International Journal of Scientific Engineering and Applied Science (IJSEAS) – Volume-2, Issue-5,May 2016
ISSN: 2395-3470
www.ijseas.com

Self Modifying Semantic Focused Crawler for Mining Services Knowledge Discovery

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

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

***Abstract*- Knowledge discovery is the process of finding new information from the internet, over the internet there is a tremendous amount of information available, we will extract only particular information from these set of huge information. In our research work, we will mined any particular word based information. First it will check with our self adaptive database, if information is already available then it will show the information otherwise it will fetch the information from word net and if the information related to that word is not found in word net also then it will take the information from the internet up to ten lines and finally, update this information in our self adaptive database for later used. In this way, in offline mode we can extract the information from our self adaptive database. The crawler we designed is used to solve the issue relating to the heterogeneity, ubiquity and ambiguity and machine learning will increase the performance of the crawler and also perform prediction on data.**

***Keywords*- Prediction; Knowledge Discovery; Crawler; Word net.**

I. INTRODUCTION

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]

Heterogeneity which provides diversity of services in the real world [2], 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].

The 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].

We propose the framework of a novel self modifying 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.

II. LITERATURE SURVEY

In this context 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 archives.It has drawback regarding, the ontology based semantic focused crawler is that the crawling performance crucially depends on the quality of ontologies.

III. SYSTEM WORKFLOW

Here, we will explain the system workflow of the self modifying 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 self modifying 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 and self adaptive database, next step is crawling and term extraction in which it will first input a query then crawling and term extraction [9] will be done on the internet , extract the k term from internet for further used.

Next steps is term processing [11], in which term will extract from internet and perform the processing on that term, if term get matched with first self adaptive database then metadata generation and association take place otherwise check with existing word net [12] 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 [13] and their related information in self adaptive database both these keywords and related information keep separately in database for further used. If the algorithm based string matching will not performed then that term will be filtered out [14]. Following are the algorithms used in this process- steps are given below-

A. *Self Adaptive Dictionary*

Step 1. Accept input word
Step 2. Connect with adaptive databases
Step 3. Send query
Step 4. If result generated then return result
else
return null.

B. *Word net*

Step1. Accept input word
Step 2. Retrieve data from word net
Step3.Preprocess data (Remove unwanted pattern)
Step 4. Remove HTML tag
Step 5. Return Result

C. *Internet*

Step 1. Accept input word
Step 2. Retrieve contents by crawling
Step3. Gather the links, if related data present in the links according to semantic similarity, retrieve it
Step 4. Remove invalid characters
Step 5. Remove HTML tag
Step 6. Return results

D. *Combine Algorithm*

Step 1. Accept input keyword
Step 2.Search it in adaptive word net (self adaptive dictionary)
Step 3. If word found then return
Step 4. If word not found then search it in word net
Step 5. If word found in word net then return
Step 6. If word not found in word net then search on net by crawling
Step 7. If word found on net then return
Step 8. If word not found then filter out.

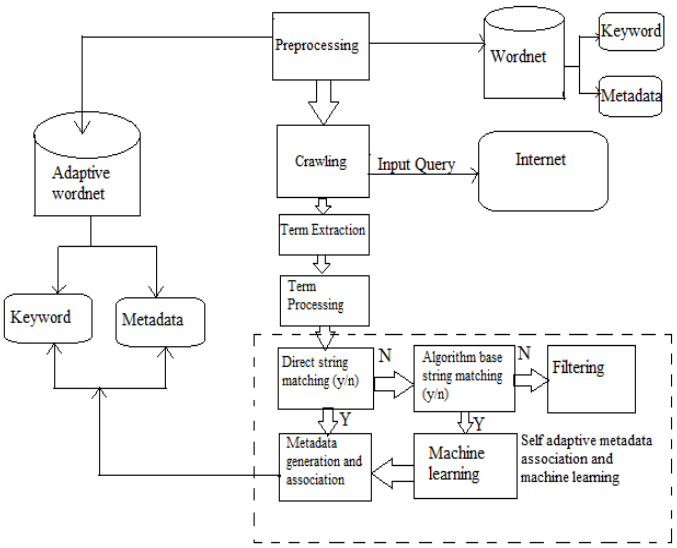
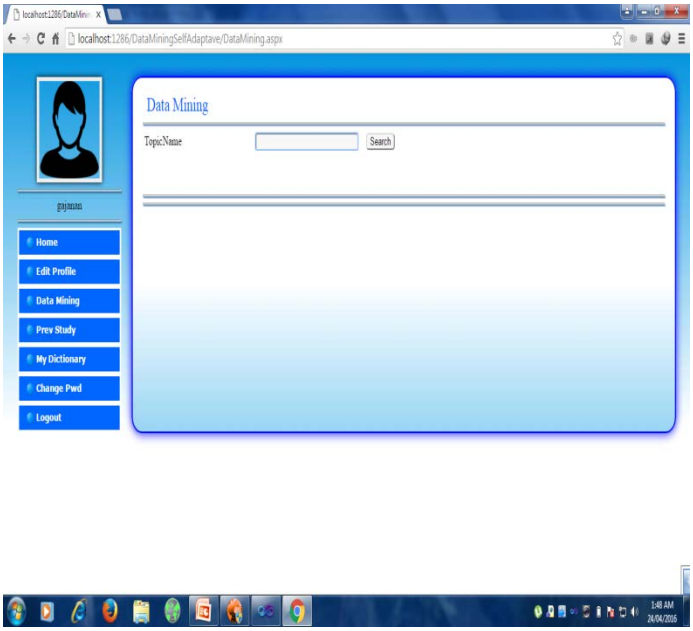


Fig 1: System architecture of the proposed self modifying semantic focused crawler

.

IV IMPLEMENTATION

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



International Journal of Scientific Engineering and Applied Science (IJSEAS) – Volume-2, Issue-5,May 2016
ISSN: 2395-3470
www.ijseas.com

Fig 2: Screen shot of web crawler page

In this web crawler we have design number fields which contains home, edit profile, data mining, previous study, 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 previous study field in which existing work is implemented, 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 Wikipedia

In fig 3 data is retrieved for word Wikipedia. first of all it will check with our adaptive dictionary if suppose not found then that word is also check with word net and also it is not match then, then information for that services is searched from internet i.e. mainly from Wikipedia up to ten lines and other information will be filtered out. If suppose, no word will be found on adaptive dictionary, word net and that on internet then that word will be filtered out.

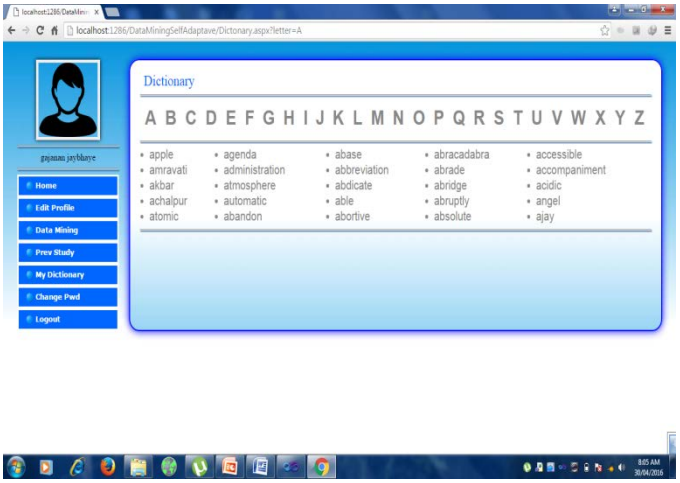


Fig 4. My dictionary

V. CONCLUSION

In this paper, we have design Self Modifying 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. In which filtering is done on irrelevant data and up to ten line will shown of any services that you want mine. Algorithms are used to fetch the data from adaptive dictionary, word net and internet it works efficiently. 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.