Automatic Service Discovery using Ontology Learning Semantic Focused Crawler for Mining

[1]D.Saranya, [2]S.Thulasidass, [3]D.Gomathi

[1]M.E Computer Science & Engineering
[2]Assistant Professor / Computer Science & Engineering
[3]Assistant Professor / Information Technology
Author Correspondence: Mailam Engineering College, Mailam-604304, India.

Abstract—World Wide Web has grown to be the leading marketplace in the world and extremely accepted for online service marketing advertisement with various service industries. Service marketing advertisements are effectual carriers of mining service information in service industry. Yet, there existed ubiquitous, ambiguous and heterogeneous of service information when service users searching for mining service information over the web. In order to resolve these issues, in this paper, present a theoretical structure for an Ontology learning Semantic Focused Crawler (OLSF), by the use of automatically discovering the services, also annotating, classifying and indexing mining service information over the web. This structure combines the idea of semantic focused crawler technology is used to solve the issues of heterogeneity, ubiquity and ambiguity of mining service information by means of the Semantic net technologies and ontology learning technology is used to maintain the high performance of crawling in the uncontrolled Web environment.

Index Terms — **Ontology learning, service metadata classifier, focused crawler, Service Metadata Generator, Information retrieval.**

I. INTRODUCTION

It is well accepted that information technology has a thoughtful effect on the way business is led, and the World wide web has grown to be the wide marketplace in the world novel business experts have understood the business utilizations of the web both for their clients and key accomplices, transforming the World Wide Web into a massive shopping center with an huge list. Purchasers have the capacity to peruse an huge scope of items and administration promotions over the World wide Web, and purchase this products specifically through online exchange frameworks.

Service Marketing advertisements can be registered by service providers through various service registries, When a service provider publishes a service entity to service registries, the service entity can be annotated by alternative Semantic Web markup languages such as Resource Description Framework or Web Ontology Language (OWL), etc., and categorized by domain-specific ontologies provided within the service industry, by referencing the Uniform Resource Identifier (URI) of the service metadata to the ontological concepts. However, Service industries ignore the issue that, service entities had already been ubiquitous in the service environment and were heterogeneous without sufficient ontological support. And also Most of the online service advertising information is embedded in a huge 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 reliable format and standard, and varies from Web page to Web page.

Ontology-based semantic information retrieval is a hotspot of current research [13]. Ontology is a conceptualization of a domain where the knowledge is human comprehensible, but machine-readable format comprising of entities, attributes, relationships, and axioms. It is used as a standard knowledge depiction for the Semantic Web

The Semantic Web provides domain-knowledge-based classification tools [7]. From that OWL is one of the popular domain classifications Web Ontology Language. It follows the principle of Description Logic that represents domain knowledge by defining concepts and specifying relations between concepts, which is working for the classification of concepts and individuals. The classification of concepts refers to defining a hierarchical structure of concepts determined by the concept–sub concept relationship. The classification of instances is used to determine whether an individual is an instance of a concept.

A focused crawler may be described as a crawler which returns relevant web pages on a traversing the web pages. Web Crawlers are one of the most vital parts used by the Search Engines to collect pages from the Web and store in web page pool database.

The main objective of this paper is to gather information about what is available on public web pages to satisfy user requirements by proposing an Ontology Learning Semantic Focused (OLSF) Crawler Most Web Crawlers use Keywords base approaches for retrieving the information from Web. Other than they retrieve many irrelevant pages using crawler. Present the structure of a novel ontology Learning semantic focused crawler – OLSF crawler, by the use of automatically discovering the services, also annotating, classifying and indexing mining service information over the web with taking into account the heterogeneous, ubiquitous and ambiguous nature of mining service information available over the Internet. The OLSF structure combines the technologies of semantic based focused crawling and ontology learning, on order to preserve the performance of crawler.

The rest of this paper will be organized as follows. In Section II, we will briefly review the current research in the field of semantic focused crawlers, Ontology learning and state the issues in this area. In Section III, we will present the overall framework of the proposed Ontology Learning semantic focused crawler. In Section IV, we will present the work flow this proposed framework. In Section V we will validate the framework by means of a series of experiments. Conclusions are outlined in the final section.

II. RELATED WORK

A lot of work has been done by various researchers on semantic focused crawler. Here, we give an overview about the domain of semantic based focused crawling and ontology learning, and review related work done on ontology learning semantic 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.

The main purpose of semantic focused crawlers is to exactly and proficiently extract and download related Web information by automatically understanding the semantics underlying the Web information and the semantics underlying the previous fields. Taking into consideration the previously present service information becomes an essential issue in service industries. In order to solve this problem, a theoretical structure for a semantic based focused crawler, with the use of automatically discovering, the services also annotating and classifying the service information with the Semantic Web technologies was presented.

Ontology-based semantic focused crawlers refer to a cluster of focused crawlers that link web documents with related ontology concepts, with the purpose of filtering and categorizing web documents. Ontology is a well-formed knowledge representation, ontology-based focused crawling approach. It uses predefined concept weights for the calculation the relevant scores of web pages. Although, through the crawling process it is not easy to get the most favorable concept weights in order to maintain a stable harvest rate. To deal with this issue, we proposed a learnable focused crawling framework based on ontology.

M.Yuvarani. [5] proposed LS Crawler, which is a distinctive model of ontology based focused crawlers. It analyzes the semantic similarity between domain concepts and web page (URLs) Uniform Resource Locators. Domain concepts are stored into ontology base in the structure of ontologies. The compatible ontology will be retrieved from the ontology base when the query is entered. Then query is sent to the search engines for retrieving the relevant URLS. Next, a multithread crawler will be generated with the intention of fetch Web pages based on these URLs. After the Web pages have been crawled, all URLs and their surrounding texts will be extracted from the Web pages. These texts will then be matched with the concepts of a compatible ontology, which determine the relevance of URLs to the query.

M.Halkidi. [6] proposed thematic subsets; this system aims to organize web pages by linking their URLs to hierarchical ontology concepts. The working procedure of this crawler is first, the crawler extracts the URLs and their descriptive texts from the initial set of documents; then, the descriptive text of one URL is matched with one of the ontological concepts, and the URL is linked to the concept. A threshold of maximum times of recursions or maximum number of documents is set as an ending requirement.

The limitation of the ontology based focused crawlers can be described as follows. (i)Crawler fetched surrounding texts sometimes cannot be used to properly or adequately describe the URLs, which may increase the fault rate of the similarity computing.

Suet al. [12] proposed unconfirmed ontology learning based focused crawler in order to compute the relevance scores between topics and Web documents. Given specific domain ontology and a topic represented by a concept in this particular ontology, the score that is relevant between a Web document and the topic is the weighted sum of the occurrence frequencies of all the concepts of the ontology in the Web document. Next, this crawler makes use of reinforcement learning, a probabilistic framework for learning optimal decision making from rewards or punishments, in order to train. The learning step follows an unsupervised paradigm, in which the crawler is used to download a number of Web documents and learn statistics based on these Web documents. The learning step can also be repeated many times. This approach is capable of classifying Web documents by means of the concepts in ontology, in order to learn the weights of relations between concepts. The limitations of Su et al.‟s approach are: (i) It cannot be used in enriching the vocabulary of those created ontologies. (ii) Though the unsupervised learning pattern can work in an uncontrolled Web environment, it does not work well when there are numerous new terms emerge or when ontologies have a limited range of vocabulary.

By means of a comparative analysis of those ontology based focused crawlers, a common drawback is found, which is that none of the crawlers is able to really evolve ontologies by enriching their vocabulary contents. It is revealed that both of the approaches attempt to use learning models to deduce the quantitative relationship between the occurrence frequencies of the concepts in ontology and their topic, which may not be applicable in the real world Web development environment. When large number of changeable new terms outside the scope of the vocabulary of ontology emerges in Web documents, these approaches cannot determine the relation between the new terms and their topic, and it cannot make use of those new terms for which the determination of relatedness, which could result in the decline of their performance.

III. PROPOSED SYSTEM

The proposed system involves the concept of semantic focused crawler techniques for automatic services. These service industries have service entities for publishing, classification, and management. Till the service factories service entities are used which involves publishing, annotation, classification by alternative Semantic Web markup languages such as Resource Description Framework. Although Service Factories ignore the issue that, ubiquitous in the Business service environment, were heterogeneous without enough ontological support and Most of the online service advertising information is embedded in a huge 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. This existing service information cannot easily be retrieved. Reasons for this difficulty are the lack of semantic service annotation and service domain knowledge oriented classification. As a result, there is a lack of methodologies for discovering those service entities.

Still there is no methodology particularly designed for classifying the ubiquitous service entities, which hampers the retrieval performance for these entities. So there is a need for an automatic and ontology based service entity discovering, annotating, and classifying methodology. Advantage of these methodologies could be used to discover and categorize service entities, there by resulting in an improvement in service entity collection, management, and retrieval.

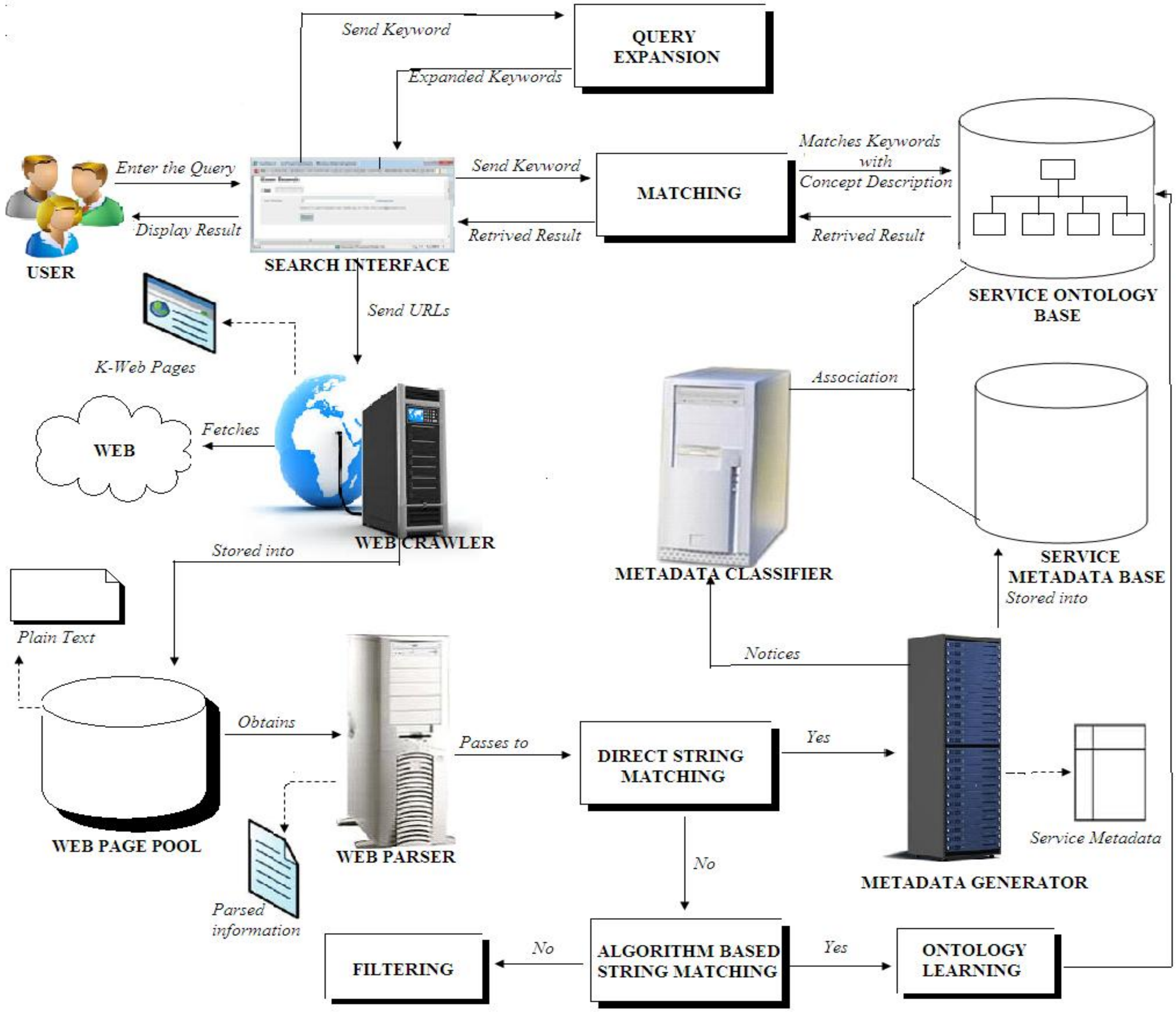

Fig 1: System Architecture

A) METHODOLOGY

Automatic service entity discovery, annotation, and classification can be achieved by using semantic focused crawling and ontology-based learning.

The proposed functions of the crawler are should able to retrieve the information regarding service entities from the Web, which corresponds to the functionality of service discovery in service industries, then able to annotate the information of services with the purpose of semantics and to store the semantic service information, which corresponds to the functionality of service annotation in service industries and also able to filter by using algorithm based string matching and classify the annotated service information by means of specific service domain knowledge, which corresponds to the functionality of service classification.

B) SYSTEM ARCHITECTURE

The system architecture is primarily composed of three components they are a Service ontology knowledge base, a semantic focused crawler and search interface. Fig.1 has shown this architecture.

Service Ontology Knowledgebase

The Service Knowledgebase is primarily composed of two components—a Service Metadata Base and a Service Ontology Base. The former is designed to store the generated service metadata, and the latter is employed to store the predefined service domain ontologies. The semantic relationship has two rules, they are.

1) A service concept may semantically relate to arbitrary service metadata.

2) A service metadata may semantically relate to arbitrary service concepts.

The structure of the service ontology concepts is a layered hierarchy. The initial layer root of the hierarchy is represents the theoretical concept of all services in a domain. The next layer is the beginning specialization of the theoretical service concept, which represents the service subdomain concepts. The third layer is the extra specialization of the theoretical service concepts, which represents the abstract service concept in each subdomain. The actual service concept is represented in bottom layer, which corresponds to the actual services in the real environment. The major difference between the theoretical concepts and the actual concepts is that only the latter can link to service metadata.

The Service Ontology is defined as the service conceptualization, which is identified by a Service Name, defined by a Service Annotation. The Service domain Ontology as the combines the name of ontology and a tuple where the tuple elements be able to complex elements as defined as follows. Service

[Service Name, Service Description] where

Service Name refers to the name that can be used to uniquely identify a service.

Service Annotation refers to the definitional descriptions of a service. The normal form of a service description is a set of words (noun, adjective, or adverb). A service concept may have many service descriptions. The purpose of setting the property of service description is to calculate the semantic similarity values between service concepts and query terms, which will be introduced later. The service domain ontology is the definition of the service concept in the root of the service concept hierarchy. All remaining concepts in this hierarchy automatically inherit its properties in the children concepts.

In addition, actual concepts have one extra property defined as follows.

Linked Metadata refers to the service URIs of semantically linked metadata to a concept.

The main purpose of service metadata is to extract meaningful information with regard to a service provided through a service provider in the real social environment. Service provider provide the service metadata, which means that even the similar service provided through two service providers are conceptualized to two metadata.

The metadata can be represented as a tuple where the tuple elements can be complex elements as defined as follows.

[Linked Concepts, Service Provider Name, Provider Address, Provider Contact Details, Service Description] where

Linked Concepts refers to the service URIs of semantically related concepts to the metadata.

Service Provider Name refers to the name of the person or organization that provides a service.

Provider Address refers to the address where a service provider can be located.

Provider Contact Details refer to the information regarding how a service provider is contacted, for instance, mail box, phone number, fax number, website, and so on.

Service Description refers to the detailed text description with regard to the content of a service. Service description can be used for matching with a service concept.

Semantic Focused Crawler

The semantic focused crawler consists of following mechanisms.

1) Web Crawler

Crawler is a software agent that can search and download Web information for certain specific topics. The Webpage crawler can download the Web pages linked by the given list of URLs. In addition, it can extract the URLs from the downloaded Web pages and send them to for further analysis. Webpage crawler is to configuring numerous policies for exploring the maximum depth to a website.

2) Webpage Pool

Webpage Pool is a database; which is used to store the downloaded web pages through the Webpage crawler. Here Web pages are stored in the form of plaintexts, so all the surrounded Web page markup language tags and scripting language tags are detached from the Web pages.

3) Webpage Parser

The job of the Webpage Parser is to extract meaningful information snippets from the Webpage Pool stored Web pages. This is done by the set of heuristic rules for the text processing. Heuristic rules are used to deal with the information heterogeneity of Web pages. Rules are referring the layouts of Web pages and a general format of service metadata for maintains a consistent style.

4) Metadata Generator

Metadata Generator is to generate metadata by annotating the information snippets obtained by the Webpage Parser with the ontology markup languages.

5) Metadata Classifier

Metadata Classifier is to classify the generated metadata by structured domain knowledge, and this task is done by associating the metadata with predefined service domain ontology concepts. The association is based on the similarities between the metadata and each service concept.

6) Metadata association and ontology learning

Metadata association is assign URI of concepts and service metadata into their linked property of service metadata and concepts respectively. Ontology learning is learned Concept Description properties of a concept and service Description property of metadata.

IV. SYSTEM WORK FLOW

Therefore, the whole working process of the system can be described as follows.

Step 1: initially user enter their query into search interface simultaneously the entered query is configure the initial URLs of visiting web sites and exploring depth to the Web sites by using web Policy Centre.

Step 2: The Webpage Crawler will start to fetch Web pages after it receives the URL list. Then will extract the URLs in the Web page and send them to the Policy Centre for further analysis. The Web pages are stored into Web page Pool.

Step 3: All the surrounded tags in the web pages will be removed, and stored these web pages in the form of plain texts to Webpage Pool.

Step 4: The Webpage Parser will extract the meaningful term information snippets from each Web pages stored in Web page Pool, and also Process that terms.

Step 5: Compare the Processed terms by using Direct String Matching algorithm. Whether processed term matches the domain then passes information to Service Metadata Generator. Otherwise processed terms entered into Algorithm based string matching.

Step 6: Algorithm Based String Matching compares concept description with metadata, whether it is matched then perform ontology learning. Otherwise filter out the Web Page.

Step 7: The Metadata Generator will create service metadata by explain the information snippets by using the ontology markup languages. Then generated service metadata is stored into the Service Metadata Base and also send to Metadata Classifier.

Step 8: The Service Metadata Classifier is perform the similarity computation between the generated service metadata and each ontology stored concepts. Then compare the similarities with threshold value, if it is above threshold value, the service concept is relevant to the service metadata. The Service Metadata Generator is associate the service metadata through the service concept and stores all these information to service ontology base.

Step 9: User entered keyword is send to matching module for matches' keywords with concept description.

Step 10: A matching algorithm is run by the matching module to perform similarity computation between the service knowledge base stored service ontology concepts and the query terms.

Step 11: the concepts with higher similarity values will be returned to the search interface and ranked according to their similarity values.

A) CONCEPT-METADATA SIMILARITY COMPUTATION AND ASSOCIATION ALGORITHM

When metadata is generated by metadata generator, then the Service Metadata Classifier is calculate the similarity between the service ontology concepts and generated metadata. Here, we make use of H. Dong. [2]'s Extended Case-based Reasoning (ECBR) model to attain this objective, which can be represented by equations:

$$sim(S,C) = \max_{sd_i \in S}\left(\max_{cd_j \in C}\left(\sum_{t_{jk} \in cd_j} \frac{f(sd_i, t_{jk})}{l_{cdj}} \right)\right) \quad (1)$$

$$f(sd_{i,} t_{jk}) = \begin{cases} 1, & if \quad t_{jk} \in sd_i \\ 0, & otherwise \end{cases} \quad (2)$$

where S - service metadata, $sd_{i\text{-}}$ value of a service Description property of the metadata S, C - service concept, cd_j - value of a concept Description property of the concept C, l_{cdj} - total number of terms that occurs in the concept Description cd_j *and* t_{jk} - term that occurs in the concept Description $cd_{j.}$

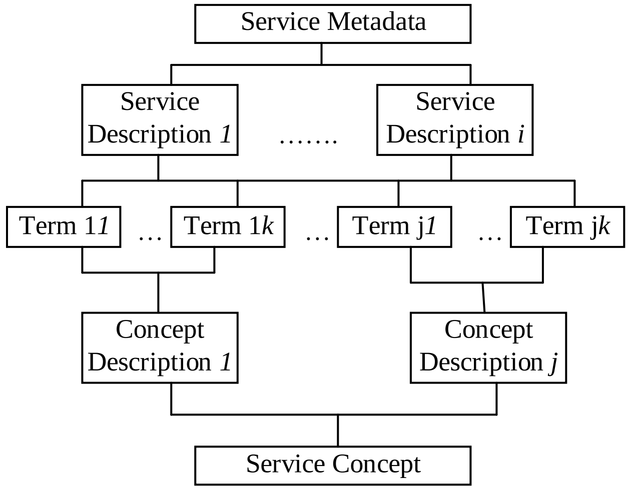

Fig. 2: Similarity computation process

```
Input: C= (C_1, C_2…C_m) is the complete set of the service
       ontology concepts, S is a service metadata,
Output: Associate the S and C to their property

Get the serviceDescription values of S
Store the values in a two dimensional array sd
for each concept in service ontology j =1 to m
    Get the conceptDescription values of $C_j$
    Store the values in a two dimensional array $cd_j$
    Calculate the similarity between $cd_j$ and sd by the
        ECBR and store the value in $s_{ij}$.
    if $s_{ij}$ > thresholds then
        linkedConcepts property of S = URI of $C_j$
        linkedMetadata property of $C_j$ = URI of S
    end if
end for
```

Fig. 3: Association algorithm

In order to get the maximum similarity between two properties of objects the ECBR model is comparing the terms obtained in the service metadata Description properties and in the service concept Description properties, Fig.2 shown this similarity computation process

After the computation of concept- metadata similarity, then perform association process using the association algorithm, Fig. 3 shown this association algorithm.

B) ONTOLOGICAL CONCEPTS AND QUERIES SIMILARITY COMPUTATION

Calculate the similarity between a service concept *C* and a user query *Q* by using the following mathematical formula.

$$sim(C,Q) = \max_{SD_i \in C}\left(\sum_{t_{ik} \in SD_i} \frac{match(t_{ik},Q) + match(t_{ik},S)}{l_{SD_i}} \right) \quad (3)$$

Where,

$$match(t_{jk},Q) = \begin{cases} 1, & if \quad \exists q_t \mid (t_{ik} = q_t) \wedge (q_t \in Q) \\ 0, & otherwise \end{cases} \quad (4)$$

$$match(t_{jk},S) = \begin{cases} 0.5, & if \quad \exists s_h \mid (t_{ih} = q_h) \wedge (s_h \in S) \\ 0, & otherwise \end{cases} \quad (5)$$

where SD_i - service description of the service concept *C*, t_{jk} - a term that appears within SD_i, l_{SDi} - the terms frequency appearing in SD_i, *S* - synonyms of *Q*, which consists of a set of terms s_k, q_t - a term that appears within the query *Q*,

The ECBR model is simple to implement, and also it saves the preprocessing time so which does not require to generate terms of index before matching It can updates the ontologies frequently, which frequently require the index term regeneration in the majority of the index term based algorithms. Because the model is independent of index terms, it does not have the problem of index term dependence.

V. EXPERIMENTAL RESULTS

A) Crawling Time

Crawling time is one of the factors which are used to measure the efficiency of a crawler. Evaluate the proposed system structure OLSF crawling time, perform the comparison of various data sets on traditional system and proposed system.

To measure proposed system crawling time various queries are used and their respective crawling times are recorded. Here the time is measured in seconds. And compare traditional system with proposed system. Figure 4 shows a graph for crawling time. A graph is plotted with X-axis against Y-axis. X-axis represents the queries of traditional and proposed system whereas Y-axis represents time in milliseconds.

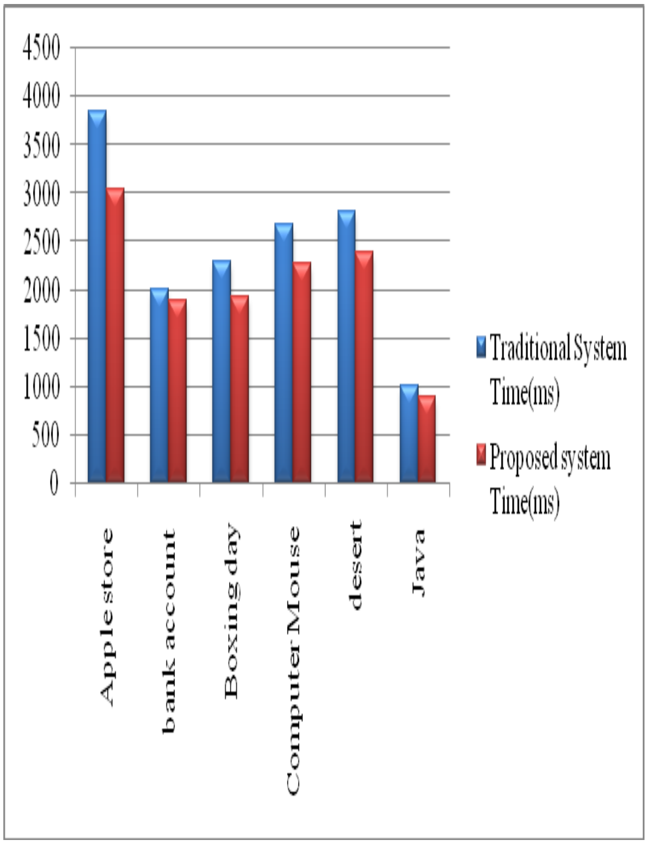

Fig. 4: Crawling Time

B) Execution Time

The execution time is another important factor for evaluating the performance of proposed system. It required for various queries. Execution time is measured in milliseconds. The execution time for traditional system and the propose system is evaluated by perform compassion of various datasets.

Figure 5 shows a graph for Execution time. A graph is plotted with X-axis against Y-axis. X-axis represents the query of traditional and proposed system whereas Y-axis represents time in milliseconds.

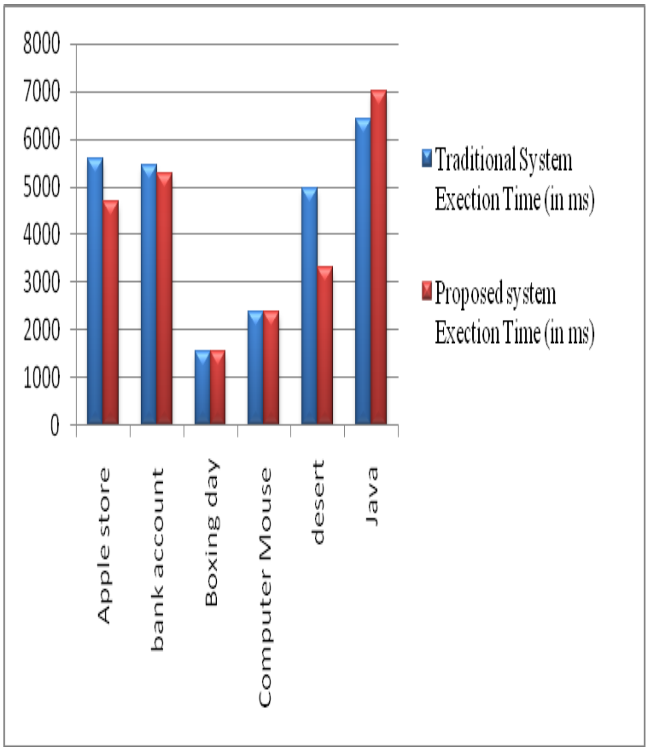

Fig. 5: Execution Time

C) Accuracy

Accuracy is measured by perform the comparison of various data sets on traditional system and proposed system.

Fig 6 shows a graph for Accuracy. A graph is plotted with X-axis against Y-axis. X-axis represents the query of traditional and proposed system whereas Y-axis represents time in milliseconds.

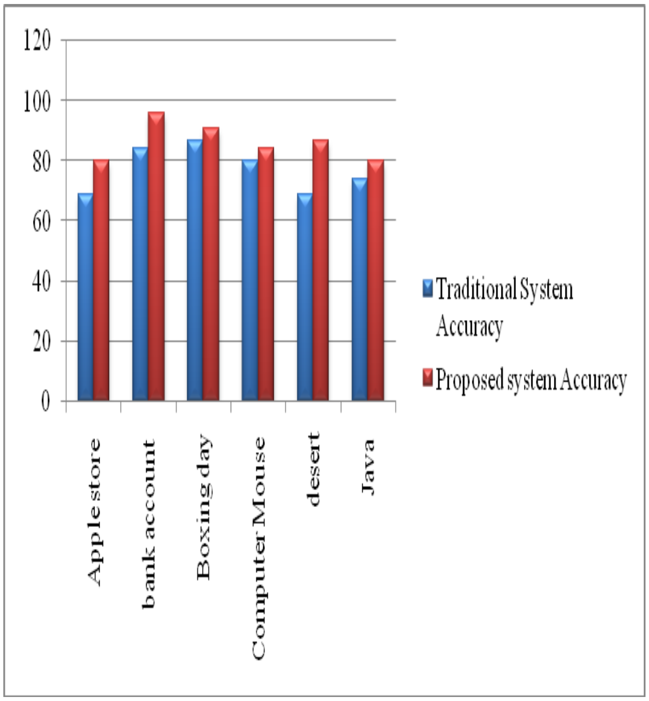

Fig. 6: Accuracy

VI. CONCLUSION

In this paper, achieve the goal of automatic service discovery in the service industries services. The main functions of the crawler include the following: discovering service information from the Web parsing, service metadata annotating, and storing the service information; and classifying the annotated service information based on specific service domain knowledge. It defined a common layout for service metadata and concept, which enables the similarity computation and the association process between service metadata and concepts. In addition, an ECBR model is used to calculate the similarity between service metadata and concepts. The service search interface retrieves ontological concepts for users' queries using ECBR model. This employs semantic similarity to improve the search results and fetch the top N - results returned by search engine, and use semantic similarities between the query and the candidate to re-rank the results. The crawler, for service information discovery in the mining service process, by taking into account the heterogeneity, ubiquity and ambiguity nature of the mining service information available over the World Wide Web.

REFERENCES

[1] Hai Dong, Member, IEEE, and Farookh Khadeer Hussain, Self-Adaptive Semantic Focused Crawler for Mining Services Information Discovery , Vol. 10, MAY 2014

[2] H. Dong and F. K. Hussain, "Focused crawling for automatic service discovery, annotation, and classification in industrial digital ecosystems," IEEETrans.Industrial .Electronics, vol.58, no.6, pp.2106–2116, Jun. 2011.

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

[4] 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, DOI: 10.1109/TIE.2009.2031186.

[5] M. Yuvarani, N. C. S. N. Iyengar, and A. Kannan, "LSCrawler: A framework for an enhanced focused web crawler based on link semantics," in *Proc. IEEE/WIC/ACM Int. Conf. WI*, 2006, pp. 794–800.

[6] M. Halkidi, B. Nguyen, I. Varlamis, and M. Vazirgiannis, "THESUS: Organizing web document collections based on link semantics," *VLDB J.*, vol. 12, no. 4, pp. 320–332, Nov. 2003.

[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] Y. Takama and S. Hattori, "Mining association rules for adaptive search engine based on RDF technology," *IEEE Trans. Ind. Electron.*, vol. 54, no. 2, pp. 790–796, Apr. 2007.

[9] C. L. Giles, Y. Petinot, P. B. Teregowda, H. Han, S. Lawrence, A. Rangaswamy, and N. Pal, "eBizSearch: A niche search engine for e-business," in *Proc. SIGIR*, Toronto, ON, Canada, 2003, pp. 413–414.

[10] S. Bechhofer *et al.*, OWL Web Ontology Language Reference Feb. 2004, W3C. [Online]. Available: http://www.w3.org/TR/owl-ref/

[11] H. Dong, F. K. Hussain, and E. Chang, "A transport service ontology based focused crawler," in *Proc. SKG*, Beijing, China, 2008, pp. 48–55.

[12] 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.

[13] Jun Zhai, Yiduo Liang, Yi Yu and Jiatao Jiang, "Semantic Information Retrieval Based on Fuzzy Ontology for Electronic Commerce," *Journal Of Software*, Vol. 3, No. 9, pp. 20-27, DECEMBER 2008