International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064
Index Copernicus Value (2013): 6.14 | Impact Factor (2013): 4.438

Mining Service Using Ontology Learning

R. Santhosh Gowtham[1], S. J. Vivekanandan[2]

[1]M.Tech CSE II Year, Department of Computer Science, PRIST University, Thanjavur – 613 403

[2]Research Scholar, Sathyabama University, Chennai Assistant Professor, Department of Computer Science, PRIST University, Thanjavur – 613 403

Abstract: ***It is well recognized that the Internet has become the largest marketplace in the world, and online advertising is very popular with numerous industries, including the traditional mining service industry where mining service advertisements are effective carriers of mining service information. However, service users may encounter three major issues – heterogeneity, ubiquity, and ambiguity, when searching for mining service information over the Internet. In this paper, we present the framework of anovel 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. This framework incorporates the technologies of semantic focused crawling and ontology learning, in order to maintain the performance of this crawler, regardless of the variety in the Web environment. The innovations of this research lie in the design of an unsupervised Framework for vocabulary-based ontology learning, and a hybrid Algorithm for matching semantically relevant concepts and metadata. A series of experiments are conducted in order to evaluate the performance of this crawler.***

Keywords: Mining service industry, ontology learning, semantic focused crawler, service advertisement, service, Information discovery

1. Introduction

It is well recognized that information technology has a profound effect on the way business is conducted, and the Internet has become the largest marketplace in the world. It is estimated that there were over 2 billion Internet users in 2011,with an estimated annual growth of over 16%, compared with 360 million users in 20001. Innovative business professionals have realized the commercial applications of the Internet both for their customers and strategic partners, turning the Internet into an enormous shopping mall with a huge catalogue. Consumers are able to browse a huge range of products and service. advertisements over the Internet, and buy these goods directly through online transaction systems [1]. Service advertisements form a considerable part of the advertising which takes place over the Internet and have the following features:

A. Heterogeneity

Given the diversity of services in the real world, many schemes have been proposed to classify the services from various perspectives, including the ownership of service instruments [2], the effects of services [3], the nature of the service act, delivery, demand and supply [4], and so on. Nevertheless, there is not a publicly agreed scheme available for classifying service advertisements over the Internet. Furthermore, whilst many commercial product and service search engines provide classification schemes of services with the purpose of facilitating a search, they do not really distinguish between the product and the service advertisement; instead, they combine both into one taxonomy.

B. Ubiquity

Service advertisements can be registered by service providers through various service registries, including 1) global business search engines, such as Business.com2 and Kompass3, 2) local business directories, 3) domain-specific business search engines, such as healthcare, industry and tourism business search engines, and 4) search engine advertising, such as [5]. These service registries are geographically distributed over the Internet.

C. Ambiguity

Most of the online service advertising information is embedded in a vast amount of information on the Web and is described in natural language, therefore it may be ambiguous. Moreover, online service information does not have a consistent format and standard, and varies from Web page to Web page. Mining is one of the oldest industries in human history, having emerged with the beginning of human civilization. Mining services refer to a series of services which support mining, quarrying, and oil and gas extraction activities. Australian GDP between 2007 and 2008, to which the field of mining services contributed 7.65% [6]. Since the advent of the information age, mining service companies have realized the power of online advertising, and they have attempted to promote themselves by actively joining the service advertising community. It was found that nearly 50,000 companies worldwide have registered their services on the Compass website. However, these mining service advertisements are also subject to the issues of heterogeneity, ubiquity and ambiguity, which prevent users from precisely and efficiently searching for mining service information over the Internet. Service discovery is an emerging research area in the domain of industrial informatics, which aims to automatically or semi-automatically retrieve services or service information in particular environments by means of various IT methods.

Many studies have been carried out in the environments of wireless networks [7]–[9] and distributed industrial systems [10]. However, few studies have been planned for industrial service advertisement discovery in the Web environment, by taking into account the heterogeneous, ubiquitous and ambiguous features of service advertising information.

In order to address the above problems, in this paper, we propose the framework of a novel self-adaptive semantic focused (SASF) crawler, by combining the technologies of

semantic focused crawling and ontology learning, whereby semantic focused crawling technology is used to solve the issues of heterogeneity, ubiquity and ambiguity of mining service information, and ontology learning technology is used to maintain the high performance of crawling in the uncontrolled Web environment. This 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. The rest of this paper is organized as follows: in Section II, we review the related work in the field of ontology-learning-based focused crawling and address the research issues in this field; in Section III, we present the framework of the SASF crawler, including the mining service ontology and metadata schema and workflow of this crawler.

In Section IV, we deliver a hybrid concept-metadata matching algorithm to help this crawler semantically index the mining service information; in Section V,we conduct a series of experiments in order to empirically evaluate the framework of the crawler; and in the final section, we discuss the features and the limitations of this work and propose of our future work.

- Our previous research work created a purely semantic focused crawler, which does not have an ontology-learning function to automatically evolve the utilized ontology. This research aims to remedy this shortcoming.
- Our previous work utilized the service ontology's and the service metadata formats, especially designed for the transportation service domain and the health care service domain.

In this research, we design a mining service ontology and a mining service metadata schema to solve the problem of self-adaptive service information discovery for the mining service industry.

An overview of the system architecture and the workflow is shown in Fig. 1. As can be seen, the SASF crawler consists of two knowledge bases – a Mining Service Ontology Base and a Mining Service Metadata Base, and a series of processes, as well as a workflow coordinating these processes. The following figures shows the full architecture and workflow of proposed self adaptive semantic focused crawler and also the mining service ontology.

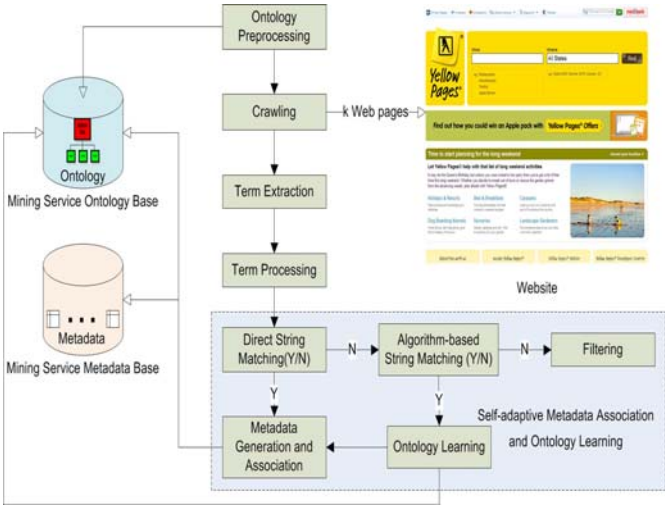

Figure 1: System Architecture

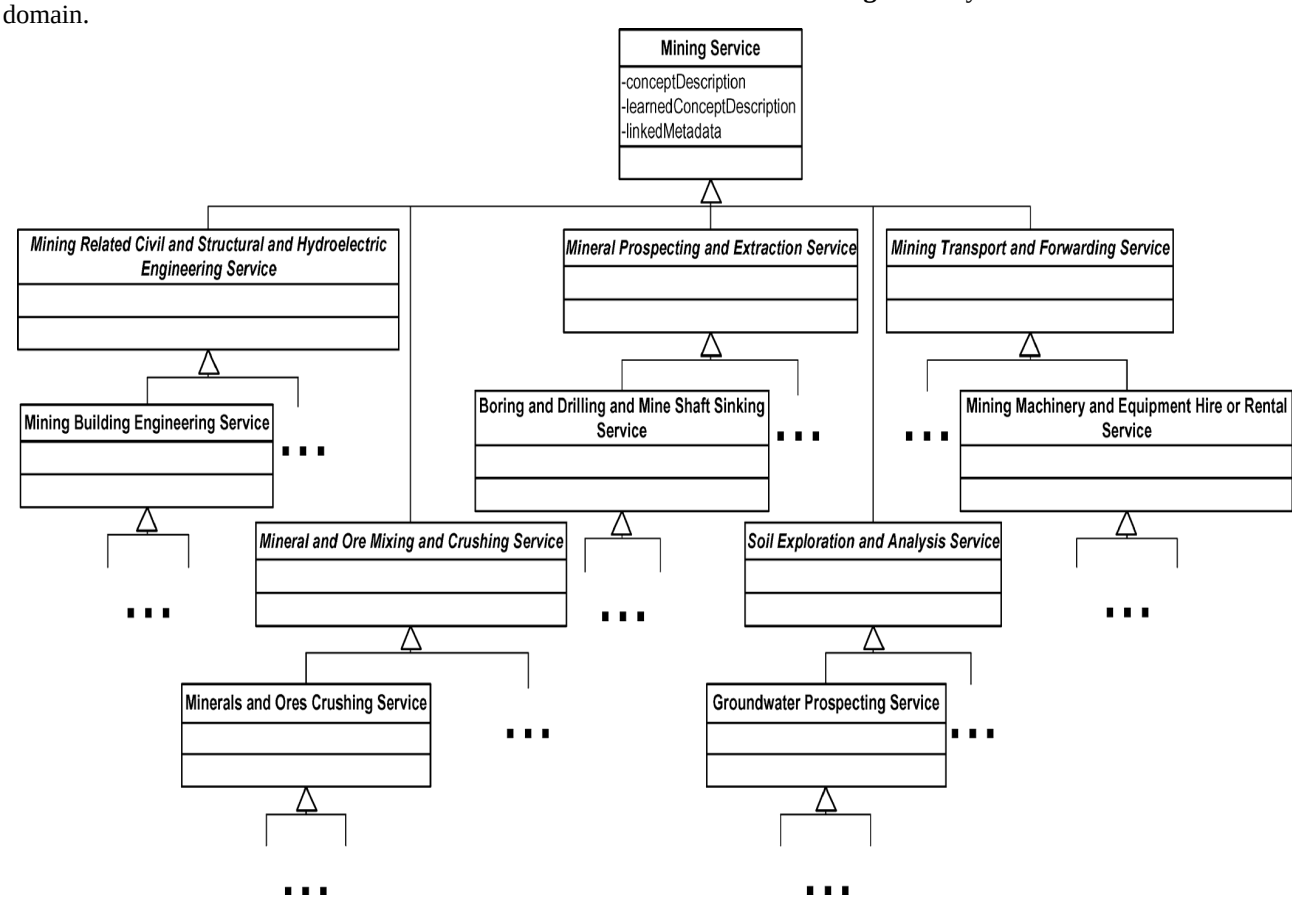

Figure 2: The Mining Service Ontology

2. Related Work

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

A semantic focused crawler is a software agent that is able to traverse the Web, and retrieve as well as download related Web information on specific topics by means of semantic technologies [11], [12]. Since semantic technologies provide shared knowledge for enhancing the interoperability between heterogeneous components, semantic technologies have been broadly applied in the field of industrial automation [13]–[15]. The goal of semantic focused crawlers is to precisely and efficiently retrieve and download relevant Web information by automatically understanding the semantics underlying the Web information and the semantics underlying the predefined topics. A survey conducted by Dong *et al.* [16] found that most of the crawlers in this domain make use of ontology's to represent the knowledge underlying topics and Web documents.

However, 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 issues. The first issue is that, as it is well known that an ontology is the formal representation of specific domain knowledge [17] and ontology's are designed by domain experts, a discrepancy may exist between the domain experts' understanding of the domain knowledge and the domain knowledge that exists in the real world. The second issue is that knowledge is dynamic 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 issues of differentiation and dynamism.

The reflection of this problem in the field of semantic focused crawling is that the ontology's used by semantic focused crawlers cannot precisely represent the knowledge revealed in Web information, since Web information is mostly created or updated by human users with different knowledge understandings, and human users are efficient learners of new knowledge. The eventual consequence of this problem is reflected in the gradually descending curves in the performance of semantic focused crawlers.

In order to solve the defects in ontology's and maintain or enhance the performance of semantic-focused crawlers, researchers have begun to pay attention to enhancing semantic- 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's [18]. Various techniques have been designed for ontology learning, such as statistics-based techniques, linguistics (or natural language processing)-based techniques, logic-based techniques, etc. These techniques can also be classified into supervised techniques, semi-supervised techniques, and unsupervised techniques from the perspective of learning control.

Obviously, ontology-learning-based techniques can be used to solve the issue of semantic-focused crawling, by learning new knowledge from crawled documents and integrating the new knowledge with ontology's in order to constantly refine the ontology's. In the rest of this section, we will review the two existing studies in the field of ontology learning-based semantic focused crawling.

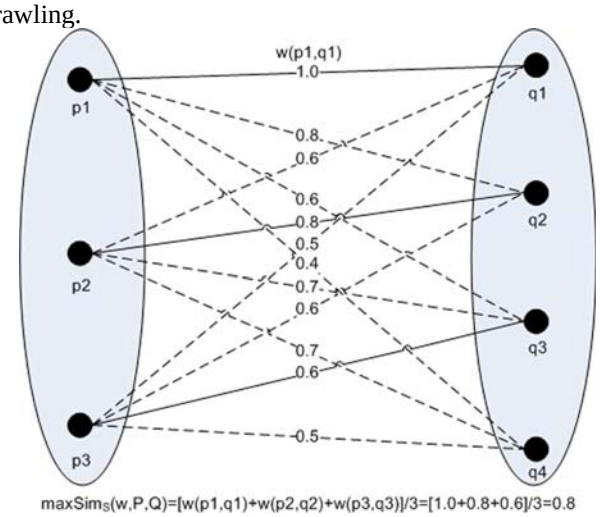

Figure 3: Graphical representation of the assignment in the bipartite graph Problem

3. Proposed Method

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 description and a service description. This algorithm follows a hybrid pattern by aggregating a semantic-based string matching (SeSM) algorithm and a statistics-based string matching (StSM) algorithm.

3.1 System Evaluation

One common defect of the existing ontology-learning-based focused crawlers is that these crawlers are not able to work in an uncontrolled Web environment with unpredicted new terms. In this section, we will evaluate our SASF crawler by comparing its performance with that of the existing

ontology- learning-based focused crawlers of Zhen *et al.* and Su*et al.*

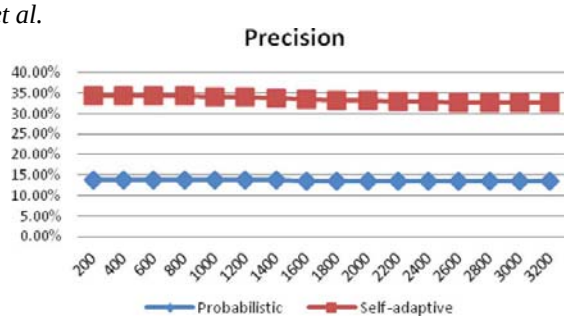

Figure 4: Comparison of the ontology-learning-based focused crawling models on precision

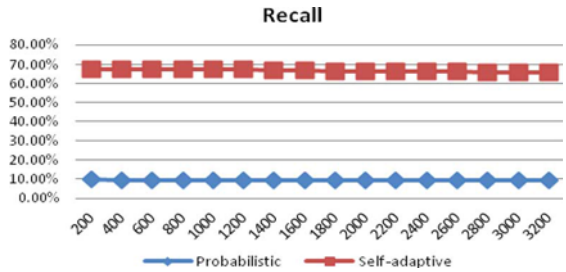

Figure 5: Comparison of the ontology-learning-based focused crawling models on recall

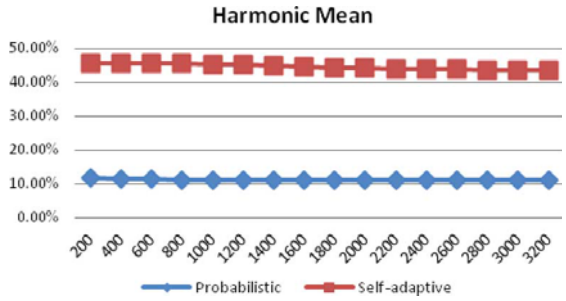

Figure 6: Comparison of the ontology-learning-based focused crawling models on harmonic mean

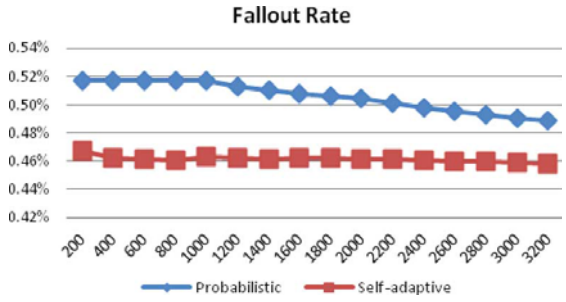

Figure 7: Comparison of the ontology-learning-based focused crawling models on fallout rate

3.2 Precision

The graphic representation of the comparison of the probabilistic and self-adaptive models on precision, along with the increasing number of visited Web pages, is shown in Fig 4. It can be observed that the overall precision of the self adaptive model is 32.50%, and the overall precision of the probabilistic model is 13.46%, which is less than half of that of the former. This is because the self-adaptive model is able to filter out more non-relevant mining service Web pages based on its vocabulary-based ontology learning function. This proves that the self-adaptive model significantly enhances the preciseness of semantic focused crawling.

3.3 Recall

The graphic representation of the comparison of the probabilistic and self-adaptive models on recall, along with the increasing number of visited Web pages, is shown in Fig 5. It can be seen that the overall recall for the self-adaptive model is 65.86%, compared to only 9.62% for the probabilistic model. This is because the self-adaptive model is able to generate more relevant mining service metadata based on its vocabulary-based ontology learning function, and thus improves the effectiveness of the semantic focused crawler.

3.4 Harmonic Mean

The graphic representation of the comparison of the probabilistic and self-adaptive models on harmonic mean, along with the increasing number of visited Web pages, is shown in Fig 6. As an aggregated parameter, the overall harmonic mean values for both of these models are below 50% (11.22% for the probabilistic model, and 43.51% for the self-adaptive model), due to their low performance on precision. Since the self-adaptive model outperforms the probabilistic model on both precision and recall, it is not surprising that the overall harmonic mean of the former is nearly four times as high as that of the latter.

3.5 Fallout Rate

The graphic representation of the comparison of the probabilistic and self-adaptive models on fallout rate, along with the increasing number of visited Web pages, is shown in Fig 7. It can be seen that the overall fallout rate of the self-adaptive model is 0.46%, and for the probabilistic model is around 0.49%. This indicates that the former generates fewer false results than the latter, which proves the low inaccuracy of the self-adaptive model.

4. Future Work

We describe a limitation of this approach and our future work as follows in the evaluation phase, it can be clearly seen that the performance of the self-adaptive model did not completely meet our expectations regarding the parameters of precision and recall. We deduce two reasons that caused this issue as follows.

Firstly, in this research, we try to find a universal threshold value for the concept-metadata semantic similarity algorithm in order to set up a boundary for determining concept metadata relatedness. However, in order to achieve optimal performance, each concept should have its own particular boundaries, namely particular threshold values, for the judgment of the relatedness.

Consequently, in future research, we intend to design a semi-supervised approach by aggregating the unsupervised approach and the supervised ontology learning-based

approach, with the purpose of automatically choosing the optimal threshold values for each concept, while keeping the optimal performance without considering the limitation of the training data set. Secondly, the relevant service descriptions for each concept are manually determined through a peer-reviewed process; i.e., many relevant service descriptions and concept descriptions are determined on the basis of common sense, which cannot be judged by string similarity or term co-occurrence. Hence, in our future research, it is necessary to enrich the vocabulary of the mining service ontology by surveying those unmatched but relevant service descriptions, in order to further improve the performance of the SASF crawler.

4.1 Advantages

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

5. Conclusion

In this paper, we presented an innovative ontology-learning based focused crawler – the SASF crawler, for service information discovery in the mining service industry, by taking into account the heterogeneous, ubiquitous and ambiguous nature of mining service information available over the Internet. 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 ontology's used by the crawler have a limited range of vocabulary. Subsequently, we conduct a series of experiments to empirically evaluate the performance of the SASF crawler, by comparing the performance of this approach with the existing approaches based on the six parameters adopted from the IR field.

References

[1] H. Wang, M. K. O. Lee, and C. Wang, "Consumer privacy concerns about Internet marketing," *Common. 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.
[14] S. Runde and A. Fay, "Software support for building automation requirements engineering—An application of semanticweb technologies in automation," *IEEE Trans. Ind. Informat.*, vol. 7, no. 4, pp. 723–730,Nov. 2011.
[15] M. Ruta, F. Scioscia, E. Di 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.
[16] 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, 2009, vol. 5593, pp. 910–924.
[17] T. R. Gruber, "A translation approach to portable ontology specifications," *Knowledge Acquisition*, vol. 5, pp. 199–220, 1993.
[18] 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.
[19] H.-T. Zheng, B.-Y. Kang, and H.-G. Kim, "An ontology-based approach to learnable focused crawling," *Inf. Sciences*, vol. 178, pp. 4512–4522, 2008.
[20] C. Su, Y. Gao, J. Yang, and B. Luo, "An efficient adaptive focused crawler based on ontology learning," in *Proc. 5th Int. Conf. Hybrid Intell. Syst. (HIS '05)*, Rio de Janeiro, Brazil, 2005, pp. 73–78.

[21] J. Rennie and A. McCallum, "Using reinforcement learning to spider the Web efficiently," in *Proc. 16th Int. Conf. Mach. Learning (ICML '99)*, Bled, Slovenia, 1999, pp. 335–343, 1626 IEEE TRANSACTIONS ON INDUSTRIAL INFORMATICS, VOL. 10, NO. 2, MAY 2014.
[22] P. Resnik, "Semantic similarity in a taxonomy: An information-basedmeasure and its application to problems of ambiguity in natural language,"*J. Artif. Intell. Res.*, vol. 11, pp. 95–130, 1999.
[23] H. Dong, F. K. Hussain, and E. Chang, "A context-aware semantic similarity model for ontology environments," *Concurrency Comput.: Practice Exp.*, vol. 23, pp. 505–524, 2011.
[24] P. Plebani and B. Pernici, "URBE:Web service retrieval based on similarity evaluation," *IEEE Trans. Knowl. Data Eng.*, vol. 21, no. 9, pp. 1629–1642, Nov. 2009.
[25] H. Dong, F. K. Hussain, and E. Chang, "Ontology-learning-based focused crawling for online service advertising information discovery and classification," in *Proc. 10th Int. Conf. Service Oriented Comput.(ICSOC 2012)*, Shanghai, China, 2012, pp. 591–598.

Author Profile

Mr.R.Santhosh Gowtham received B.E in Computer Science P.R.Engineering College, Thanjavur in 2013. He is currently doing M. Tech in Computer Science and Engineering from PRIST University, Thanjavur, India. He has presented papers in international conferences and published papers in international journals.

S. J. Vivekanandan, pursuing PhD in Sathyabama University Chennai, Tamilnadu., India. He received B.Tech in Information Technology in 2008 and Management from SASTRA University and M.Tech in Computer Science and Engineering from SASTRA University in 2010. He is currently working as Assistant professor in PRIST University Thanjavur, India. His research interests are Data Mining and Utility Mining, Data Structures and Database.