the information retrieval is used to measure the crawling ability of a crawler. Harvest rate is the ratio of logically linked metadata to the whole collection of metadata. It can be defined below: [7]

$$\text{Harvest rate} = \frac{\text{number of logically linked metadata}}{\text{number of metadata}}$$

It can be seen that the proposed model has the optimal performance (52%), compared to the self adaptive model (around 46%).

Precision in the information retrieval is used to measure the preciseness of a retrieval system. Precision for a single concept is the ratio of logically linked and relevant metadata to all logically linked metadata. It can be defined below: [7]

$$\text{Precision(S)} = \frac{\text{number of logically linked relevant metadata}}{\text{number of logically linked metadata}}$$

The whole precision is the sum of precision for each concept in the collection and it can be defined below: [7]

$$\text{Precision(W)} = \frac{\sum_{i-1}^{n} \text{Precision}(S_i)}{n}$$

It can be observed that the overall precision of the proposed model is 34.20%, and the overall precision of the self adaptive model is 31.59%.

Recall in the information retrieval is used to measure the effectiveness of a query system. Recall is the ratio of logically linked and relevant metadata to all relevant metadata. It can be defined below: [7]

$$\text{Recall(S)} = \frac{\text{number of logically linked relevant metadata}}{\text{number of relevant metadata}}$$

The total recall is the sum of recall for each concept in the collection. It can be defined below: [7]

$$\text{Recall(W)} = \frac{\sum_{i-1}^{n} \text{Recall}(S_i)}{n}$$

It can be seen that the overall recall for the proposed model is 65.59%, compared to only 63.21% for the self adaptive model.

The fallout rate is ratio of logically linked and non relevant metadata to the whole collection of non relevant metadata to the concept. It is defined below: [7]

$$\text{Fallout rate(S)} = \frac{\text{No. of logically linked non-relevant metadata}}{\text{No. of non-relevant metadata}}$$

The whole fallout rate is the sum of fallout rate for each concept in the collection. It can be represented as: [7]

$$\text{Fallout rate(W)} = \frac{\sum_{i=1}^{n} \text{Fallout rate}(S_i)}{n}$$

The fall rate value is lower, the crawler's performance is better. It can be seen that the overall fallout rate of the proposed model is 0.43%, and for the self adaptive model is around 0.46%.

Harmonic Mean is defined as the total performance of precision and recall value. Harmonic Mean is define as below: [7]

$$\text{Harmonic Mean} = \frac{\text{Precision} + \text{Recall}}{\text{Precision} \,.\, \text{Recall}}$$

As an aggregated parameter, the overall harmonic mean values for both of these models are below 50% (43% for the self adaptive model and 44.46% for the proposed model).

Crawling time is used to measure the efficiency of a crawler. The crawling time for a Web page is defined as the time interval from processing the Web page from the Crawling process to the Metadata Generation and Association process or to the Filtering process. [7] Overall, the average crawling speed for the proposed model is 77.00 ms/page and for the self adaptive model is 70.00 ms/page.

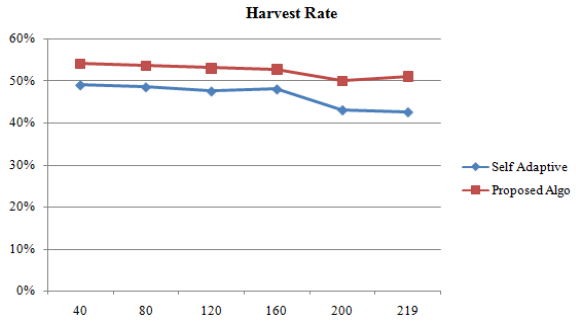

Fig. 2. Comparison of models on Harvest Rate

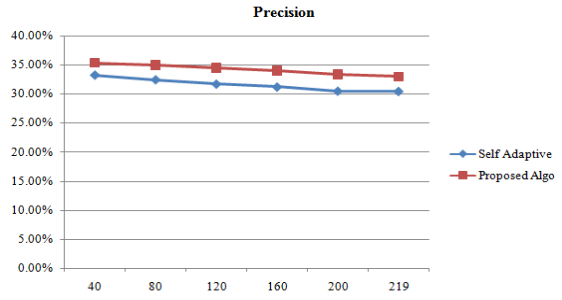

Fig. 3. Comparison of models on Precision

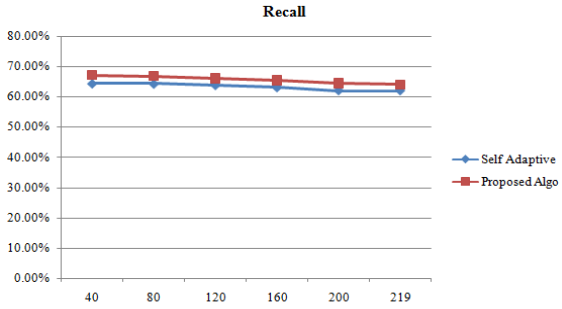

Fig. 4. Comparison of models on Recall

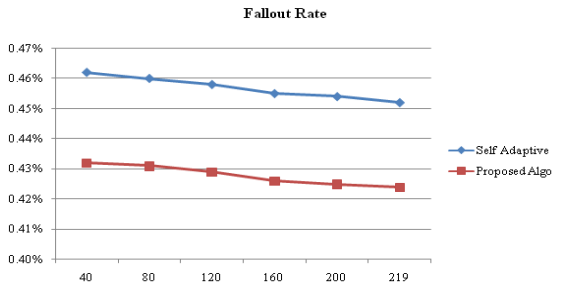

Fig. 5. Comparison of models on Fallout Rate

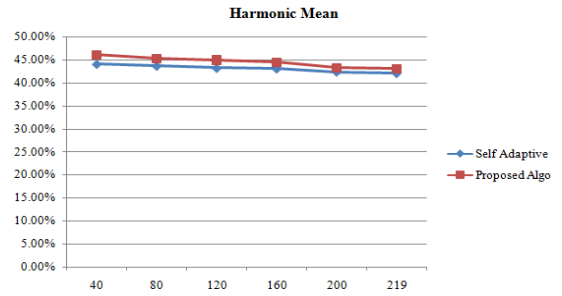

Fig. 6. Comparison of models on Harmonic Mean

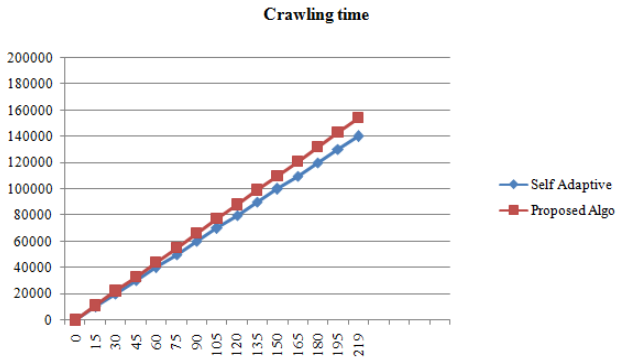

Fig.7. Comparison of models on Crawling time

V. Conclusion

It is concluded that the semantic focused crawler using ontology learning is better among all the existing methods. This semantic focused crawler is used for the purpose of precisely and efficiently and effectively discovering and formatting and indexing service information over the internet. It can be design a semi supervised approach Transductive Support Vector Machine by aggregating unsupervised approach and supervised approach. This approach is used to improve the perfprmance of crawler. According to the analysis, the proposed work perfoms better than the existing methods. It gives efficient results in terms of retrieving most relevant information and quality web pages from the web using semantic focused web crawler.

VI. References

[1] Lu LIU, Tao PENG "Clustering-based topical Web crawling using CFu-tree guided by link-context" in Higher Education Press and Springer-Verlag Berlin Heidelberg 2014

[2] Hai Dong, Farookh Khadeer Hussain, and Elizabeth Chang "Ontology-Learning-Based Focused Crawling for Online Service Advertising Information Discovery and Classification" in Springer-Verlag Berlin Heidelberg 2012

[3] Rodolfo Zunino, Roberto Surlinelli "An Analyst-Adaptive Approach to Focused Crawlers" in 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining

[4] Hai Dong, Farookh Khadeer Hussain "Self-Adaptive Semantic Focused Crawler for Mining Services Information Discovery" in IEEE Transactions On Industrial Informatics, Vol. 10, No. 2, May 2014

[5] Hardik P. Trivedi, Gaurav N. Daxini, Jignesh A. Oswal, Vinay D. Gor, Swati Mali "An Approach to Design Personalized Focused Crawler" in International Journal of Computer Science and Engineering Volume-2, Issue-3 E-ISSN: 2347-2693

[6] Bireshwar Ganguly, Devashri Raich "Performance Optimization of Focused Web Crawling Using Content Block Segmentation'' in 978-1-4799-2102-7/14 $31.00 © 2014 IEEE DOI 10.1109/ICESC.2014.69

[7] Hai Dong, Farookh Khadeer Hussain, Elizabeth Chang "A Transport Service Ontology based Focused Crawler" in 2008 IEEE

[8] R.Eswaramoorthy, M.Jayanthi "A Survey on Detection of Mining Service Information Discovery Using SASF Crawler" in International Journal of Innovative Research in Computer and Communication Engineering Vol. 2, Issue 10, October 2014

[9] Boser BE, Guyon IM, Vapnik VN. A training algorithm for optimal margin classifiers. Proceedings of the Fifth Annual Workshop on Computational Learning Theory, ACM: Pennsylvania, United States, 1992; 144-152.

[10] http://www.eclipse.org/

[11] http://www.csie.ntu.edu.tw/~cjlin/libsvm/

[12] Osmar R. Zaïane, "Introduction to Data Mining" in CMPUT690 Principles of Knowledge Discovery in Databases

[13] Trupti V. Udapure, Ravindra D. Kale, Rajesh C. Dharmik, "Study of Web Crawler and its Different Types" in IOSR Journal of Computer Engineering (IOSR-JCE)

[14] http://cacm.acm.org/blogs/blog-cacm/153780-data-mining-the-web-via-crawling/fulltext

[15] http://upload.wikimedia.org/wikipedia/commons/thumb/d/df/WebCrawlerArchitecture.svg/300px-WebCrawlerArchitecture.svg.png