DOI: http://dx.doi.org/10.26483/ijarcs.v8i7.4471

ISSN No. 0976-5697

Volume 8, No. 7, July – August 2017

International Journal of Advanced Research in Computer Science

RESEARCH PAPER

Available Online at www.ijarcs.info

A SEMANTIC MODEL FOR WEB SEARCH AND DATA MATCHING USING LDA

Ms. Prachi Mhatre
Computer Department
MGM's College of Engineering and Technology
Navi Mumbai, India

Prof. Vilas Jadhav
Computer Department
MGM's College of Engineering and Technology
Navi Mumbai, India

Abstract: An abstract architecture for semantic web services is requisite to present a conceptual model to assist in the principled design and implementation of semantic web service applications. The architectural structure we portray stands on the shoulders of two promising technical conceptions: Web Services and the Semantic Web. Automatic discovery is acknowledged as a key task and semantic description is the foundation for automatic service discovery. We portray a novel approach for automatic discovery of semantic Web services which utilize LDA to match a user request, articulated in service discovery language, with a semantic Web service description. Our approach to semantic based web service discovery entails semantic-based service categorization and semantic enhancement of the service request. We intend a solution for accomplishing functional level service categorization based on an ontology framework. Furthermore, we exploit clustering for precisely categorizing the web services based on service functionality.

Keywords: Semantics; Automatic Discovery; Service Categorization; Semantic Enhancement.

I. INTRODUCTION

A large amount of web services exist that support and facilitate the development, deployment and invocation of various distributed applications over the web. The web services may be related to e-commerce, educational, commercial and many other fields. With the rapid emergence of web service, there is a need of a proficient technique for efficient retrieval of required web data and services. The main drawback is that majority of service descriptions are syntax based. The traditional web service discovery process was syntactic in nature. It used keyword matching techniques to search the appropriate web services. Due to this, there results could not accurately match the given service request.

Semantic description is the base for automated service discovery. Web Service Description Language (WSDL) [4] document contains semantic tagged service descriptions that are required for web service discovery. The service requester is unaware of all this semantic concepts. Hence, many web services relevant to the service request may not be considered for the service discovery process. To address the above issues, we introduce the Hybrid Web Search (HWS) approach which overcomes these issues interrelated to the discovery process. HWS uses Semantic Service Description instead of the traditional syntax based keyword matching technique because it combines semantics along with syntax to get better results.

II. REVIEW OF LITERATURE

Linda I. Terlouw *et al.* [1] have proposed an approach for service specification which contributed a method for service-orientation. The Web Service Definition Language (WSDL) and Universal Description Discovery Integration (UDDI) form the basic standards of service interfaces and service registries. But they prove to be inefficient to present the exact behaviour and understanding of a service. Based on the work proposed by the authors, the concept of service definition, service classification, and service specification framework are enlightened. K. Tamilarasi *et al.* [2] have proposed an indexing mechanism to be combined with the traditional UDDI for efficient discovery of web services. The mechanism is based on two QoS parameters namely response time and relevancy which forms an efficient framework for efficient discovery of web services. The authors have evaluated the proposed framework by implementing the algorithm and then comparing its performance with the traditional UDDI discovery. The proposed algorithm retrieves and decreases the response time of discovery. Also only related web services are fetched to the requestor thus eradicating the irrelevant web services. Wenge Rong and Kecheng Liu [3] have proposed the work which presents a summary of the field of context aware web service discovery and discuss its importance in discovery domain. Their approach faced some challenges in its service discovery process and discussed the possibilities that will enhance the overall process. Seemal Asif and Philip Webb [4] have presented an overview of some of middleware technologies which can be used to integrate different software systems. After analysing different ways of software integration it is clear that every technique and architecture has its own implications. This review helped the research of the system which is under development in the Aero-structure Assembly and Systems Installation Research Group of Cranfield University.

Nicoleta Preda *et al.* [5] have proposed an algorithm which aims to regain critical data from the database. The task of Information Extraction (IE) from the databases is a critical issue due to web service symmetry. Existing web services proposed by various authors are black box and hard-coded. The traditional IE usually takes benefits of NLP techniques such as lexicons and grammars, whereas Web IE adapts machine learning and pattern mining techniques. Considering the above issue the authors have proposed a scheme that tries to overcome the problem of web service symmetry which results in faster execution of query since it dynamically executes the services. Anthony Zukas and Robert J. Price [6] have presented an overview of the Latent Semantic Indexing (LSI) technique which is used for document categorization. The benefit of using LSI technology is that it does not depend on auxiliary structure and is also language independent. The use of LSI actuates automatic document categorization, information retrieval in a conceptual and cross-lingual manner. The motive of this research is to develop systems that can consistently categorize documents using LSI technology. Priyadharshini.G *et al.* [7] provided a survey on the available service discovery methods in the field of web service technologies. The authors have discussed various issues related to the traditional techniques used for service discovery and had enlightened the semantic based web service discovery approaches which overcomes the drawbacks of traditional techniques. Mourougaradjane Puthupattan *et al.* [8] have described different approaches used for semantic web service discovery. The authors have focused on the key issues of the semantic web discovery approaches such as dynamism, security, privacy, trust, negotiation, context awareness, Quality of Service (QoS) attributes. Sheila A. McIlraith and David L. Martin [9] have introduced the concept of semantics to web services with the motive to improve robustness and excellence of web service discovery and invocation. Introducing semantics in web services enables a wide range of automated tasks in the discovery process. Their work focused on the need to develop a language which supports the semantics in web services. A.V. Paliwal *et al.* [10] have presented a scheme for web service discovery which combines ontology linking with Latent Semantic Indexing (LSI).

III. OVERVIEW OF PROPOSED APPROACH

A. Overview of Hybrid Web Search (HWS)

Our data discovery process is based on the semantic analysis of the web service [9] and its relevant service request. In our proposed framework, execution of the data discovery process is defined under the semantic categorization of the data and semantic enhancement of the service request. Initially, the semantic categorization of the data is done offline at UDDI. The services published in the UDDI are classified depending on its functional characteristics. The categorization unit is executed just once and is not dependent on any service request. Semantic categorization corresponds to the Ontology Framework and individual web services are symbolized as vector depending on the input and output parameters of services and also their service descriptions. The service description vector along with the ontology concepts set the base of semantic categorization. The data is further labelled with their providers name in order to identify each group individually based on their parameters. The tModel based on the ontology concept are created in the registry using which we can fetch the data to signify the associated semantic descriptions [10].

A key aspect of the data discovery process is enhancement of the service request. The unprocessed service request undergoes a process of enhancement and is distorted to a service request vector. At the initial stages, ontological terms [11] are used for enhancing the service request. Further, semantic ranking is retrieved from the ontology linking to transform the service request into an enhanced service request. This enhanced service request is then matched against the refined collection of data in order to select the most appropriate data from the database. Semantic similarity-based matching [12] [13] does the task of matching the data from the set of database with the service request based on the requestor’s functionality. We use Latent Dirichlet Allocation (LDA) Model [14] [15] to measure the similarity metric for semantic similarity-based service matching.

Most web services are described in the WSDL file. The service descriptions available in the WSDL file is a key feature for semantic-based categorization and semantic enhancement of service request. So including semantic descriptions in web services which extracts the functional behaviour will lead to easy discovery and integration of services.

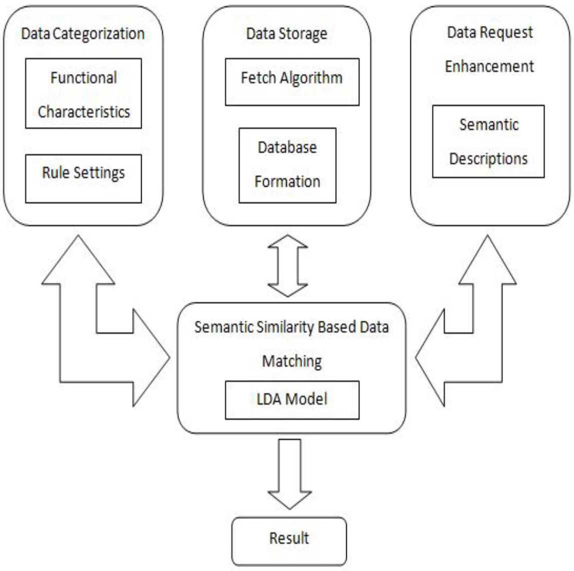

Figure 3.1 Overview of Hybrid Web Search (HWS)

B. Latent Dirichlet Allocation (LDA)

Latent Dirichlet Allocation is a generative statistical model in natural language processing that permits sets of observations to be described by unobserved groups which explain why some piece of data are similar [18]. LDA is a common method used for topic modelling that helps to figure out the type of the document. LDA treats a document just as a collection of topics where each topic has some particular probability of generating a particular word.

LDA generative process:
Step 1: For each document:
Step 1.1: Choose a distribution over topic weights.

Step 2: For every word in the document:
Step 2.1: Choose a topic from the distribution of topics.
Step 2.2: Chosen a topic, draw a word from the distribution over vocabulary.

C. Architecture of Hybrid Web Search (HWS)

This section describes the architecture and the key steps of the proposed approach. The HWS approach illustrates the Semantic Based Data Discovery Process [19] under three stages.

1. Data Categorization

The first step of the HWS approach involves setting the outline for semantic based data categorization. Data categorization deals with the organization of the database under different functional categories. The movie information which will be fetched at the second stage of the process will be stored under the categorical layout created at stage one to enhance the semantic value of the system.

The semantic data categorization is carried out by the Rule Settings section of the system and the Provider's which work under the portal unit also plays an important role in the functional categorization of data. The provider creates different sections in the database to ease the search of the required data during the result generation [20]. Rules are set for each channel handled by the provider which acts as a filter while fetching the result. The rules support categorization of the data under its behavioural categories which boosts the semantic value of the movie data.

2. Data Storage

The semantic based data storage is the second stage of the HWS approach. The movie data is fetched and stored in the database under the functional categories which we had set at stage one. The movie data is fetched from the OMDb API which is an open movie database. The task of fetching the data from this legitimate web service is done by the Package Exporter unit of the system. The data is fetched by the Fetch Algorithm which extracts the movie details from the OMDb API site and save the details in XML format in our database. XML emphasizes simplicity and is both human readable and machine readable which therefore supports our semantic approach.

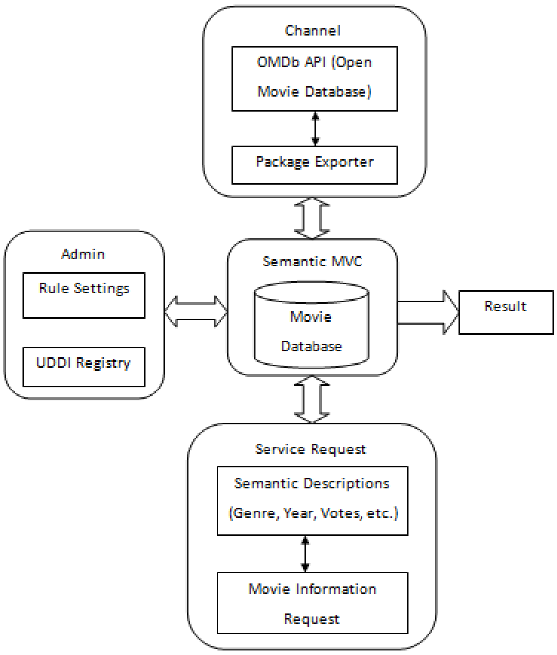
Channel
OMDb API (Open Movie Database)
Package Exporter
Admin
Rule Settings
UDDI Registry
Semantic MVC
Movie Database
Result
Service Request
Semantic Descriptions (Genre, Year, Votes, etc.)
Movie Information Request

Figure 3.2 Detailed Architecture of Hybrid Web Search

3. Data Request Enhancement

The semantic based data request enhancement does the augmentation of the service request to filter the possible outputs. The data request is enhanced by adding the data description related to the requested movie data which helps to specify the required output. For instance, the genre of the movie can be specified and its value like action or comedy will be mentioned in the search section to narrow the search in specific section. The enhancement of the service request is done by adding description to the search which relates to the ontology properties.

4. Semantic Similarity based Data Matching

LDA is a generative model in which a statistical approach is projected for modelling text documents by determining latent semantic topics in huge collections of text corpora. LDA utilizes the semantics in the text as the words in the document enclose strong semantic information.

IV. DETAILS OF IMPLEMENTATION

This section explains the flow within the modules of the web service and gives propaganda of the modules of the system. The system contains three modules which are Admin, Channel and Portal. The Admin is the chief module which handles the various units of the system. The units under Admin module are Channel Registration, Portal Registration, Rule Settings, UDDI Registry and Result which are shown in figure which show flow of implementation. Initially, the channel and portal registration procedure is performed offline by the backend users. Users are registered for the available channel and the portal assigns providers for each channel which creates a base for categorization of the data. Rule Settings are set for each channel which categorizes the movies according to the semantics of the movie information available. The rule settings unit under the admin module defines the rules for different channels based on which the validation at the UDDI unit is done. These activities create a layout for the categorization of data in the database.

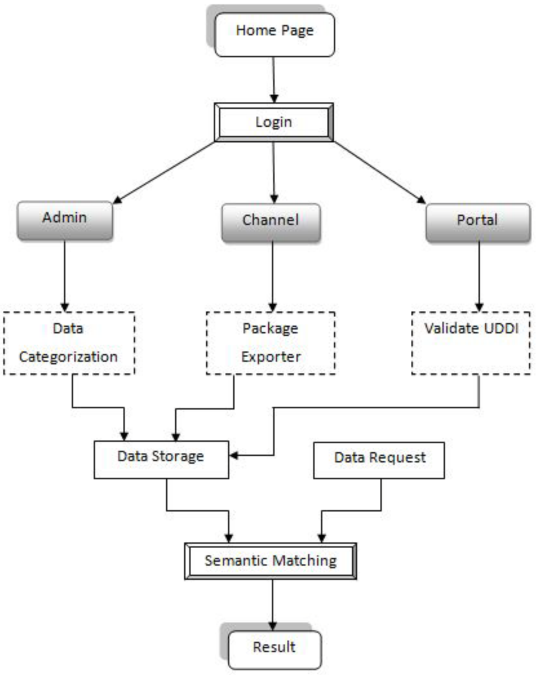

Figure 4.1 Flow of Implementation of HWS

The channel module contains the Package Exporter unit whose task is to fetch the data. The Fetch algorithm extracts the movie details from the OMDb API site and saves it in the movie database in a format that supports the semantics of the web service. The portal module validates the movie details against the rule settings. The details of all the fetched movies can be viewed at the UDDI registry unit and for each movie fetched a unique ID is assigned. For each package ID of a movie, a tModel is to be created based on the semantic descriptions present.

The task of the web service is to output all the related movie information which contains the semantic descriptions mentioned as input. The searching pattern is based on the semantic descriptions so the web service is termed as a semantic web service. Semantic enhancement of the data request is performed at the result unit by semantic filtration of the service request. Matching the data request with the required information in the database is the task of the LDA algorithm [15] which accomplishes the semantic matching.

V. RESULTS AND ANALYSIS

The Result analysis of our approach is explained in this section and it is measured on the basis of some parameters. The parameters considered here are Accuracy and Time and the algorithms to be compared are LSI (existing system algorithm) and LDA (proposed system algorithm). These algorithms are used for faster and more accurate data retrieval and the main task of these algorithms is indexing.

A. Accuracy

LSI Algorithm and LDA Algorithm both describe mathematical models that are designed to be used for information retrieval i.e. returning search results. The functioning of LSI is different from that of LDA though they have some common features in them. LSI examines the words used in a document and looks for their relationship. LSI has major weakness i.e. ambiguity as it cannot differentiate between for example office and Microsoft office.

LDA groups the words into topics and they can exist in more than one topic. LDA tackles ambiguity by comparing a document to two topics and determining which topic is closer to the document. The results achieved by LDA algorithm are more accurate than those compared to LSI algorithm. The graphical analysis shows that the LDA algorithm is approximately 14% more efficient in terms of accuracy than the LSI algorithm.

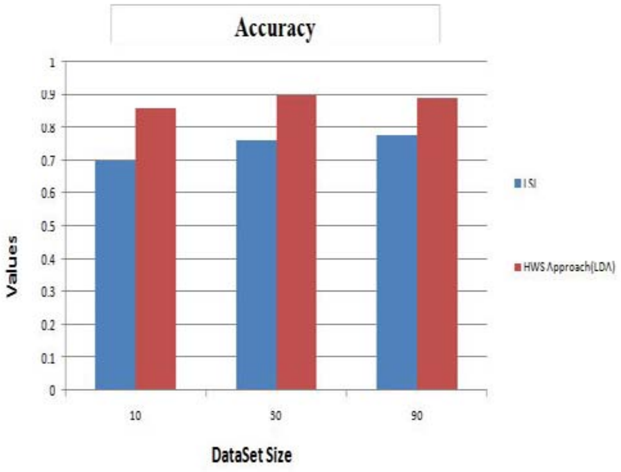

Figure 5.1 Accuracy Measurement Chart

B. Time

We propose an approach which uses a probabilistic model called LDA for information retrieval. LDA uses the bag of words concept, thus it is a collection of words under a topic. So it is more time efficient than LSI in consideration to the time constraint.

The below graph give an idea about the comparison between LDA and LSI algorithm in terms of time constraint. The analysis shows that the LDA algorithm gives approximately 62% more efficient results as compared to LSI algorithm in consideration to the time parameter.

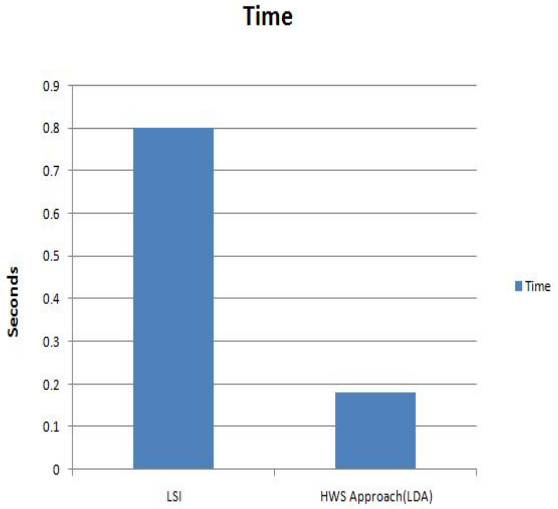

Figure 5.2 Time Comparison Chart

VI. CONCLUSION AND FUTURE SCOPE

The traditional syntax-based approach for service discovery does not emphasize the use of semantics in the service discovery process due to which only few services which match the exact keywords are focused. Using our HWS approach the service discovery process starts encapsulating the semantic aspects of the web descriptions leading to an automated service discovery process. Our HWS approach is derived using the features extracted from semantic service categorization and semantic service enhancement. The categorization of the services depends on the ontology framework which classifies the web services according to their functional characteristics. This leads in better matching process by creating a narrow group of web services which helps to retrieve the appropriate services based on the service descriptions specified. The LDA technique is used for semantic similarity matching between the web service description and the web service request.

In general, multiple services have to be discovered so that they collectively match a service request. It should be possible to exploit ontologies, and explicitly return the sequence of individual service invocations to be executed in order to achieve the required composite service. When no full match is feasible, a flexible corresponding approach could be formed to return partial matches and imply additional inputs that would fabricate a full match by capturing the dependencies amid the matched services. This has several fascinating research issues. Another avenue for future work is to construct an interactive, intellectual service composer that is semantically guided to position the target service components step by step. We also aim to extend our ontology framework and explore additional mapping tools to better convey a service request to search for appropriate concepts. Finally, as part of the service discovery procedure we will discover associating semantic weights to the retrieved set of web services for valuable semantic ranking of the results.

VII. ACKNOWLEDGMENT

We are thankful to MGMCET who gave us the opportunity to complete the research work. We would like to express our gratitude to our Professors, who guided us in the light of their previous professional experiences and knowledge. Special thanks to our colleagues, friends and family for providing us useful comments, suggestions and continuous encouragement.

VIII. REFERENCES

[1] Linda I. Terlouw and Antonia Albani, "An Enterprise Ontology-Base Approach to Service Specification," IEEE Transactions on Services Computing, Vol. 6, No. 1, January-March 2013, pp. 89-101.

[2] K. Tamilarasi and Dr. M. Ramakrishnan, "Indexing Traditional UDDI for Efficient Discovery of Web Services," in Indian Journal of Computer Science and Engineering (IJCSE), Vol. 6, No.1, Feb-Mar 2015, pp. 14-20.

[3] Wenge Rong and Kecheng Liu, "A Survey of Context Aware Web Service Discovery: From User's Perspective," in Fifth IEEE International Symposium on Service Oriented System Engineering (SOSE), June 2010, pp.15-22.

[4] Seemal Asif and Philip Webb, "Software System Integration – Middleware – An Overview," in International Journal of Computer Applications, Volume 121, No. 5, July 2015, pp. 27-29.

[5] Nicoleta Preda, Fabian Suchanek, Wenjun Yuan and Gerhard Weikum, "SUSIE: Search Using Services and Information Extraction," IEEE Transactions on Knowledge and Data Engineering Year 2013.

[6] Anthony Zukas and Robert J. Price, "Document Categorization Using Latent Semantic Indexing," 2013.

[7] Priyadharshini.G, Gunasri.R and Saravana Balaji.B, "A Survey on Semantic Web Service Discovery Methods," in International Journal of Computer Applictions, Volume 82, No. 11, November 2013, pp. 8-11.

[8] Mourougaradjane Puthupattan and Dinadayalan Peruma, "A Comparitive Study on Semantic Web Service Discovery Approaches," (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 6 (2), 2015, pp.1584-1587.

[9] Sheila A. McIlraith, David L. Martin, "Bringing Semantics to Web Services," in Intelligent Systems, IEEE, Volume: 18, Issue: 1, 2003, pp.90-93.

[10] Aabhas V. Paliwal, Basit Shafiq, Jaideep Vaidya, Hui Xiong and Nabil Adam, "Semantic-Based Automated Service Discovery," IEEE Transactions on Services Computing, Vol. 5, No. 2, April-June 2012, pp. 260-275.

[11] M. Suchithra and M. Ramakrishnan, "A Survey on Different Web Service Discovery Techniques," in Indian Journal of Science and Technology, Vol 8(15), July 2015, pp. 1-5.

[12] M. Deepa Lakshmi and Dr. Julia Punitha Malar Dhas, "An User-Friendly and Improved Semantic-Based Web Service Discovery Approach Using Natural Language Processing Techniques," in International Journal of Innovative Research in Computer and Communication Engineering, Vol. 1, Issue 10, December 2013, pp. 2435-2442.

[13] Arunachalam, R. Swarnalakshmi, R. Sangeetha and B. Pradheepa, "An Ontology-Based Service Discovery Framework For An Enterprise," International Journal of Computer Science and Mobile Applications, Vol.1 Issue. 5, November- 2013, pp.65-75.

[14] Alexander Ihler and David Newman, "Understanding Errors in Approximate Distributed Latent Dirichlet Allocation," IEEE Transactions on Knowledge & Data Engineering, Vol. 24, Issue no. 5, May 2012, pp. 952-960.

[15] Anima Anandkumar, Dean P. Foster, Daniel Hsu, Sham M. Kakade and Yi-Kai Liu, "A Spectral Algorithm for Latent Dirichlet Allocation," Volume 72, Issue 1, May 2015, pp. 193-214.

[16] Quan Wang, Jun Xu, Hang Li and Nick Craswell, "Regularized Latent Semantic Indexing," SIGIR '11 Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval, July 2011, pp. 685-694.

[17] Mehmet Yalcinkaya and Vishal Singh, "Patterns and trends in Building Information Modeling (BIM) research: A Latent Semantic Analysis," in Automation in Construction, September 2015, pp. 68-80.

[18] A.V. Paliwal, N. Adam, and C. Bornhoevd, "Adding Semantics through Service Request Expansion and Latent Semantic Indexing," in IEEE International Conference on Services Computing, July 2007, pp.106 – 113.

[19] Soodeh Pakari, Esmaeel Kheirkhah and Mehrdad Jalali, "Web Service Discovery Methods and Techniques: A Review," International Journal of Computer Science, Engineering and Information Technology (IJCSEIT), Vol. 4, No.1, February 2014.

[20] Abhishek Pandey and S.K.Jena, "Dynamic Approach for Web Services Selection," Proceedings of the International MultiConference of Engineers and Computer Scientists (IMECS), Vol I, March 2009.