ISSN 2394-3777 (Print)
ISSN 2394-3785 (Online)
Available online at www.ijartet.com
International Journal of Advanced Research Trends in Engineering and Technology (IJARTET)
Vol. 3, Special Issue 20, April 2016

An Efficient and Query Specific Semantic Latency Algorithm for Web Based Information Retrieval Incorporating URL Normalization

Gerard Deepak, Rushikeshav G
Department of Computer Science and Engineering
Bangalore University
University Visvesvaraya College of Engineering
Bangalore, India

***Abstract*— The World Wide Web is growing and expanding owing to its enormous amount of information as well as large number of users getting added to it in the era of Internet and Smart Phones.Extraction of useful and needed information from the World Wide Web is a highly tedious task. The present day Web is loaded with information and is growing at a very high rate. Satisfying the users' requirement with relevant information is a cumbersome and a crucial task. To overcome this, Personalization of Web Search is highly recommended. A query driven strategy for Personalized Web Search is proposed using a Semantic Web Strategy.The Web Usage Information is elicited, analyzed using the Semantic Latency methodology. The proposed methodology computes the Semantic Heterogeneity to find and recommend the Web Pages of users' interest. An algorithm for Semantic Personalized Web Page Recommendation is proposed based on the input query and the Web Usage Information specific to a user. An average precision of 0.74 is achieved which is much better than the average precision of the existing approaches.**

Keywords*— *Personalized Web Search, Recommendation Systems, Semantic Latency, Web Usage Information.

1. INTRODUCTION

The present era is the age of information and knowledge. With the advancement in the Smart Phone Technology, several users are being added to the World Wide Web connected through the Internet via Social Networking Websites. With the explosion of data and information on the Web, it is becoming more difficult for information retrieval systems to yield information based on the users' needs. Web Searches are one of the most important and most common strategies attached to the World Wide Web. There are millions of users of the World Wide Web who search the World Wide Web for required information regularly. Today, Searching the Web has become the most common way extracting information owing to its ease of exraction and user friendliness.

A Web Search engine must relate with the strategy with which the user has written the input query. The search context must be understThe Search Engines extract the Web pages using a crawler which is a Web Spider which automatically loads the web pages. The Web Crawler is a bot which is a software by itself which proactively loads the web pages from the World Wide Web. The obtained Web Pages are then re-ranked as per the query relevance. Today the Web is available to almost everyone. This makes the Web a very large repository and contains heterogenous data. The data on the Web is Structured, Unstructured and even Semi-Structured. The information on the Web has no end and keeps growing almost every second as the Web users' keep uploading data.Only a good search engine with an efficient searching algorithm can catch relevant web pages and can satisify the users' needs.

There are several types of search engines. The most common ones are the Generic Search engines which are traditional search engines which are designed for catching relevant Web Pages as per the query. The Specialized Search engines are the ones which are designated for a particular purpose such as Search Engines for Videos alone or Search Engines for Audio content alone, etc. There are several strategies for Web Searches. The most common and the most classical of the Web Search strategies is the Keyword Driven Web Search which mainly deals with Keyword matching paradigm. Furthermore, several Graph Based Strategies like dynamic construction of graphs and graph like structures were successful in recommending Web Pages. Also, Several Statistical and Probabilistic strategies which used statistical techniques were employed. Then can hybrid methods which constituted the combination of two or more methodologies. In the present day times, with popularity of Semantic Web Technologies, several Semantic Strategies for Web Page Recommendation is needed as Semantic Web is intelligent and has a better performing quotient than the traditional World Wide Web.

ISSN 2394-3777 (Print)
ISSN 2394-3785 (Online)
Available online at www.ijartet.com
International Journal of Advanced Research Trends in Engineering and Technology (IJARTET)
Vol. 3, Special Issue 20, April 2016

The relevance with respect to the query is appreciable but the users' needs also must be satisified. To overcome the problem of yielding needed information for a specific user, personalization of Web Search is definitely the key concept that needs to be implemented. A personalized Web Search is a technique of Search which is been designed for a specific user. Personalized Web search engine serves as a cognitive bridge for the generalized web search and user specific web search.Personalization or Customization of Web Search could be a more authentic technique to improve the search potency and correctness when it comes to the relevance of search results. The search results for individuals with a totally different data search goals; the Web is one amongst the necessary applications to cater for information from the Web. Most users' experience unsatisfactory results from the Web Search engines and are unfulfilled as their data expectations are never met. In such instances, Personalization of Web Search is the only key to satisfy the needs of the users'. The index quality and the pedagogy of information usage by the search engines is the key characteristic of improving the quality of the search results. Personalizing the web search for a specific user is the best strategy to solve this problem.

Motivation: The major motivation for the proposed work is Customizing the Web Search Engine for a specific users' interest. There is a need for satisfying the users to their query searches and satisify the information needs of the users. Also there is a mandate to yield a much higher relevance of results specifically according to the query input and also the users' needs. There is a need for a semantic based Personalization of Web Searches for the users' interest in the era of Semantic Web. Lastly, it is highly recommended to improve on the quality of Web Searches and yield results with less noise and more accuracy. The overall quality of the search results must always be kept high.

Contribution: A Semantically driven strategy for personalization of Web Image Searches is proposed. The extraction of the user preferences from Web Usage Data and Semantically Analyzing the same for customized web search results is proposed. The Semantic Heterogenity is determined based on a strategy of Latent Semantic Analysis. The Semantic Similarity is measured between the query term and the occurance of the similar term in the Web Usage Information of a specific user.An algorithm for Semantically Driven Personalized of Web Search is proposed. The overall Average Precision of the System is increased.

Organization*:* The remaining paper organization is as follows. The Section 2 provides a brief overview of Related Literature of the research conducted.Section 3 presents the System Archiecture. Section 4 decscribes the implementation in detail. Strategy Incorporated is discussed in Section 5.Section 6 presents the Results and Performance Evaluation. Finally, Conclusions are depicted in section 7.

2. RELATED LITERATURE

Dou et al.,[1] through his study investigated whether personalization of web search is consistently effective or not. Specifically, he studied it on different queries for different users, and under different search contexts. Dou also put forth the idealogy of how personalization of Web Searches will increase the overall relevance of the search results as per the users' search expectations.Tan et al., [2] used the long term search history which contained rich information about a user's search preferences. They prodosed statistical language modelling based methods to mine contextual information from long-term search history. The experimentation which were carried on a web search test depicts that the algorithms are effective in improving retrieval accuracy for new as well as repetitive queries. The inference was that the best performance is achieved when using the combination of related and as well as past searches through related data.

Xu et al., [3] proposed a personalized web search for searching by task based results for users with single goals. Users show a much secure preference details in search engines. This paper presents about users rich profiles which automatically builds uo to change for users.This methodology has a few disadvantages like populating the result space with irrelevant and noisy search results.Ramanathan et al., [4] proposed personalized information retrieval and search methodology for improving the overall Internet experience. An important requirement for building personalized web applications is by building user profiles which represent the users 'interests. Two representations commonly used for user profiles. This methology is computationally expensive as the complexity in the web search is higher.

Shen et al., [5] studied that the major limitation of most existing and retrieval models and systems is that the retrieval decision is made based solely on the query and document collection information. The information about the actual user and search context is largely ignored. They also studied on how to exploit implicit feedback information, including previous queries and click through information, to improve retrieval accuracy in an interactive information retrieval setting. Ramya et al., [6] proposed a personalized internet search by adopting the meta search approach that replies on one among the meta search engines like Bingo, Yahoo and google. The client receive's the request from the user's and submit to the server and displays the results based on his/her profile details and favourite search history. The server manages the tasks and forwards the request to search engines. The user details are stored in user profile that preserve the privacy. That makes client-server model to communicate in a faster way and provides more efficiency in results based on the user query.

ISSN 2394-3777 (Print)
ISSN 2394-3785 (Online)
Available online at www.ijartet.com
International Journal of Advanced Research Trends in Engineering and Technology (IJARTET)
Vol. 3, Special Issue 20, April 2016

Stibu et al., [7] proposed a personalized web search which is used to improve the web search services on the internet. But still users facing a problem to get relevant data based on key words and less effectiveness as well. To overcome those problem they designed a frame work called UPS. The designed framework helps to generalize the profiles and improve the efficiency for getting the required search in a faster way based on the user interest.Sun et al., [8] have proposed a unique methodology of personalized web search based on clickthrough data.The relationship between the various objects in the click-throigh data is analyzed to determine the actual interest of the user.A CubeSVD approach is implemented in this paper for personalization of Web Searches based on tensor structure reconstruction strategy. Christo Ananth et al. [9] discussed about a method, Wireless sensor networks utilize large numbers of wireless sensor nodes to collect information from their sensing terrain. Wireless sensor nodes are battery-powered devices. Energy saving is always crucial to the lifetime of a wireless sensor network. Recently, many algorithms are proposed to tackle the energy saving problem in wireless sensor networks. There are strong needs to develop wireless sensor networks algorithms with optimization priorities biased to aspects besides energy saving. In this project, a delay-aware data collection network structure for wireless sensor networks is proposed based on Multi hop Cluster Network. The objective of the proposed network structure is to determine delays in the data collection processes. The path with minimized delay through which the data can be transmitted from source to destination is also determined. AODV protocol is used to route the data packets from the source to destination.

Shafiq et al., [10] have proposed a methodology of personalization of web search using Online Social Network analysis.The preferences of the users are obtained analytically by mining the Social Network Behaviour of Users as well as their interest groups and communities and Web searches are driven based on these dynamically extracted preferences of users.

Soldaini et al., [11] have proposed a domain based personalization of Web Search based on a technique of query clarification. Strategies like mapping synonums and task based retrieval are incorporated to clarify the users' queried to provide more relevant and accurate results as per the users' choices and preferences.Wang et al., [12] have proposed a strategy of recommending the preferences itself for personalized Web Searches to increase the appropriateness of the results. Conditional Preference Networks are used for recommending the preferences more explicitly and improve the accuracy of the personalized web searches.The user preferences are first obtained before recommending the Web Pages to the users.

Rophie et al., [13] have proposed an innovative methodology for personalization of mobile search engines using based on drilling through the clickthrough information.An ontological approach is used for re-ranking the final search results.An inventive methodology for customizing the web searches is proposed here. The re-ranking methology proposed is not very innovative but it definitely focuses on the complete analysis of the clickthrough data for better results.Kakulapati et al., [14] have proposed a unique strategy of privacy incorporated personalized web search using a re-ranking approach. An RPS framework is proposed here in order to recommend web based based on a specific re-ranking scheme. The greatest challenge here was to incorporate privacy which was done based on the previous existing algorithms like the GreedyIL.

Goel et al., [15] have proposed an innovative strategy of multi-user search scheme for autocomplete query.Concepts like token generation and identifier determination are incorporated for completing the query automatically based on Personalized Web format. This approach is quite innovative and mainly focuses on the completion of the query auomatically rather than on personalizing the overall web search experience.Abu-Dalbouh et al., [16] have proposed a unique methodology of incorporating end user privacy in Personalized Web Search Systems. A study is conducted and end to end privacy in human computer systems is proposed in this paper for better retrieval privacy in Web Search Systems. The main inference from the study is that Privacy is a criterion which must accompany the Web Searches as a Web Search without privacy is incomplete and risky.

3. SYSTEM ARCHITECTURE

ISSN 2394-3777 (Print)
ISSN 2394-3785 (Online)
Available online at www.ijartet.com
International Journal of Advanced Research Trends in Engineering and Technology (IJARTET)
Vol. 3, Special Issue 20, April 2016

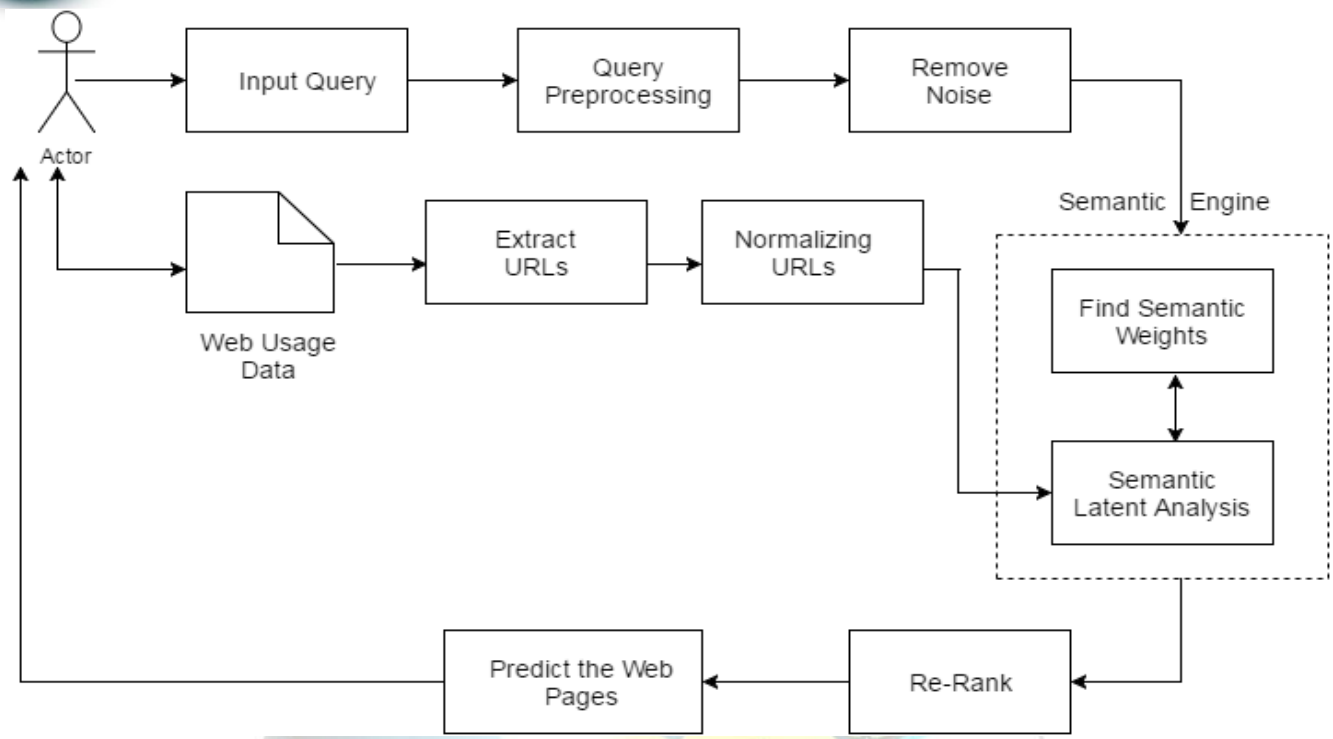

Fig 1: System Architecture

The Proposed System Architecture which is depicted in the Fig. 1 inputs the search query from the users. The Search Query is subject to Query Preprocessing where the explicit noise is removed by Parsing and Tokenization. The Stop Words in the Query are further removed through Stemming. The pre-processed query is subject to a Semantic Engine. The Semantic Engine is specific to the system which works in a two-step process. The Semantic Engine initially computes the Semantic Weights and Uses Semantic Latent Analysis to find out the Semantic Heterogeneity.The proposed methodology also involves the elicitation of the Web Usage Information from the Users' Local Web Browser.

The Web Usage Information that is extracted is the Historical Data from the Users' Browser. The Web Usage Information is extracted using a customized link crawler. The Web Usage Data is filled with several hyperlinks which are URLs. The URL data is highly linked, noisy and redundant with other associated information like the Date, Time and many more details. The only information that is required by the proposed system is that the URLs in their simplest format with the frequency of their visit. This leads to the process of normalizing the URLs where the unwanted links in the Data is eliminated and the frequency of their visit is also computed.
Furthermore, the URLs are parsed and fed into the Semantic Engine for Latent Semantic Analysis. Based on the Semantic Similarity Measure, the deviations is computed in the Search Query and the available URL data, Considering 0.25 as the threshold value for semantic heterogeneity, the Search Content is decided from the List of URLs. Finally, the URL list is re-ranked based on the frequency of the visit before the Web Page Links are recommended to the User.

5. STRATEGY INCORPORATED

Latent Semantic Analysis

Table 1: Steps to Compute the Latent Semantic Heterogenity

Step 1: The Query Term *qt* that is input is used to count the term frequency in the underlying data set,ie, the Web Usage Information.
Step 2: Formulate, the Term Frequency Matrix T where the element (x,y) depicts the occurance of query term qt_x in the Web Usage Information y.
Step 3: Perform Singular Value Decomposition (SVD) to T and reduce the total number of rows preserving the column structure.
Step 4: Find the Cosine Similarity of the angle between the two resulting vectors to deduce the Semantic Similarity.

ISSN 2394-3777 (Print)
ISSN 2394-3785 (Online)
Available online at www.ijartet.com
International Journal of Advanced Research Trends in Engineering and Technology (IJARTET)
Vol. 3, Special Issue 20, April 2016

The strategy incorporated is the Latent Semantic Analysis or the Semantic Latency Approach to compute the Semantic Similarity between a the query term and a set of terms which are extracted from the Web Usage Information. Semantic Latency measurement is a methodology of Distributional Semantics where a Term Count Matrix is Formulated. In this work, the query term count in the Web Usage Information is counted and is represented in the form of the Matrix. Furthermore, Singular Value Decomposition technique is applied to the Term Count Matrix to minimize the number of rows in the Original Matrix. To the resulting matrix, the cosine of the resultant angle between a pair of vectors is computed to deduce the semantic similarity. The step by step procedure to compute the Latent Semantic Heterogenity

4. IMPLEMENTATION

The implementation is done in C#. Net using Visual Studio as an IDE. Basic C# String Tokenizer is incorporated for tokenization. Statistical Methods for Semantic Latent Analysis is implemented in C# for computing the Semantic Heterogeneity. Abot which is a C# based Web Crawler is incorporated for implementing Crawling. Explicit URL analysis for normalization is done based on a custom written code in C#. The experimentation was done for Web Usage Information of 10 users. The Web Usage Information was loaded into the local repository of the system developed and the user was asked to browse. They definitely experienced the relevance of web pages which they visited on regular basis. The Web Usage Data almost comprised of 104 unique URLs of User 1. User 2 had 87 unique URLs.User 3 had 293 unique URLS while User 4 and User 5 had 187 and 239 URLs respectively. User 6, User 7 and User 8 had 184, 192 and 298 unique URLs. However User 9 and User 10 had 198 and 289 unique URLs. The System Worked without any lags for the URLs mentioned. The Algorithm for the Proposed System is depicted in Table 2.

Table 2: Algorithm for Semantic Personalized Web Search

Input: Initial user specific query strings S and the Web Usage Information set WU_i.

Output: Reordered Web Page URLs based on the personal preferences of the user.

Begin

Step 1: Initialize the query string S; preprocess S to remove the redundancy by Tokenization and Stemming to yield the tokens T(S).

Step 2: Dynamically compute the Semantic Weight of the T(S) and Store it in a Hash Table.

Step 3: (a) Process set WU_i for eliminating the noise and normalize it.

(b) Find the URL frequency of the hyperlinks present in set WU_i

Step 4: Compute the Semantic Similarity between the T(S) and the URL set WU_i.

Step 5: Filter the Search Results Based on the Semantic Similarity Value and the Semantic Weights and construct the URL space vector V_{url}

Step 6: Re-Rank the final result in V_{url}.

Step 7: Recommend the Web Pages in the exact order as in V_{url}.

End

6. RESULTS AND PERFORMANCE ANALYSIS

The Personalized Web Search for 5 different users' based on their Web Usage Historical Data is tabulated in detail in table 3. The tabulation comprises of the User Details, Total Number of URLs Visited, Number Unique URLs, and Number of Queries by the users.

Table 3: Details of Users', URLs and Queries incorporated.

User Details	Total Number of URLs Visited	Number of Unique URLs	Number of Unique Queries
User 1	342	104	12
User 2	446	87	16
User 3	618	293	21
User 4	258	187	4
User 5	348	239	14
User 6	336	184	11
User 7	471	192	8
User 8	845	298	15
User 9	786	198	17
User 10	693	289	24

$$Precision = \frac{Number\ of\ relevant\ Tags\ Recommended}{Total\ number\ of\ \ Tags\ in\ Tag\ Space} \quad (1)$$

Precision was chosen as the necessary performance metric. The reason why precision was chosen is that it's a factor which directly comprehends to the relevance. Higher the precision value, better is the search result relevance and accuracy. This is the reason why precision was chosen as the primary performance metric. Precision was calculated for the Users as well as the query separately. Finally, the average of the precisions were taken inorder to arrive at the average precision. Equation (1) was used to calculate the precision.The

ISSN 2394-3777 (Print)
ISSN 2394-3785 (Online)
Available online at www.ijartet.com
International Journal of Advanced Research Trends in Engineering and Technology (IJARTET)
Vol. 3, Special Issue 20, April 2016

average precision for all the five users' are depicted in Table 4. The detailed precision for individual queries of User 4 is explained in Table 5. User 4 is chosen as the number of queries which was used by him is much lesser than all the users'.

Table 4: Average Precision for Individual Users'

User Details	Average Precision
User 1	0.72
User 2	0.77
User 3	0.71
User 4	0.79
User 5	0.75
User 6	0.71
User 7	0.74
User 8	0.70
User 9	0.78
User 10	0.73
Average	**0.74**

Table 5: Query Details in Detail for User 4

Query	Precision
Mobile Computing	0.74
Genius	0.79
Indians in America	0.85
Christ the King	0.78
Tea Estates in the World	0.81
Blue Whale Size	0.84
Indian Pranksters	0.72
Technology Today	0.79
Word of God	0.84
Women in Power	0.81
Average	**0.81**

.

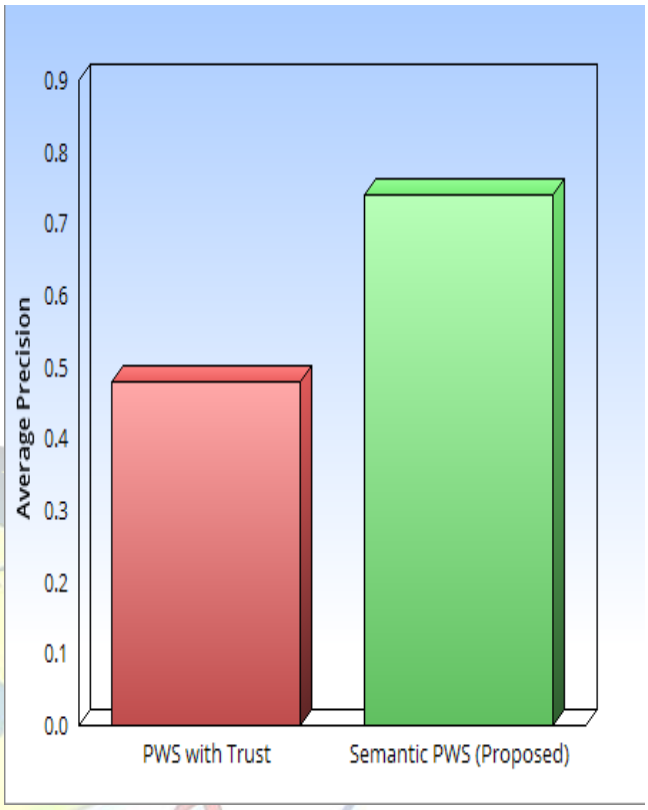

Fig 2: Comparison of the Proposed System with the existing systems.

The Comparison of the average precision of the proposed system and the existing PWS system with trust [17] is graphically depicted in Fig 2. The reason for a much higher average precision value in the Proposed System is that a Semantic Web Approach is followed. When the Semantic Value which is the Latent Semantic Value is considered, the semantic heterogeneity becomes the computation factor for filtering the search results.Moreover, the Normalization of the URL removes excess noise and deviations and thereby increasing the average precision value of the proposed system. The average precision of the existing PWS with trust is 0.48 whereas the average precision of the proposed system is 0.74. This clearly indicates that the proposed system outperforms the PWS with trustThe screenshots of the implementation of the proposed system is depicted in Fig 3. The figure shows how the results which are URL hyperlinks are obtained after the Semantic Latent Analysis in the URL Space Vector are displayed.

ISSN 2394-3777 (Print)
ISSN 2394-3785 (Online)
Available online at www.ijartet.com
International Journal of Advanced Research Trends in Engineering and Technology (IJARTET)
Vol. 3, Special Issue 20, April 2016

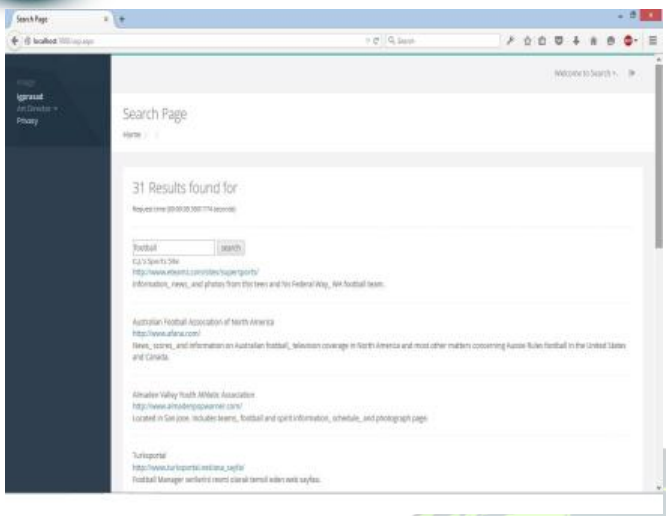

Fig 3: Screenshots of Implemented System

7. CONCLUSIONS

The problem of increasing the relevance in yielding the needed and useful information to the users' is overcome in this methodology. In this paper, the concept of URL Normalization and Semantic Latency Analysis is proposed for finding semantic similarity between the input query and the Web Usage Information. An Algorithm for Semantics Based Personalized Web Search is proposed. The problem of inefficient query searches over historical data is overcome in the proposed system by re-ranking of hyperlinks based on the relevance. An Average Precision of 0.74 is achieved in the proposed method which is much better than the existing systems like PWS with Trust whose average precision is just 0.48.The possible future work for this can be inclusion of privacy into this methodology of Web search. Other possible enhancements could be increasing the average precision value for the semantic based approach.

ACKNOWLEDGEMENT

I take this opportunity to thank God the Almighty and Eternal Father for giving me strength and grace to complete my work. I thank my parents and sister who supported me to do this work. I thank all my friends who directly and indirectly were involved in publishing this work.

REFERENCES

[1] Z. Dou, R. Song, and J.-R. Wen, "A Large-Scale Evaluation and Analysis of Personalized Search Strategies," Proc. Int'l Conf. World Wide Web (WWW), pp. 581-590, 2007.

[2] B. Tan, X. Shen, and C. Zhai, "Mining Long-Term Search History to Improve Search Accuracy," Proc. ACM SIGKDD Int'l Conf.Knowledge Discovery and Data Mining (KDD), 2006.

[3] Y. Xu, K. Wang, B. Zhang, and Z. Chen, "Privacy-Enhancing Personalized Web Search," Proc. 16th Int'l Conf. World Wide Web(WWW), pp. 591-600,2007.

[4] K. Ramanathan, J. Giraudi, and A. Gupta, "Creating Hierarchical User Profiles Using Wikipedia," HPLabs, 2008.

[5] X. Shen, B. Tan, and C. Zhai, "Privacy Protection in PersonalizedSearch," SIGIR Forum, vol. 41, no. 1, pp. 4-17, 2007.

[6] Ramya V and Gowthami S, "Enhance Privacy Search in Web Search Engine using Greedy Algorithm," in International Journal of Scientific Research Engineering & Technology (IJSRET) Vol 3 No. 8, pp.1106-1109, 2014.

[7] Stibu Stephen and A Venugopal, "Supporting privacy protection in personalized web search for knowledge mining" in International Journal of Advanced Research in Computer and Communication Engineering Vol. 4, Issue 1,pp. 75-78, 2015.

[8] Sun, Jian-Tao, Hua-Jun Zeng, Huan Liu, Yuchang Lu, and Zheng Chen. "Cubesvd: a novel approach to personalized web search." In Proceedings of the 14th international conference on World Wide Web, pp. 382-390. ACM, 2005.

[9] Christo Ananth, T.Rashmi Anns, R.K.Shunmuga Priya, K.Mala, "Delay-Aware Data Collection Network Structure For WSN", International Journal of Advanced Research in Biology, Ecology, Science and Technology (IJARBEST), Volume 1,Special Issue 2 - November 2015, pp.17-21

[10] Shafiq, Omair, Reda Alhajj, and John G. Rokne. "On personalizing Web search using social network analysis." Information Sciences 314 (2015): 55-76.

[11] Soldaini, Luca, Andrew Yates, Elad Yom-Tov, Ophir Frieder, and Nazli Goharian. "Enhancing web search in the medical domain via query clarification." Information Retrieval Journal (2016): 1-25.

[12] Wang, Hongbing, Shizhi Shao, Xuan Zhou, Cheng Wan, and Athman Bouguettaya. "Preference recommendation for personalized search." Knowledge-Based Systems (2016).

[13]Rophie, A. Smilien, and A. Anitha. "User Preferences Based Personalized Search Engine." In the International Journal of Research in Computer Applications and Robotics, Vol. 4, Issue 3 pp.6-10, (2016).

[14] Kakulapati, Vijayalakshmi, and Sunitha Devi Bigul. "A Re-ranking Approach Personalized Web Search Results by Using Privacy Protection." InInformation Systems Design and Intelligent Applications, pp. 77-88. Springer India, 2016.

[15] Goel, Samir, Franck Chastagnol, and Abhishek Agrawal. "Multi-user Search System with Methodology for Personalized Search Query Autocomplete." U.S. Patent 20,160,055,185, issued February 25, 2016.

[16] Abu-Dalbouh, Hussain. "Implementing End-User Privacy through Human Computer Interaction for Improving Quality of Personalized Web." Computer and Information Science 9, no. 1 (2016): 75.

[17] Suruchi Chawla. Article: Trust in Personalized Web Search based on Clustered Query Sessions. International Journal of Computer Applications 59(7):36-44, December 2012.

Harvard