RESEARCH PAPER | Engineering | Volume : 6 | Issue : 5 | May 2016 | ISSN - 2249-555X | IF : 3.919 | IC Value : 74.50

Design of Automatic Adaptive Semantic Focused Crawler for Mining Services

KEYWORDS: Mining services; Word Net; Crawler; machine learning; information discovery.

Gajanan V. Jaybhaye
M-tech Student at Computer Science and Engineering Department Government College of Engineering Amravati, India

Prof. Anil V. Deorankar
Associate Professor at Computer Science and Engineering Department Government College of Engineering Amravati, India

ABSTRACT *There is huge amount of data over the internet. We want to mined actual data or important data from the internet. But, there is some problems at the time of mining any kind of services or data that contains-heterogeneity, ubiquity and ambiguity. To tackle such a problem we have design automatic adaptive semantic focused crawler to mine any kind of services over the internet. This crawler will finding, arranging and indexing the data and it will increase the performance of the crawler. This structures combines the automatic adaptive semantic focused crawler with machine learning. This designed crawler will solve the issues related to the heterogeneity, ubiquity and ambiguity and machine learning will increase the performance of the crawler and perform prediction on data. Here, also given implementation of the project with their snapshots.*

I. INTRODUCTION

A significant number of people use Web search engines to formulate queries and review a list of suggested answers. Search engines are built from practical implementations of information retrieval techniques devised to handle large-scale Web collections. An increasing interest in the use of new specialized search engines has focused many efforts in the development of vertical search technologies [2].

Heterogeneity which provides diversity of services in the real world, Ubiquity in which service providers can be registered the service advertisements through various service registries. Ambiguity means amount of information present over the internet is described in natural language therefore it may be unclear [1].

Semantic Focused Crawler

A semantic focused crawler could assist us to solve the problem. Semantic focused crawlers are a subtype of the focused crawlers enhanced by various semantic web technologies with the purpose of crawling web documents under specified topics [10]. The emerging semantic focused crawlers can be primarily classified into two categories as follows:

The first category is ontology-based focused crawlers. These crawlers are able to utilize ontology to classify web documents by computing the similarity values between ontology concepts and descriptions of URLs of web documents [13, 15]. Courseware Watchdog was developed by Tane et al. [14], which has one special feature whereby users can specify their preferences on certain ontology concepts by assigning corresponding weights to the preferred concepts. Then the weights of concepts are aggregated with the similarity values between concepts and web documents in order to obtain user-preferred web documents.

The second category is metadata data abstraction crawlers. These crawlers are able to automatically generate metadata based on web contents by parsing web documents and annotating them with ontology markup languages [11,12].

This crawler is designed with the motive of helping search engines to precisely and capable of search mining service information by semantically finding, arranging, and indexing information [3].

Also, here we are using machine learning, Machine learning traverse the study and construction of algorithms that can learn from and make prediction on data [4].

II PROBLEM DEFINITION

In our paper we addressing the three major problem-heterogeneity, ubiquity and ambiguity. We propose the framework of a novel automatic adaptive semantic focused crawler, by combining the technologies of semantic focused crawling and machine learning. whereby semantic focused crawling technology is used to solve the issues of heterogeneity, ubiquity and ambiguity of mining service information and machine learning technology is useful to maintaining the high performance of crawling in the uncontrolled Web environment. Here, we proposed a crawler-is designed with the purpose of helping search engines to precisely and efficiently search mining service information by semantically discovering, formatting, and indexing information. Machine learning will perform the prediction on data and increase the performance of the crawler in uncontrolled network environment.

III. LITERATURE SURVEY

In this section we briefly describes the prevoius works-H. Dong et al.[1] proposed a self adaptive semantic focused crawler for mining services information discovery. It is based on ontology learning approach [5]. It uses the ontology as repository and generate the metadata [6].It has drawback regarding the performance of the self adaptive model did not completely meet expectations regarding the parameters of precision and recall. W. Wong et al.[7] proposed a crawler in which attention is towards the enhancing semantic focused crawling technologies by combining them with ontology learning technologies. It contains drawback ralating to the differentiation and dynamism. Dong et al.[8] proposed a crawler in which a large portion of the crawler in this space make utilization of ontology to speak to the information fundamentals themes and web

Volume : 6 | Issue : 5 | May 2016 | ISSN - 2249-555X | IF : 3.919 | IC Value : 74.50

archives.It has drawback regarding, the ontology based semantic focused crawler is that the crawling performance crucially depends on the quality of ontologies.The prime idea of this crawler is to construct an artificial neural network model to determine the relatedness during a web documents and an ontology.It does not have the function of classification. It cannot be used to resolve ontologies by enriching the vocabulary of ontologies [9].

IV. SYSTEM WORKFLOW

Now, we will explain the system workflow of the automatic adaptive semantic focused crawler step by step as shown in fig 1. The initial goals of this crawler include- to generate mining service metadata from web pages and to exactly associate between the semantically pertinent mining service concepts and mining service metadata with relatively low computing cost. In fig 1. system architecture of the proposed automatic adaptive semantic focused crawler is shown it is based on the machine learning approach.

The first step is preprocessing in which processing is done on word net, next step is crawling in which it will download the k web pages or k term from internet for further used.

Next steps are term extraction and term processing, in which term will extract from web and perform the processing on that term, if term get matched with existing word net then metadata generation and association take place otherwise algorithm based string matching will done and generate the new term with help of machine learning and put that keyword and their related information in mining service word net base and mining service metadata base database for further used. If the algorithm based string matching will not performed then that term will be filtered out.

Machine Learning

It performed prediction on data by using some algorithm. Also it focuses more on exploratory data analysis. Machine learning tasks include- unsupervised learning, supervised learning and reinforcement learning. It is a sub domain of computer science that evolved from the practice of pattern recognition and computational learning theory in artificial intelligence

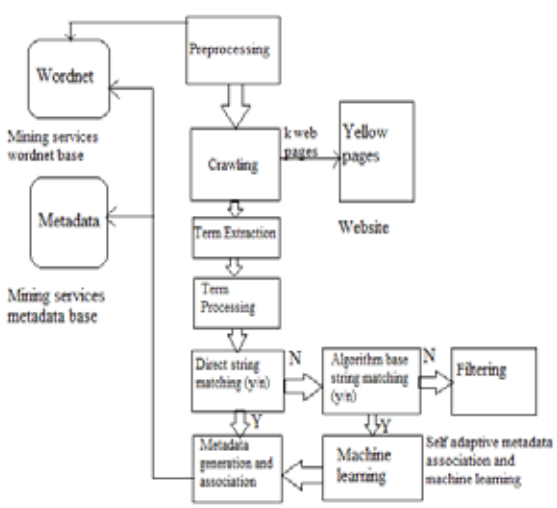

Fig 1: System architecture of the proposed automatic adaptive semantic focused crawler

V IMPLEMENTATION

In this section, we have design a web crawler which is shown in fig 2.

Fig 2: Screen shot of web crawler page

In this web crawler we have design number fields which contains home, edit profile, data mining, my dictionary, change password and log out field. Lets see one by one- First user must have to login using their user name and password, then on the basis of gender image will display on the screen whether login user is male or female, then the home page field in which user can go to the home page directly, after we can edit the profile of the user, and next field is data mining in which, which word is to mine is put up over there get the information of that word, next is my dictionary in which all the words in automatic adaptive dictionary will be shown over there, next is change password field in which we can change the password of the user and last is log out field, in which user can log out.

Fig 3. Screen shot of Retrieving data for word network

In fig 3 data is retrieved for word network. first of all it will check with our adaptive dictionary if suppose not found then that word is also check with wordnet and also it is not match then, then information for that services is searched

Volume : 6 | Issue : 5 | May 2016 | ISSN - 2249-555X | IF : 3.919 | IC Value : 74.50

from internet i.e. mainly from Wikipedia up to ten lines and other information from word net will filtered out.

VI. CONCLUSION

Here, we developed Automatic Adaptive Semantic Focused Crawler to mined any kind of services. It based on real time system to avoid Heterogeneity, Universality and Ambiguity. Also in snapshot we have shown any services are to be mined shown in fig 3. In which filtering is done on irrelevant data and up to ten line will shown of any services that you want mine. Because of this performance of the crawler increase more than previous one. Further, in future research, it is important to enrich the vocabulary of mining service word net by surveying those unmatched but relevant data, in order to improve the performance of the crawler.

References

[1] Hai Dong, member, IEEE, and Farookh Khadeer Hussain, "Self Adaptive Semantic Focused Crawler for Mining Services Information Discovery" IEEE Transactions on Industrial,Informatics,vol.10,No.2, pp.1616-1626, May 2014.

[2] C. H. Lovelock, "Classifying services to gain strategic marketing insights," *J. Marketing*, vol. 47, pp. 9–20, 1983.

[3] H. Dong, F. K. Hussain, and E. Chang, "A service search engine for the industrial digital ecosystems," *IEEE Trans. Ind. Electron.*, vol. 58, no. 6, pp. 2183–2196, Jun. 2011.

[4] Mining Services in the US: Market Research Report IBISWorld2011.

[5] H. Dong and F. K. Hussain, "Focused crawling for automatic service discovery, annotation, and classification in industrial digital ecosystems," *IEEE Trans. Ind. Electron.*, vol. 58, no. 6, pp. 2106–2116, Jun. 2011.

[6] J. L. M. Lastra and M. Delamer, "Semantic web services in factory automation: Fundamental insights and research roadmap," *IEEE Trans. Ind. Informat.*, vol. 2, no. 1, pp. 1–11, Feb. 2006.

[7] W. Wong, W. Liu, and M. Bennamoun, "Ontology learning from text: A look back and into the future," *ACM Comput. Surveys*, vol. 44, pp. 20:1–36, 2012.

[8] H. Dong, F. Hussain, and E. Chang, O. Gervasi, D. Taniar, B. Murgante, A. Lagana, Y. Mun, and M. Gavrilova, Eds., "State of the art in semantic focused crawlers," in *Proc. ICCSA 2009*, Berlin, Germany, vol. 5593, pp. 910–924,2009.

[9] P. Resnik, "Semantic similarity in a taxonomy: An information-based measure and its application to problems of ambiguity in natural language," *J. Artif. Intell. Res.*, vol. 11, pp. 95–130, 1999.

[10] H. Dong, F.K. Hussain, E. Chang, State of the art in semantic focused crawlers, in: Computational Science and Its Applications – ICCSA 2009, pp. 910–924, 2009

[11] E. Francesconi, G. Peruginelli, Searching and retrieving legal literature through automated semantic indexing, in: ICAIL' pp. 131–138, 07, 2007

[12] C.L. Giles, Y. Petinot, P.B. Teregowda, H. Han, S. Lawrence, A. Rangaswamy, N. Pal, eBizSearch: A niche search engine for e-business, in: SIGIR' pp. 213–214, 03, 2003.

[13] M. Halkidi, B. Nguyen, I. Varlamis, M. Vazirgiannis, THESUS: Organizing web document collections based on link semantics, VLDB J. 12 (2003) 320–332, 2003.

[14] J. Tane, C. Schmitz, G. Stumme, Semantic resource management for the web: An e-learning application, in: WWW2004, 2004, pp. 1–10, 2004.

[15] M. Yuvarani, N.C.S.N. Iyengar, A. Kannan, LSCrawler: A framework for an enhanced focused web crawler based on link semantics, in: The 2006 IEEE/WIC/ACM International Conference on Web Intelligence (WI'06), pp. 794–800, 2006.