REVIEW ON SELF ADAPTIVE SEMANTIC INFORMATION BASED CRAWLERS FOR DATA MINING SERVICES

Gajanan V. Jaybhaye Computer Science and engg. Department GCOEA , gajanan.jaybhaye@gmail.com

Prof. Anil V. Deorankar Associate Professor, Department of CSE GCOEA, avdeorankar@gmail.com

***Abstract*- In today's world, online advertisements are very famous with number of industries, which includes mining service industry. Web crawler are of the most critical components used by the search engines to collect pages from the web. It is an intellectual technique of browsing used by the search engines. Service users may faces the three major problems- heterogeneity, ubiquity and ambiguity at time of searching for information over the internet. The framework of a self-adaptive semantic focused crawler i.e. SASF crawler have the purpose of precisely and efficiently discovering, formatting and indexing mining ploy information over the internet by taking into account these three major problems. This framework incorporates the technologies of semantic focused crawling and ontology learning to maintain the performance of this crawler. In this paper number of literature survey are taken into account with their drawback.**

Index Terms— *Service advertisements, Mining service, Ontology learning, SASF crawler.*

1 INTRODUCTION

The Internet has become the largest marketplace in the world and now days online advertisements are very popular with number of industries which includes mining service industries where, mining service publications are effective carriers of mining service information . Service users may faces three major problems- heterogeneity, ubiquity and ambiguity when searching for information over the internet. The framework of self-adaptive semantic focused crawler i.e. SASF crawler have the purpose of precisely and efficiently discovering, formatting and indexing mining service information onto the internet by taking into account the three major problem i.e. this framework assemble the technologies of semantic focused crawling and ontology learning to maintain the performance of this crawler indifferent of the variety in the web environment. The innovations lie in the design of an unsupervised framework for vocabulary basis ontology learning, and a hybrid algorithm for matching semantically relevant concepts and metadata [1].

Service publications form a considerable part of the advertising which takes place over the Internet and have the some features-

Heterogeneity- which provides diversity of services in the real world, there is number of schemes have been proposed to distribute the services from various perspectives, which includes ownership of service instruments, the effect of services, the nature of the service act, delivery, demand and supply and so on. But, there is not a publicly agreed scheme obtainable for classifying service advertisements over the Internet. Furthermore, while many commercial product and service search engines provide classification plan of services with the purpose of facilitating a search, they do not surely distinguish between the product and the service advertisement; instead, they combine both into one taxonomy [2].

Ubiquity-service providers can be registered the service advertisements through various service registries which includes global business search engines, such as Business.com2 and Kompass3 and other service registry is local business directories like Google Local Business Center4 and local Yellowpages5, also there is another registry domain specific business search engines such as healthcare, industry and tourism business search engines. These service registries are geographically distributed over the Internet [3].

Ambiguity-amount of information present over the internet is described in natural language therefore it may be unclear. Moreover, online service information does not have a consistent format and standard, and differs from Web page to Web page. Mining is one of the oldest industries in human history, having appeared with the beginning of human civilization. Mining services refer to a series of services which support mining, quarrying, and oil and gas extraction activities. Since the arrival of the information age, mining service companies have bethink the power of online advertising and they have attempted to promote themselves by actively joining the service advertising community [4].

In order to address the above problems the framework of a novel self-adaptive semantic focused (SASF) crawler, by mixing the technologies of semantic focused crawling and ontology learning is design, whereby semantic focused crawling technology is used to solve the issues of heterogeneity, ubiquity and ambiguity of mining ploy information, and ontology learning technology is used to maintain the high performance of crawling in the uncontrolled network environment. This crawler is designed with the purpose of helping search engines to precisely and capable of search mining service information by semantically discovering, formatting, and indexing information.

2. BASICS OF CRAWLER

A web crawler is an internet bot which systematically browses the world wide web, typically for the purpose of web indexing. A web crawler may also be called a web spider an ant, an automatic indexer or a web scutter.

Web search engines and some other sites uses web crawling or spidering software to update their web content or indexes of others sites web content. Web crawlers can duplicate all the pages they visit for later processing by a search engine which indexes the downloaded pages such the users can search much more efficiently. Crawlers can validate hyperlinks and HTML code, they can also be used for web scrapping.

A crawler must not only have a good crawling strategy, but it should too have a highly optimized architecture. Shkapenyuk and Suel noted that While it is fairly motile to build a dull crawler that downloads a few pages per second for a short period of time, building a high performance system that can download hundreds of millions of pages over several weeks presents a number of challenges in system design, I/O and network capability, and robustness and manageability. Web crawlers are a central part of search engines, and details on their algorithms and architecture are kept as business secrets. When crawler designs are published, there is often an massy lack of detail that prevents others from reproducing the work. There are also emerging concerns about search engine spamming, which prevent major search engines from publishing their ranking algorithms [5].

3. LITERATURE REVIEW

The semantic focused crawler approach is studied in various papers. I briefly introduce the fields of semantic focused crawling and review previous work on ontology learning based focused crawling. A semantic focused crawler is a software agent that is used to traverse the web and retrive as well as download related web information on specific topics by means semantic technologies [6].

Since semantic technologies provide shared knowledge for enhancing the interoperability during heterogeneous components, semantic technologies have been broadly applied in the field of industrial automation [7]. The goal of semantic focused crawlers is to precisely and efficiently retrieve and download relevant Web information by automatically grasping the semantics underlying the Web information and the semantics underlying the predefined topics.

H. Dong et al.[1] proposed a self adaptive semantic focused crawler for mining services information discovery. It is based on ontology learning approach. It uses the ontology as repository and generate the metadata.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.[8] 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.[9] 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.

Zheng et al.[10] proposed a supervised ontology learning based focused crawler that aims to maintain the harvest rate of the crawler in the crawling process. The main idea of this crawler is to construct an artificial neural network model to determine the relatedness between a web documents and an ontology.It does not have the function of classification. It cannot be used to evolve ontologies by enriching the vocabulary of ontologies. The supervised learning may not work within an uncontrolled network environment with unpredicted new terms.

Kang et al.[10] proposed a crawler this approach is able to distribute web documents by means of the concepts in an ontology to learn the weights of relations between concepts and to work in an uncontrolled network environment. It is not able to really evolve ontologies by enriching their contents, namely their vocabularies.

C. su et al.[11] proposed an unsupervised ontology learning based focused crawler in order to enumerate the relevance scores between topics and web documents. Given a specific domain ontology and a topic represented by a concepts in this ontology, the relevance score between a web documents and the topic is weighted sum of the occurrence frequencies of all the concepts at the ontology in the web documents. Also this crawler makes use of reinforcement learning, which is probabilistic framework for learning optimal decision making from rewards or punishments [12], in order to train the weight of each concept. It has drawback like, it cannot be used to enrich the vocabulary of ontologies.

4. CONCLUSION

In the above papers the semantic focused crawler approach is given. They have given the innovative ontology learning based focused crawler – the SASF crawler, for service information discovery in the mining service industry, by appealing into account the heterogeneous, ubiquitous and ambiguous nature of mining service information available onto the Internet. This approach involved an innovative unsupervised ontology learning framework for vocabulary-based ontology learning, and a new concept-metadata matching algorithm. The reviews are given above with its drawbacks. Every paper proposes a method or a theoretical approach which has some drawbacks in it. They don't focus on some aspects which can be useful for it. So the drawbacks are given which can be their future work.

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] https://webcrawler-wikipedia,the free encyclopedia.html.

[6] 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.

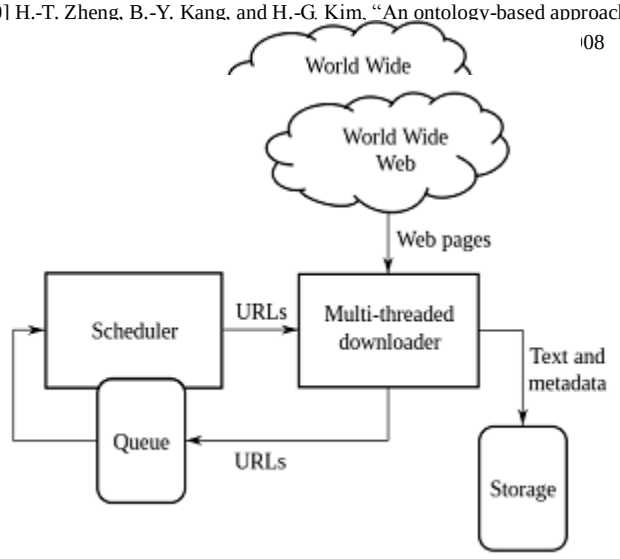
[7] J. L. M. Lastra and M. Delamer, "Semantic web services in factory automation: Fundamental insights and research roadmap," *IEEE Trans. Ind. Informat.*, vol. 2, no. 1, pp. 1–11, Feb. 2006.

[8] 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.

[9] 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.

[10] H.-T. Zheng, B.-Y. Kang, and H.-G. Kim, "An ontology-based approach to

08



rawler
. *(HIS*

e Web
l, Slo-