ISSN 2395-1621

Ontology Based Self-Adaptive Crawler for Mining Services Information Discovery

Vaibhav Nangare[#1], Kiran Shirsath[#2], Vishakha Lothe[#3]

[1]vaibhavnangare5@gmail.com

[2]kirans3500@gmail.com

[#123]Dept. of Computer Engineering, G.S.M.C.O.E.
University of Pune, Maharashtra, India.

ABSTRACT

It is well known that the Internet has become the best marketplace all over the world, and online advertising is well known with so many industries, including the traditional mining service industry where mining service advertisements are effective carriers of mining information about the services. However, all users may encounter three major issues – variety, universality, and uncertainty, when searching for mining service information over the Internet. In this paper, we are presenting the framework of a novel self-adaptive crawler– SASF crawler, with the intend of strictly and easily finding, changing, and listing mining service information over the Internet, by considering the three main issues. This framework includes the technologies of semantic focused crawling and ontology learning, in order to preserve the enforcement of this crawler, unconcerned of the difference in the Web environment. The purpose of this research lies in the design of an unsupervised framework for vocabulary-based ontology discovering. Algorithm for matching semantically related concepts and metadata. A list of experiments is conducted in order to evaluate the proper working of this crawler. The proposed work and the future work are given in the final section.

***Keyword*- Mining service industry, ontology learning, semantic focused crawler, service advertisement, service information discovery.**

ARTICLE INFO

Article History

Received : 30th March, 2015

Received in revised form :

2nd April, 2015

Accepted : 5th April, 2015

Published online :

8th April 2015

I. INTRODUCTION

The internet has becoming the large number of unstructured data for gaining access to information over the documents. With the very quick progress of electronic text from the complex the WWW, more and more knowledge you need is added. But, the large amount of text also gives so much problem to people to determine useful information. For example, the standard Web search engines have low accuracy, since typically some related Web pages are returned with mixed different pages, which is mainly because of situation that the topic-related features may be placed in different contexts. So, one proper way of organizing this overwhelming amount of documents is necessary. The World Wide Web is an architectural framework for accessing linked documents spread out over millions of machines all over the Internet.

A crawler is a software program used to create search engine index entries by visiting Web sites. It systematically reads the web pages and retrieves information from those pages for web indexing. Web crawling software is used in web search engines to update the web content of the web sites. Web search engine indexes the downloaded pages by crawler for later usage by the user, making the search quicker. Web Crawler is incessant running programs which download pages at regular intervals from internet. For assembling Web content locally, crawlers are used as tools. Web crawlers are used in many applications where large number of pages is quickly fetched into a local repository and is indexed based on keywords. Since crawlers extract information from web sites, they are used in Web Scrapping.

A semantic focused crawler is s software agent that is able to traverse the Web, and download the related information related web information on specific topics. Since the semantic technology provide shared knowledge for enhancing the interoperability between heterogeneous component and semantic technologies have been broadly applied in the field of industrial automation. The main goal of semantic focused crawlers is to efficiently and precisely retrieve and download relevant Web information automatically understanding the semantic underlying the predefined topics. The survey conducted by Dong et al. found that most crawler in domain use of ontology to understanding topics and Web information. The main issue

in ontology-based semantic crawler is that the performance is depends on the quality of ontology. Ontology mainly affected by two things. The first is that ontology is the formal representation of specific domain knowledge. Ontology designed by expert, a inconsistency between the domain experts' understanding of domain knowledge and the real worlds' domain knowledge. The second issue is that knowledge is dynamic and is continuously evolving, compared to related static ontology. These two contradictory situations could lead to the problem that ontology sometimes cannot precisely represent real-world knowledge.

To enhance and maintain semantic-based crawler and solve the defect in ontology, need to pay attention on enhancing semantic-based-focused crawling technologies by integrating them with ontology learning technologies. The goal of ontology learning is to semi-automatically extract facts or patterns from a corpus of data and turn these into machine-readable ontology.

Various techniques have been designed for ontology learning obviously, ontology-learning-based techniques can be used to solve the issue of semantic-focused crawling, by learning these new knowledge from crawled documents and integrating the new knowledge with ontology in order to constantly understand the new ontology technologies is necessary task.

II. Organization

The paper is organized as follows: Related work is presented in Section II. We present our scheme in Section III. The Mathematical background in section IV.The Future works in Section V. We conclude in Section VI.

III. Related Work

In this section, we are introducing the fields of semantic focused crawling and ontology-learning-based focused crawling, and review previous work on ontology learning-based focused crawling.

IV. Existing System

The limitation of the ontology-based semantic focused crawlers is that the crawling performance crucially depends on the quality of ontology's. Furthermore, the quality of ontology's may be affected by two problems. The first is that, as it is well known that ontology is the formal representation of specific domain knowledge and ontology's are designed by domain experts, a conflict may exist between the domain experts' understanding of the domain knowledge and the domailn description that stay in the real world. The second one is that knowledge is changing or active and is constantly evolving, compared with relatively static ontology's. These two contradictory situations could lead to the problem that ontology's sometimes cannot precisely represent real-world knowledge, considering the problem of difference and dynamism.

The observation of this problem in the field of semantic focused crawling is that the ontology's used by semantic focused crawlers cannot accurately give the knowledge given in web information,since Web is widely created or updated by human users with different knowledge understandings, and human users are productive learners of new knowledge. The concluding result of this problem is reflected in the gradually descending curves in the performance of semantic focused crawlers.

V. PROPOSED METHODOLOGY

The primary goals of this crawler include: 1) to generate mining service metadata from Web pages; and 2) to precisely associate between the semantically relevant mining service concepts and mining service metadata with relatively low computing cost.

The second goal is realized by: 1) measuring the semantic relatedness between the concept Description and learned Concept Description property values of the concepts and the service Description property values of the metadata; and 2) automatically learning new values, namely descriptive phrases, for the learned Concept Description properties of the concepts.

It uses a novel concept-metadata semantic similarity algorithm to judge the semantic relatedness between concepts and metadata in the algorithm-based string matching process. The major goal of this algorithm is to measure the semantic similarity between a concept and a service information. This algorithm uses a hybrid pattern by aggregating a semantic-based string matching (SeSM) algorithm and a statistics-based string matching (StSM) algorithm

VI. PROPOSED Architecture

The Objective of this project

- To retrieve the data from global repository using web crawler
- To create ontology database
- To display more personalized result in the browser

VII. Overall Flow of Proposed System

The Fig.1.shows how proposed system will performs the operations.

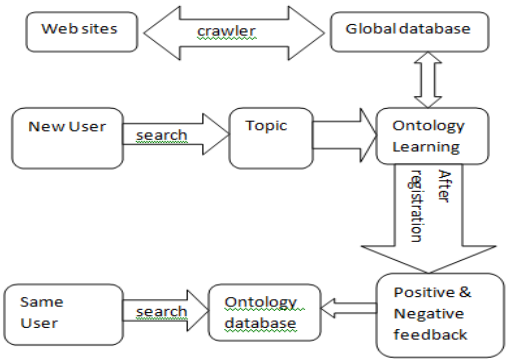

Fig. 1 Propose Architecture

This model is the implementation of the proposed ontology model. The input to this model was a topic and the output was a retrieval of data for the particular search term based on the user profile. For example in the local database if a person has given his/her interest under cricket as Indian team when that particular person searches in the global

database it will retrieve the accurate content about the Indian teams instead of retrieving about the contents about the cricket as a common search. In his proposed model query execution time and navigation cost is reduced.

VIII. MATHEMATICAL BACKGROUND

This uses a graph partitioned clustering algorithm to group users. A directionless graph based on the linked between each pair of web pages that are we used. We are taking into consideration each and every edge of the graph for giving a weight. Association Time used to tally the degree of visit ordering for each two pages in a session.

Fig.2 Equation (1)

Ti is the time duration of ith session that uses both a,b pages, we are using Tab to handle only inequality between a , b pages requested time in the currently ongoing time, f(k)=k if there are web page on that location k. Frequency measures the occurrence of two pages in each sessions(equation 2).

Fig.3 Equation (3)

Where Nab gives how many number of sessions using both page a and also for b page. We have given Na and also Nb for the different session of a page also for page b. Both the formulas select all values to time and frequency is between 0 and 1. Both these are tested as two clues of the degree of connectivity for each pair of web pages and is calculated using Equation(3).

Fig.4 Equation (4)

The data structure can be used to store the weights is an adjacency matrix M where each entry Mab contains the value Wab computed as given in the (3) for limiting the number of edge for these type of graph ,form of Mab whose value is less than a threshold are too little correlated and thus we are not using. This threshold is named as MinFreq in this contribution.

Algorithm:

Step 1: L[p] = A; // Assign all URLs to a list of web pages.

Step 2:For each (Ai, Aj) L[p] do //every web page pair

Step 3: M(i, j) = Weight Formula(Ai, Aj); //computing the weight based on Equation (3)

Step 4:Edge (i, j_) = M (i, j); End for For all Edge (a, b) Graph (E, V) do //removing all edges that its weight is below than MinFreq

Step 5: If Edge (a, b) < MinFreq then
Remove (Edge (a, b));
End if End for all vertices (a) do Cluster[i] = DFS (a);
//used for DFS If cluster[i] < MinClusterSize
//delete cluster whose length is below MinClusterSize

Step 5: Delete(Cluster[i]);
End if i = i + 1 end for return (Cluster) .

IX. FUTURE WORK

We describe a limitation of this approach and our future work as follows in the judgment phase, it can be clearly seen that the throughput of the self-adaptive model did not completely meet our expectations regarding the parameters of accuracy and remembrance. We figured out two reasons that caused this problem as follows.

Firstly, in this research, we try to search a universal threshold value for the concept-metadata semantic similarity algorithm in order to set up a boundary for determining concept metadata significance. However, in order to get the optimal performance, each concept must have its own exact outer limit, namely specific threshold values, for the judgment of the similarities.

Consequently, in future research, we aims to design a semi-supervised approach by combining the unsupervised approach and the supervised ontology learning-based approach, with the intend of automatically choosing the optimal threshold values for each concept, while maintaining the optimal performance without knowing the limitation of the training data set. Secondly, the relevant service information for each concept are manually found through a peer-reviewed process; i.e., many related service and concept information are determined on the basis of general knowledge, which cannot be judged by string similarity or term co-occurrence. Hence, in our future research, it is needed to add the vocabulary of the mining service ontology by surveying that unmatched but related service information, in order to further improve the performance of the Self-Adaptive crawler.

X. ADVANTAGES

In this project, It has a mining service ontology and a mining service metadata schema to solve the problem of self-adaptive discovery of data related to service for the mining industry.

This method allows the crawler to work in an uncontrolled environment where the numerous new terms and ontology's used by the crawler have a limited range of vocabulary.

XI. CONCLUSION

In this paper, we presented an innovative ontology-learning based focused crawler – the Self-Adaptive crawler, for appropriate information analysis in the mining service industry, by considering variety, universality, and uncertainty nature of mining service information available over the Internet. This method involves creative unsupervised ontology learning framework for vocabulary based ontology learning, and a new concept-metadata algorithm, which adds together a semantic-similarity based SeSM algorithm and a probability-based StSM algorithm for associating semantically relevant mining service concepts and mining service metadata. This method allows the crawler to work in an uncontrolled environment where the numerous new terms and ontology's used by the crawler have a limited range of glossary. Subsequently, we perform a series of tests to empirically analyse the performance of the Self-Adaptive crawler, by inspecting the performance of this approach with the existing approaches based on the six parameters adopted from the IR field.

XII. ACKNOWLEDGEMENT

To prepare this survey paper, I would like to be very thankful to my project guide Prof. Ashok Kumar, our Co-

ordinator Prof. Shrinivas And Head of the Department Prof Ratnaraj in Computer Department of Genba Sopanrao Moze College Of Engineering Affiliated to Savitribai Phule University. I would also like to thank the whole IEEE organization who helps allot to search various research papers related to my research. Because of their support only I am able to complete my research note.

REFRENCES

[1] H. Wang, M. K. O. Lee, and C. Wang, "Consumer privacy concerns about Internet marketing," *Commun. ACM*, vol. 41, pp. 63–70, 1998.

[2] R. C. Judd, "The case for redefining services," *J. Marketing*, vol. 28, pp. 58–59, 1964.

[3] T. P. Hill, "On goods and services," *Rev. Income Wealth*, vol. 23, pp. 315–38, 1977.

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

[5] 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 .

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

[7] B. Fabian, T. Ermakova, and C. Muller, "SHARDIS – A privacy-enhanced discovery service for RFID-based product information," *IEEE Trans. Ind. Informat.*, to be published .

[8] H. L. Goh, K. K. Tan, S. Huang, and C. W. d. Silva, "Development of Bluewave: A wireless protocol for industrial automation," *IEEE Trans. Ind. Informat.*, vol. 2, no. 4, pp. 221–230, Nov. 2006 .

[9] M. Ruta, F. Scioscia, E. D. Sciascio, and G. Loseto, "Semantic-based enhancement of ISO/IEC 14543–3 EIB/KNX standard for building automation," *IEEE Trans. Ind. Informat.*, vol. 7, no. 4, pp. 731–739, Nov. 2011 .

[10] I. M. Delamer and J. L. M. Lastra, "Service-oriented architecture for distributed publish/subscribe middleware in electronics production," *IEEE Trans. Ind. Informat.*, vol. 2, no. 4, pp. 281–294, Nov. 2006.

[11] 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 .

[12] H. Dong, F. K. Hussain, and E. Chang, "A framework for discovering and classifying ubiquitous services in digital health ecosystems," *J. Comput. Syst. Sci.*, vol. 77, pp. 687–704, 2011 .

[13] 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.

