International Journal of Applied Engineering Research ISSN 0973-4562 Volume 10, Number 24 (2015) pp 44166-44171

An automatic Search Result Computation and Recommendation Using Bayesian Method for QoS improvisation

Althaf Ali A

Research Scholar Department of Computer Science Bharathiar University Coimbatore-641046 Tamilnadu
althafa579@gmail.com

Dr. R. Mahammad Shafi

Research Supervisor Department of Computer Science Bharathiar University Coimbatore-641046 Tamilnadu
rmdshafi@gmail.com

Abstract
Big Data is pertaining to a huge volume, complex growing data with multiple and autonomous resources in the fast developments of the distributed applications, websites and data collection capacity. The immense increase creates challenges in suggesting users appropriate information need of their interest in terms of web link or web pages for their queries request. Most of the previous recommendation approaches are based on the user web usage data to build knowledge and suggestion of the web data, but these are satisfactory only to particular domain information. In case of distributed domains information it fails to build knowledge and recommend inaccurate web link or pages. This paper presents an automatic search result computation and recommendation approach which will cover the distributed domain information over cloud and other networks. It implement a modified Bayesian method for accurate recommendation the right value of information need to solve the information retrieval accuracy in relate to a user search query. The evaluation testing with few popular search engines and empirical measures shows a QoS improvisation in recommendation.

Keywords: Big data, Search Result, Recommendation, QoS, Cloud computing.

Introduction

In recent years popularity of the internet has grown to a great extent with nearly every person, young or old, using it for a variety of purposes. People use the internet to get information in areas of interest, do research related to work or study, get good deals for commodities or travel, increase awareness about their surroundings and the world, get latest news, etc. With each passing day, large amounts of informative web sites, web pages or web documents get added to the already huge collection to form big data over distributed cloud [1], [10], [11]. Any popular search engine returns thousands of related links to a search query. It has become highly difficult for users to get the most relevant information from this excess of related information readily available. Users often spend considerable time browsing the web pages for getting the right information [7].

Webpage recommendation [2] has become increasingly popular, and is shown as links to related stories, related books, or most viewed pages at websites [3]. But the growth of information collection at various domains over this cloud era creates an extensive computation overhead to recommend a relevant page for a query to the users. Many approaches in relevant to recommendation are based on web usage mining [14], [15], [19]. These approaches work well if it related to particular domains search or for a individual websites [8], [9]. But, in case of multi domain information search of or in open search environment the web usage mining based approaches are incompetent to provides the relevance results, due to it limitation to process different domain or website user usage log collectively for the accurate recommendation.

This paper presents an automated search results computation and recommendation approach utilizing the popular search engines results in relevance to a particular search query. The result recommendation is performed using a modified Bayesian method. It will compute the best recommendable result based on the results local rank and position of the retrieved results and through computing overall rank for each results from the extracted result and suggest the best relevance result for the recommendation.

The following paper organized in five sections. InSection-2 we present an insight on the background works, Section-3, discuss the proposed Automatic Search Result Recommendation Approach, Section-4, present the Experiment Evaluation and results and Section-5 presents the conclusion and future work.

Background Study

In literature various approaches and techniques are being discussed to model and understand the user activity based on web usage mining for the information recommendation[5], [6]. Mostly these approaches are based the traditional sequence learning and semantic learning model approaches and implements association rules and probabilistic models in sequence learning for the recommendation.

Jespersen et al. [14] proposed a hybrid approach for analyzing the visitor click-stream sequences. A combination of hypertext probabilistic grammar and click fact table approach is used for Weblog mining that could also be used for general sequence mining tasks. Mobasher et al. [15] proposed the web

personalization system, the offline tasks made mining related, it usage data and online process of automatic Web page customization based on the discovered knowledge. Chi et al. [16] built LumberJack utilizing user profiles by combining both clustering of user sessions and traditional statistical traffic analysis with *k*-means algorithm. Even some works are proposed using Markov and tree-based structuring models utilizing web pages transitions in different web sessions [17]. In [18] a tree-based approach describes that recommendation pages are processed in advance and form a Pre-Ordered linked tree which supports in recommendation. In [19], [20] a Markov model based recommendation is described which shows an significant enhancement in the performance of the recommendation.

Semantic based information processes for recommendation are also being in suggested in [21]. These approaches make use of website ontology for the recommendation process. These approaches show a good enhancement in semantically processing using web information and logs for web recommendation and web personalization system [22]. But these approaches are very much limited to a particular type of domain or a set of information, such as education domain contents, personalized e-learning or a particular subject related contents[23], [24].

L. Wei et al. [21] proposed ontology based online recommendation system. It generates an ontology which represents a website domain knowledge using the term frequencies which are extracted from the document concepts. It recommends pages through semantically comparing and searching against the user request to achieve higher accuracy rate and user satisfaction.

In spite of all these approaches the recommendations systems [12] are failed to cover the vast distribution of the information over different networks [13]. In this paper, we present a novel approach integrating multiple search engines to extracted results in related to the user query over different published information using a Bayesian probability model for automatically recommend the best result over various information domains and improve the QoS through achieving higher precision results.

Automatic Search Result Recommendation Approach

The proposed recommendation approaches is presented in the Figure-1 framework. It consists of four main components which perform different computational jobs for accomplishing the recommendation process. It has Request and Response Handler, Information Extractor using Web Services, Result Evaluator and Result Recommender.

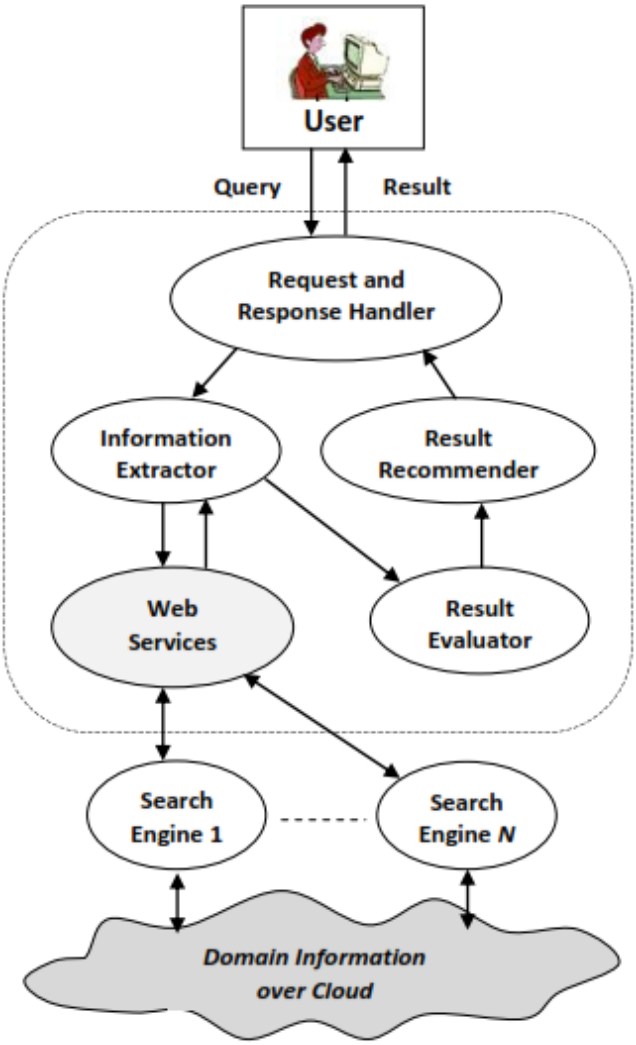

Fig.1. Framework for automatic result recommendation

A. Request and Response Handler

Request and Response Handler is an important component which handles the user request query and intern reply the recommended results provided by the result recommender. On receiving the user query it performs the initial pre-processing to prepare search key words by eliminating the prepositions, determiners, verbs and adjectives. For an example, if a user send a request as "*developments in treatment of cancers in past five years*" the key terms prepared as "*developments, treatments, cancers, past, five, years*" and the words "*in, of*" are eliminated with a supported filter library maintained.

B. Information Extractor

Information Extractor is another key component which performs information extraction using web services. It takes the key terms input from request hander to process. Web service is developed for type of processes that can be integrated into external systems through valid XML documents over Internet protocols. The dynamic nature of web services provides high scalability for the integration with

the different search engine interfaces. It acts as a software component which interfaces to communicate with other software components. It mostly follows a Remote Procedure Call (RPC) to avail the information different search engines.

An information extractor is needed to extract the correct search result records from different component search engines can be merged into a single ranked list. This program is sometimes called an extraction wrapper. Since different search engines often format their results differently, a separate result extractor program is usually needed for each component of search engine. But, here it automatically process the obtained results and generate a unique sets of results for further processing.

C. Result Evaluator

Result evaluator process the retrieved result from the from the search engine which is provided by information extractor. Typically a search engine returns one or more response pages in response to a search query. Each page consists of multiple search result records, usually 10, and it consists of a link to the source page and the title of the page and a short summary in related.Mostly the page snippets and title can provide good evidence on whether the relevant document is relevant to the query or not. In information search, search engine gives high weights to the title of pages primarily. Snippets are usually specially created for the user's query is submitted, and they are often the best fragments of text in the documents that match the query. Weight of each word in the title and snippet is calculated using term frequency and with the sum of the term weights of query and computed weight of the title is compared to find the similarity, and finally based on the similarities obtain results are presented in a descending order for recommendation process.

D. Result Recommender

All of the major search engines, implements the best information retrieval algorithm for information retrieval. But due the limitation of web crawling and indexing techniques they maintains every search engine present the results in different order. The degree of challenge is to best fit the user meets in order of the requirement arises to merge all the relevant results receiving from the different search engines.

Most of the expert systems and decision analysis system use Bayesian method to perform a specific analysis for the works [4].We enhance the Bayesian method which operates on a probability model to ranks the local results. The method performance depends on the accuracy of the training data queries. But the proposed approach compute the probability without any prior training, it works based on the runtime data obtained related to result position and overall rank.

Let's assume a user submit a query Q for the information search, E_n is the number of search engine used for the information search, and the obtained result are stored in a result set as Ralong with their rank k_i.To recommend the best result we compute result position rank as $P_i$of each result using equation-1. The computed $P_i$of each result are stored for the final computation.

$$P_i = \frac{\sum_{i=n}^{i=0} k_i(R)}{n} \quad (1)$$

On completion of P_i computation, we computes the result relevancy against the requested query using modified Bayesian method which compute the probabilities of relevancy as $A_{relv}$and irrelevancy as A_{irrv} for the query Qresults data sets using equation-2 and 3 and it's optimal relevance as OA_{relv} and irrelevance as $OA_{irrv}$is computed using equation-4 and 5. If a no relevance results is retrieved by any search engine then it rank will be considered as ∞and in case if zero irrelevancies OA_{irrv} will be 1.

$$A_{relv} = Prob\ (relv|r_1, \ldots\ldots r_n) \quad (2)$$

$$A_{irrv} = Prob\ (irrv|r_1, \ldots\ldots r_n) \quad (3)$$

$$OA_{relv} = \frac{p_{rel}}{n} \times 100 \quad (4)$$

$$OA_{irrv} = \frac{p_{irr}}{n} \times 100 \quad (5)$$

According, the bayesian optimal decision rule a result will be considered as relevant if and only$\frac{OA_{relv}}{OA_{irrv}} \geq 1$, and only the result which obtains the optimal ratio ≥ 1 are stored. The final relevance obtained results initially ordered based on the P_i value in ascending order and reordered based on the OA_{relv} in descending order for the final recommendation. Here, OA_{relv} results which are below a threshold limit *(TL)* will be not considered for recommendation, we consider, $TL \geq 50$ for the evaluation.

Experiment Evaluation

i. Setup

The propose framework is developed using Java Web Service, Servlet and JSP technology. To build Information extractor and deploy Web services Java Web Service Developer Pack (WSDP) is implemented. It provides a set of toolset and new API including XML Messaging (JAXM), XML Processing (JAXP), XML Registries (JAXR), XML-based RPC (JAX-RPC) and the SOAP with Attachments (SAAJ). The main advantage of Java WSDP is it supports heterogeneous platforms and has dynamic behaviour.

Web Service Description Language (WSDL) [25].file serves as a reference for a Web Service. It helps to communicate and learn how to use the corresponding service using WSDL file. Among the advantages, the most important for Web services is an XML-based platform and it allows easy data processing and exchange between different applications. The most popular HTTP transfer protocol and it supported by almost all platforms. This combined with XML standard and implementation between the Web services platforms to remove almost all the boundaries. In a decentralized and distributed environment information are exchanged using Simple Object Access Protocol (SOAP) [26]. A message describing what to do and how to do it, the application-defined data types, signals, encoding rules for expressing an envelope that defines a framework for the process, and for a

meeting: SOAP is an XML-based protocol that consists of three parts representing remote procedure calls and responses.

ii. Evaluation Test

For evaluation test we considered 4 popular search engine as Google, Yahoo, Excite and Ask for the integration and we send a request query as "*big data mining*" to the request handler. With the help of result extractor web services we retrieve the top 10 results from the 4 search engine. The initial position of the obtain from the different search engine are shown in Table-1.

Table-1: Initial Result Position of the results retrieved

Rank	*Google*	*Yahoo*	*Excite*	*Ask*
1	G1	Y1	Y1	A1
2	G2	G8	G8	G1
3	G3	G2	G2	G2
4	G4	G10	G10	G5
5	G5	Y5	Y5	A5
6	G6	Y6	Y6	A6
7	G7	Y7	Y7	G7
8	G8	Y8	Y9	G8
9	G9	Y9	G9	A9
10	G10	G9	Y8	Y6

On combining the unique obtained results it form a set of result as *Z*, such that *Z={ G1, G2, G3, G4, G5, G6, G7, G8, G9, G10, Y1, Y5, Y6, Y7, Y8, Y9, A1, A5, A6, A9}*. The obtained set *Z*, is used to compute the position ranking using equation-1. The computed position ranks are shown in Table.2.

Table.2. Unique Results Initial and New position Rank

Initial Positions	
Unique Results	P_i *Values*
G1	6.25
G2	2.75
G3	9
G4	9.25
G5	7.75
G6	9.75
G7	9
G8	5
G9	9.75
G10	7.25
Y1	6
Y5	8
Y6	8.25
Y7	9
Y8	10
Y9	10.5
A1	8.5
A5	9.5
A6	9.75
A9	10.5

New Result Positions	
Unique Results	P_i *Values*
G2	2.75
G8	5
Y1	6
G1	6.25
G10	7.25
G5	7.75
Y5	8
Y6	8.25
A1	8.5
G3	9
G7	9
Y7	9
G4	9.25
A5	9.5
G6	9.75
G9	9.75
A6	9.75
Y8	10
Y9	10.5
A9	10.5

The obtained new results in Table-2 are further used to compute the relevancy, A_{relv} and irrelevancy, A_{irrv} using modified Bayesian probability equation-4 and 5. The computed results are shown in Table-3.

Table.3. New result Relevancy and Irreverence percentage

New Results	P_i Values	A_{relv} Value	A_{relv} (%)	Recommendation Result
G2	2.75	4	100	G2
G8	5	4	100	G8
Y1	6	2	50	G10
G1	6.25	2	50	G9
G10	7.25	3	75	Y1
G5	7.75	2	50	G1
Y5	8	2	50	G5
Y6	8.25	2	50	Y5
A1	8.5	1	25	Y6
G3	9	1	25	G7
G7	9	2	50	Y7
Y7	9	2	50	Y8
G4	9.25	1	25	A1
A5	9.5	1	25	G3
G6	9.75	1	25	G4
G9	9.75	3	75	A5
A6	9.75	1	25	G6
Y8	10	2	50	A6
Y9	10.5	1	25	Y9
A9	10.5	1	25	A9

Using the computed results and final recommendation result generated as shown in Table-3, result recommender sends the top-10 recommend results from the final result set generated.

The proposed approach provide high relevance result and enhance the QoS in the information search. To measure the enhancement we measure the precision and recall rate with a variation of number of search engine as describe in below section.

iii. Evaluation Measure

Precision: Precision in the information retrieval is used to measure the preciseness of a retrieval system. In this experiment, Precision for a single query is the proportion of the relevant result associated by this query in all the results associated by this query, which can be mathematically represented as,

$$Precision(P) = \frac{No.of\ associated\ and\ relevant\ Result}{No.of\ associated\ results}$$

Recall: Recall in the information retrieval refers to the measure of effectiveness of a query system. In this experiment, Recall for a single query is the proportion of the relevant results associated by this query in all the relevant results of this query in the collection of generated results, which can be represented as,

$$Recall(R) = \frac{No.of\ associated\ and\ relevant\ Result}{No.of\ relevant\ results}$$

To measure the precision and recall of our proposed method we implement and run the simulation with varying number of search engine selection from 2 to 4. Based on the obtained result we tabulated and present the comparison analysis graphs in Fig.2.and 3.

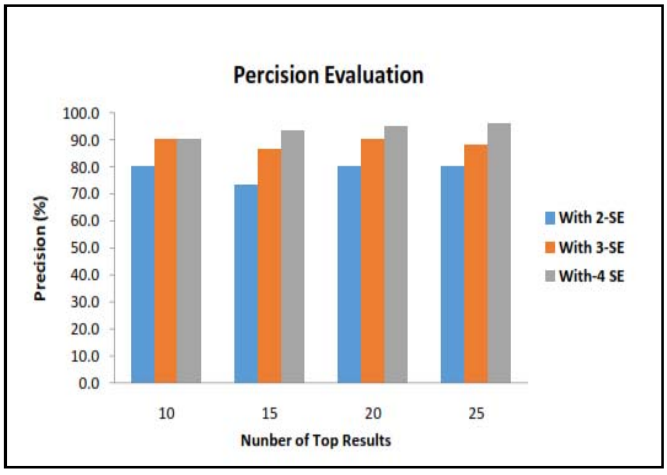

Fig.2. Precision Evaluation Results

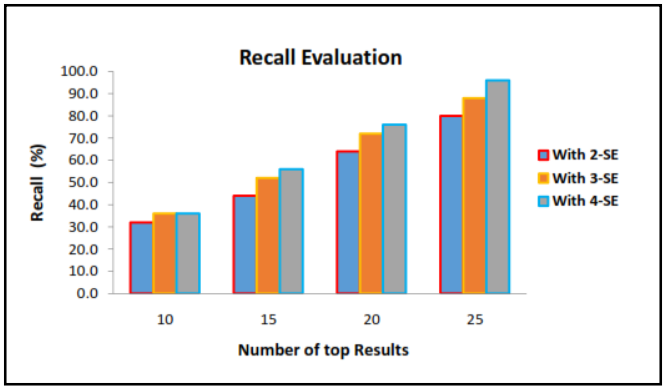

Fig.3. Recall Evaluation Results

The obtained result in Figure-2 and 3 describes that with the increasing the number of search engine for the information retrieval shows an improvisation in precision level. It is due to the selection high relevance result and discarding duplicates and irrelevant results. This concludes that proposed approach efficiently able to recommend the precise to meet the user needs. It can be an efficient approach which can be integrated with any application for searching information any distributed network and domain.

Conclusion

The vast distribution of information over cloud and web makes challenging for user to find the accurate information in related to a query. Most popular search engine implements advance algorithm for result indexing and presenting. In this paper, we present an result recommendation approach using modified bayesian method to recommend precise result for the improvisation and QoS in the information search. The experiment evaluation results shows an high precise and relevance result with integrating multiple search engine for the result extraction and recommendation. This approach works automatically to recommend the best results without any prior training knowledge. In future, it can be enhanced with web usage data for improvising and evaluate the impact on result accuracy over cloud and distributed web data.

References

[1]. Xindong Wu, XingquanZhu, Gong-Qing Wu and Wei Ding, January, 2014 "Data Mining with Big Data", IEEE Transactions On Knowledge And Data Engineering, Vol. 26, No. 1, pp 97-107.

[2]. ThiThanh Sang Nguyen, Hai Yan Lu, and Jie Lu, October, 2014"Web-Page Recommendation Based on Web Usage and Domain Knowledge", IEEE Transactions On Knowledge And Data Engineering, Vol. 26, No. 10, pp 2574-2587.

[3]. M.H. Alam, J.W. Ha, and S.K. Lee, Dec. 2012 "Novel Approaches to Crawling Important Pages Early", Knowledge and Information Systems, vol. 33, no. 3, pp 707-734.

[4]. R. Chen, K. Sivakumar, and H. Kargupta, 2004"Collective Mining of Bayesian Networks from

Distributed Heterogeneous Data", Knowledge and Information Systems, vol. 6, no. 2, pp. 164-187.

[5]. Zhengkui Wang, DivyakantAgrawal and Kian-Lee Tan OSAC September, 2013"A Framework for Combinatorial Statistical Analysis on Cloud" IEEE Transactions On Knowledge And Data Engineering, Vol. 25, No. 9.

[6]. K. Su, H. Huang, X. Wu, and S. Zhang, 2006 "A Logical Framework for Identifying Quality Knowledge from Different Data Sources", Decision Support Systems, vol. 42, no. 3, pp. 1673-1683.

[7]. Jarvelin, K. and Kekalainen, J. 2000"IR Evaluation Methods for Retrieving Highly Relevant Documents", In Proceedings of the 23rd Annual Int. ACM SIGIR Conference on Research and Development in Information Retrieval, ACM Press, pp 41-48.

[8]. I. Kopanas, N. Avouris, and S. Daskalaki, 2002 "The Role of Domain Knowledge in a Large Scale Data Mining Project", Proc. Second Hellenic Conf. AI: Methods and Applications of Artificial Intelligence, I.P. Vlahavas, C.D.Spyropoulos, eds., pp. 288-299.

[9]. E.Y. Chang, H. Bai, and K. Zhu, 2009 "Parallel Algorithms for Mining Large-Scale Rich-Media Data", Proc. 17th ACM Int'l Conf. Multimedia, (MM '09,) pp. 917-918.

[10]. S. Papadimitriou and J. Sun, 2008 "Disco: Distributed Co-Clustering with Map-Reduce: A Case Study Towards Petabyte-Scale End-toEnd Mining", Proc. IEEE Eighth Int'l Conf. Data Mining (ICDM '08), pp. 512-521.

[11]. D. Luo, C. Ding, and H. Huang, 2012 "Parallelization with Multiplicative Algorithms for Big Data Mining", Proc. IEEE 12th Int'l Conf. Data Mining, pp. 489-498.

[12]. Nuray R. and Can, F. 2006 "Automatic ranking of information retrieval systems using data fusion, " Information Processing and Management, vol. 42, issue 3, pp. 595-614.

[13]. R. Ahmed and G. Karypis, 2012 "Algorithms for Mining the Evolution of Conserved Relational States in Dynamic Networks", Knowledge and Information Systems, vol. 33, no. 3, pp. 603-630.

[14]. Jespersean S.E., Throhauge J., and Bach T., 2002 "A hybrid approach to Web Usage Mining, Data Warehousing and Knowledge Discovery", SpringerVerlag Germany, pp73-82.

[15]. BamshadMobasher, Robert Cooley, andJaideepSrivastava, 1997 "Web Mining: Information and Pattern Discovery on the World Wide Web", in Proceedings of the 9th IEEE International Conference on Tools with Artificial Intelligence.

[16]. Chi E.H., Rosien A. and Heer J., LumberJack: 2002 "Intelligent Discovey and Analysis of Web User Traffic Composition". In Proceedings of ACMSIGKDD Workshop on Web Mining for Usage Patterns and User Profiles, Canada, ACM press.

[17]. B. Zhou, S. C. Hui, and A. C. M. Fong, 2006 "Efficient sequential access pattern mining for web recommendations", Int. J. Knowl.-Based Intell. Eng. Syst., vol. 10, no. 2, pp. 155-168.

[18]. C. I. Ezeife and Y. Lu, 2005 "Mining Web log sequential patterns with position coded pre-order linked WAP-tree", Data Min. Knowl. Disc., vol. 10, no. 1, pp. 5-38.

[19]. J. Borges and M. Levene, 2005 "Generating dynamic higher-order Markov models in Web usage mining", in Proc. PKDD, Porto, Portugal, pp. 34-45.

[20]. S. T. T. Nguyen, 2009 "Efficient Web usage mining process for sequential patterns", In Proc. IIWAS, Kuala Lumpur, Malaysia, pp. 465-469.

[21]. L. Wei and S. Lei, 2009 "Integrated recommender systems based on ontology and usage mining", In Active Media Technology, vol. 5820, Eds. Berlin, Germany: Springer-Verlag, pp. 114-125.

[22]. M. Eirinaki, D. Mavroeidis, G. Tsatsaronis, and M. Vazirgiannis, 2006 "Introducing semantics in Web personalization: The role of ontologies", in Proc. EWMF, Porto, Portugal, pp. 147-162.

[23]. D. Dzemydiene and L. Tankeleviciene, 2008 "On the development of domain ontology for distance learning course", in Proc. 20th EURO Mini Conf. Continuous Optimization Knowledge-Based Technologies, Neringa, Lithuania, pp. 474-479.

[24]. J. M. Gascuena, A. Fernandez-Caballero, and P. Gonzalez, 2006 "Domain ontology for personalized e-learning in educational systems", in Proc. 6th IEEE ICALT, Kerkrade, Netherlands, pp. 456-458.

[25]. WSDL Specifications and RFCs-http: //www.w3.org/TR/wsdl.

[26]. SOAP specifications and RFCs-http: //www.w3.org/TR/soap/.