2016 IEEE International Conference on Advances in Computer Applications (ICACA)

A Differential Semantic Algorithm for Query Relevant Web Page Recommendation

Gerard Deepak
Dept of Computer Sci. & Engg.
Bangalore University
Bangalore, India

J Sheeba Priyadarshini
Dept of Computer Science
St. Josephs College
Bangalore, India

M S Hareesh Babu
University Visvesvaraya College of Engg
Bangalore University
Bangalore, India

***Abstract*— With the exponential rise in the amount of information in the World Wide Web, there is a need for a much efficient algorithm for Web Search. The traditional keyword matching as well as the standard statistical techniques is insufficient as the Web Pages they recommend are not highly relevant to the query. With the growth in Semantic Web, an algorithm which semantically computes the most relevant Web Pages is required. In this paper, a methodology which computes the semantic heterogeneity between the keywords, content words and query words for web page recommendation is incorporated. A Differential Adaptive PMI Algorithm is formulated for with varied thresholds for recommending the Web Pages based on the input query. The proposed methodology yields an accuracy of 0.87 which is much better than the existing strategies.**

Keywords—Adaptive PMI; Differential; Personalized Web Search Semantic Web; Web Page Recommendation

I. INTRODUCTION

The trends in Internet and Social Networking have increased the number of users' as well as data in the World Wide Web. A Web Search is the standard retrieval methodology of required information from the World Wide Web via a Web Search Engine. Several search engines are algorithm driven and execute a search algorithm through the Web Crawler. The Web is the World's largest storehouse of data with the highest degree of heterogeneity. Traditional Search Engines use existing statistical techniques for computing the probability of similarity and rank the web pages. A large number of search engines rely on keywords and recommend web pages based on keyword matching.

A search engine is graded by two main parameters namely relevance of information retrieved and the response time. An optimal search engine is said to have a very high efficiency if it furnishes information that is relevant to the query and has a very high response time. The World Wide Web is transforming into a perspicacious Semantic Web. Semantic Web is evolving from the existing Web [1]. A semantics based algorithm for Web Page Recommendation would be apt for the scenarios of the growing Semantic Web. Most Search Engines are designed for Personalized Web Search where the results are based on the query relevance or user profile relevance. A Personalized Web Search is a web information retrieval strategy where the search algorithm analyzes users' profile as well the query. User profiles are generally the Web Usage Data or Users' Browsing History which are analyzed for furnishing customized results.

A triadic similarity computation approach which checks the semantic similarity between the URL keywords, content words as well as the input query words is required for increasing the relevance of results. The World Wide Web is an enormous information source which makes it highly difficult to extract the required contents and satisfy the users' needs. Henceforth, a technique which personalizes the Web Search by analyzing the Web Usage data as well as the Web Page Contents would be an optimal technique for Web Page Recommendation. Web Mining Plays a vital role in several individual's life as extracting the exact required information from such an extensive and dynamically expanding Web is a challenge.

Motivation: Due to the density of data in the World Wide Web, it is quite essential to yield the most relevant data for a query. Traditional strategies like keyword based matching and usage of the age-old statistical methods to extract the required information from the Web have become obsolete and redundant. With the paradigm shift towards a much intelligent Semantic Web, there is a need for semantic methodologies to be imbibed into the process of information retrieval. Although, several semantic methods are available for semantic heterogeneity computation in information retrieval from the Web, their placement in the algorithms and their usage plays a vital role in increasing the overall relevance of the results. This further affects the overall performance of the information retrieval system.

Contribution: An innovative framework for yielding the most relevant web pages is proposed. A novel strategy called as Differential Adaptive Pointwise Mutual Information is proposed for computing the semantic heterogeneity which is one of the primary contributions to this work. The query words are used for extraction of the relevant URLs from the URL repository. From the URL structure, the keywords and content words are extracted. The semantic similarity is computed between the keywords and the content words to obtain a feasible word set. Additionally, the semantic heterogeneity is computed between the query words and feasible word set differentially to re-rank the URLs before they are yielded to the user. A higher precision, recall and accuracy are achieved for the proposed methodology.

Organization: The remaining of this paper is organized as follows. The Section II provides a brief overview of Related Work. Section III depicts the Proposed Architecture. Section IV discusses the Implementation in detail. The Performance

978-1-5090-3770-4/16/$31.00©2016 IEEE

Analysis & Results is discussed in Section V. Section VI concludes the paper.

II. RELATED WORK

Harish et al., [2] have proposed a content based relevancy algorithm where the web pages are recommended by matching the keywords and the content words using traditional comparison for matching. The algorithm adds weights for the subsequent matches and then re-ranks the Web Pages which is not a semantic methodology. Maratea et al., [3] have proposed a heuristic driven strategy that incorporates majority intelligence strategy which matches a normal human behavior. Niva et al., [4] have proposed a methodology for Web Page Recommendation that uses folksonomies and social bookmarks for recommendation of Web Pages. Nguyen et al., [5] have proposed a Web Page Recommendation system that recommends the Web Pages based on Domain Knowledge and the Web Usage Data. This methodology is a semantic strategy for Web Page Recommendation that furnished Web Pages Based on conceptual prediction for constructing a Semantic Network automatically.

Guduz et al., [6] have proposed a methodology for Web Page Recommendation based on a click stream tree formulation. The information of the order of Web Pages as well as the time spent on the links is considered as major criteria for Web page recommendation in this approach. Rose et al., [7] have proposed a methodology for Web Page Recommendation using weighted genetic algorithm. The methodology also uses a semantic keyword generation through Word Net and ontologies. Karunkuzhali et al., [8] have proposed a novel methodology for Web Page recommendation by building a metaphysics using the terms aggregated from documents as well as the Web Usage Data. Neelima et al., [9] have predicted the users' behavior by analyzing the Web Log Data for several sessions of the user using usability analysis.

Forsati et al., [10] have put forth a binary data clustering technique for Web Page Recommendation by partitioning of the binary session data into a number of clusters. This methodology also incorporates k-means algorithm to enhance the quality of the results. Mishra et al., [11] have proposed a strategy for Web Page Recommendation using sequential information. This methodology amalgamates similarity upper approximation with singular value decomposition to enhance the quality of Web Page Recommendation. Mary et al., [12] have proposed a unique methodology for Web Page Recommendation which uses genetic algorithm as a core strategy. This method constructs the recommendation tree for furnishing the Web Pages. Romsaiyud et al., [13] have amalgamated Social Tagging and Collaborative Web Search for enhancing the recommendation of Web Pages.Agarwal et al., [14] have proposed a methodology that incorporates fusion of ranks from the results of Multiple Results where a tool is developed to aggregate several search engines and the query is sent to all search engines. In Rank Fusion, the rank is assigned to the search engines and the best results are reordered before they are presented to the user.

III. PROPOSED ARCHITECURE

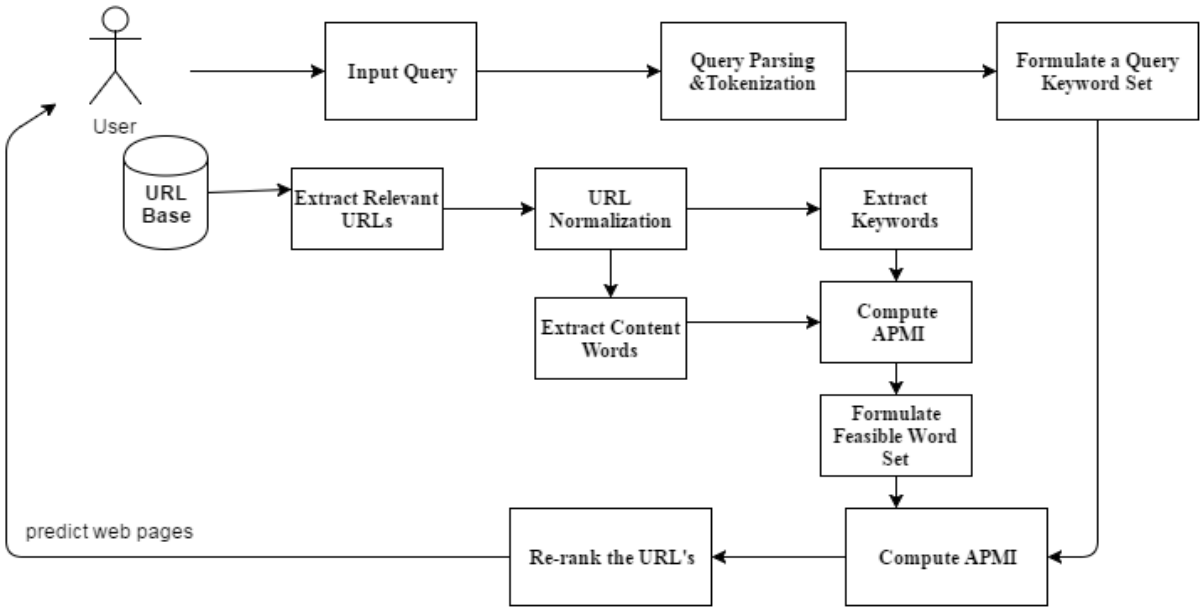

Fig. 1. Proposed System Architecture

The architecture of the proposed system is depicted in Fig. The query that is input from the user is preprocessed at first. Query Preprocessing involves parsing and tokenization of the multiword query. Furthermore, the query processing is subjected to stemming where the removal of stop words takes place. The redundant words in the query are eliminated. A query keyword set is formulated which comprises of unique query words. The proposed system architecture incorporates a

URL Base which is a very large repository that houses a large volume of URLs. The URLs in the URL Base are collected from several web sources. There are individual URLs and also the Web Log Information of various users' which are aggregated together to constitute a URL Base.

The proposed methodology extracts the query relevant URLs using the formulated set of query words. The URLs extracted are further processed to remove irrelevant URLs. The URLs are parsed and normalized. URL Normalization refers to removal of the additional continued indices to a specific URL. URL Normalization also removes the duplicates and redundancies corresponding to a particular URL. The URL structure is elicited to extract the keywords and content words. The keywords are extracted from the HTML title tag and the content words are obtained from the HTML body tags [2]. The Semantic Heterogeneity between the keywords and the content words are computed using the proposed Adaptive Pointwise Mutual Information strategy. Based on the Semantic Heterogeneity value computed between the keywords and content words, a feasible word set is formulated.

The Semantic Similarity is once again computed using the Adaptive Pointwise Mutual Information strategy between the terms in the feasible word set and the terms in the query keyword set. This computation is carried out in order to obtain the URLs with a higher relevance to the query. Furthermore, the URLs are ranked based on the scores of the final semantic similarity. The finally ranked URLs which are the Web Page Links with high relevance to the input query are recommended to the user.

The incorporated strategy is the Adaptive Pointwise Mutual Information where the existing Pointwise Mutual Information method is modified with certain parameters in order to increase the accuracy of semantic similarity between a set of terms. The proposed algorithm is termed as the Differential APMI algorithm as the semantic similarity is computed twice using the Adaptive PMI strategy with heterogeneous thresholds. The PMI between a pair of terms '*m'* and *'n'* is given by equation (1) which is an existing PMI method to compute semantic similarity. The Adaptive PMI measure depicted in equation (2) is the proposed strategy for computing the semantic heterogeneity.

$$pmi(m;n) = h(m) + h(n) - h(m,n) \qquad (1)$$

$$APMI(m;n) = \frac{pmi(m;n)}{p(m)(n)} + y \qquad (2)$$

$$y = \frac{1 + \log[p(m,n)]}{p(n)\log[p(m)] - p(m)\log[p(n)]} \qquad (3)$$

The Adaptive Pointwise Mutual Information measure is an enhanced version of the PMI measure to compute the semantic similarity. The APMI is much better than the other variants of the PMI namely the Normalized PMI strategy as it is associated with an adaptive coefficient y. The adaptive co-efficient y depicted in equation (3) is associated with a logarithmic quotient in its numerator and its denominator. The adaptive coefficient when coupled with the pmi value enhances the overall performance of the system. APMI increases the confidence in computing the semantic heterogeneity and thereby increases the overall relevance of the pages that are returned to the user.

IV. IMPLEMENTATION

The system was implemented in JAVA using Netbeans as the IDE. MYSQL was used as a backend database to store the URLs and formulate a URL base which comprised of highly diverse URLs of several topics which were correlative. The URLs that were stored in the repository served as the major data set. These URLs were randomly collected from the World Wide Web for a specific topic by the usage of standard web browsers and search engines. Also, in order to increase the complexity of data sets, URLs that were a part of several users' web usage data were collected and added into the URL repository.

A Java HashSet from the JAVA Collections framework was used to store the query word set. Similarly the feasible word set is also stored and processed using the HashSet. A HashMap is used for storing the corresponding URLs along with the semantic similarity value as a key-value pair. The key is the semantic similarity and the value is the corresponding URLs. Further an iterator is used for ranking the URLs by sorting the HashMap contents by simple sorting technique. Equations (1), (2) and (3) are realized in JAVA where the PMI value is first computed for a pair of terms. Once the PMI is obtained for a pair of terms, then the adaptive co-efficient y is computed. Finally the APMI value is calculated using (2). Based on the APMI value, the URLs are re-ranked and presented to the user.

The algorithm for the proposed strategy is depicted in Table 1. The algorithm is termed differential because the APMI value is computed twice with two different thresholds. Firstly the APMI is computed between the Keywords and the Content words. The first threshold is assumed as 0.33 for the reason of allowing the 1/3 probability. This value is not chosen as 0.5 since it allows even irrelevant terms to a larger extent. Furthermore, for the second time the threshold is chosen as 0.2 for the reason that this threshold must be lesser than the ¼ probability. Hence, a random value lesser than 0.25 is chosen. The threshold is reduced in order to increase the relevance to a higher extent and yield the result which has a very high degree of relevance to the query specified by the user.

TABLE I. DIFFERENTIAL ADAPTIVE PMI ALGORITHM

Input: Query Q that is input by the user

Output: Ranked and Highly Relevant URLs

Step 1: Input a Query Q

Step 2: Using a StringTokenizer Tokenize the query Q.

Step 3: while(tokens !=NULL)

HashSet Q_s ← tokens

Step 4: for each element t of Q_s
extract relevant URLs R_u
Parse , Normalize R_u
end for
Step 5: for each Normalized R_u
HashSet K ← extract (Keywords)
HashSet C ← extract (Content Words)
k← K.iterator();
c← C.iterator();
pmi (k, c) = h' (k) + h' (c) – h (k, c)
y=1+log[p(k,c)]
apmi(k,c)=pmi(k,c)/p(k) (c) + y
end for
Step 6: if(apmi < 0.33)
Formulate Feasible Word Set F_{ws} ←(k,c)
Step 7: for each instance of Q_s**,** F_{ws}
q← Q_s.iterator();
f← F_{ws}.iterator();
pmi (q, f) = h' (q) + h' (f) – h (q, f)
y1=1+log[p(q,f)]
apmi(q,f)=pmi(q,f)/p(q) (f) + y1
end for
Step 8 : if(apmi < 0.20)
HashMap←(apmi, f)
Step 9 : Sort the f as per apmi
Step 10 : Predict the URLs containing f

Step 11: Stop

V. PERFORMANCE ANALYSIS AND RESULTS

The experimentation was carried out and the URLs for several topics were stored in the URL Base. The details of the topics and the number of URLs collected are depicted in Table 2. The Direct URLs refer to the URLs having the terms in the topic as one of the keywords. The Indirect URLs refer to those URLs where they are related to the topic. However, these terms may be a part of their content. These URLs were extracted by the use of Search Engines and Crawlers and were manually fed into the URL Base which is a MYSQL database of the URLs.

TABLE II. DATA SETS USED FOR EXPERIMENTATION

Topics of Data Set	Number of Direct URLs	Number of Indirect URLs
Graphics	6	26
Population	12	36
Dance	5	24
Language	10	28
Movies	16	36
Universities	54	108

$$\text{Precision} = \frac{No.of\ URLs that are retrieved and relevant}{Total No.of\ URLs that are retrieved} \quad (4)$$

$$\text{Recall} = \frac{No.\ of\ URLs\ that\ retrieved\ and\ relevant}{Total\ No.\ of\ \ URLs\ that\ are\ relevant} \quad (5)$$

$$\text{Accuracy} = \frac{Precision + Recall}{2} \quad (6)$$

The Performance of the proposed system is evaluated by considering the Precision, Recall and Accuracy as the Metrics. These metrics are chosen for performance evaluation as the correctness of the methodology as well as the relevance of the results is measured in this approach. The experimentation is done specifically for five different queries. The Precision, Recall and Accuracy achieved is depicted in Table 3. Precision is depicted in equation (4) is defined as the ratio of the number of URLs retrieved and relevant to the total number of URLs that are retrieved. Recall is defined as the ratio of the number of URLs that are relevant and retrieved to the total number of relevant URLs. Recall is depicted in equation (5). Accuracy is defined as the average of precision and recall and is depicted in equation (6).

TABLE III. PERFORMANCE MEASURES OF DIFFERENTIAL APMI ALGORITHM

Query	Precision	Recall	Accuracy
Population Explosion in India	0.84	0.88	0.86
Recent Movie Reviews	0.83	0.86	0.85
Dance forms of India	0.86	0.86	0.86
State Universities in India	0.87	0.9	0.89
Languages of our Country	0.84	0.89	0.87
Average	**0.85**	**0.88**	**0.87**

In order to compare the proposed system with the other existing works, the system is tested for the Query *"Population Explosion in India"* as the existing work [2] on similar grounds is tested for the same query. It is clear that for the query *"Population Explosion in India"* the proposed algorithm, yields a Precision of 0.84 whereas an existing Content Based Relevancy Algorithm [2] yields a precision of 0.75 for the same query. Figure 2 indicates the Precision Comparison specific to the query *"Population Explosion in India"* for the Content Based Relevancy Algorithm [2] and the Proposed Differential APMI Algorithm for Web Page Recommendation. In order to further compare the overall

performance of the proposed algorithm, the Content Based Relevancy Algorithm [2] was implemented for all the Queries in Table 3 with the same URL data sets used in the proposed system.

TABLE IV. EVALUATED PERFORMANCE MEASURES OF CONTENT BASED RELEVANCY ALGORITHM [2]

Query	Precision	Recall	Accuracy
Population Explosion in India	0.75	0.8	0.78
Recent Movie Reviews	0.76	0.8	0.78
Dance forms of India	0.72	0.78	0.75
State Universities in India	0.78	0.83	0.81
Languages of our Country	0.77	0.81	0.79
Average	**0.76**	**0.8**	**0.78**

From Fig. 2 it is clearly inferable that the Precision for the Query *"Population Explosion in India"* for the Proposed Differential APMI Algorithm is 0.84 whereas the precision for the existing Content Based Relevance Algorithm is 0.75. The reason for the higher precision value for the proposed methodology is that a purely semantic approach is incorporated. The adoption of an Adaptive PMI computation technique increases the overall performance of the proposed system. Table 4 depicts the performance evaluation measures for the existing Content Based Relevancy Algorithm [2]. The precision, recall and accuracy of Content Based Relevancy Algorithm [2] were computed by actual implementation of the Content Based Relevancy Algorithm [2] considering the same data sets that was used by our algorithm. Also, the same queries were used for recording the performance measures.

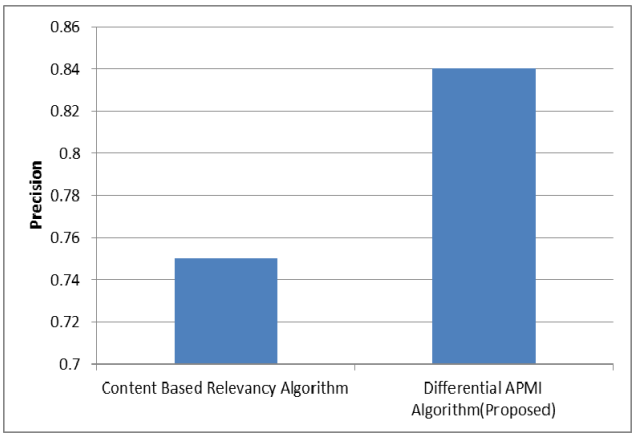

Fig. 2. Precision Comparison for Query "Population Explosion in India"

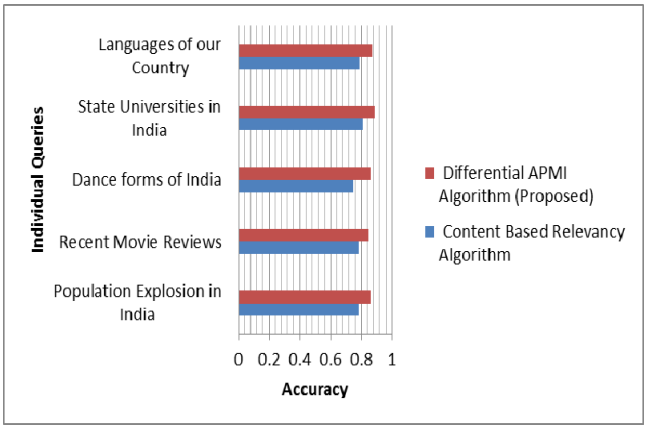

Fig. 3. Comparison of the Accuracy of Differential APMI Algorithm and Content Based Relevancy Algorithm

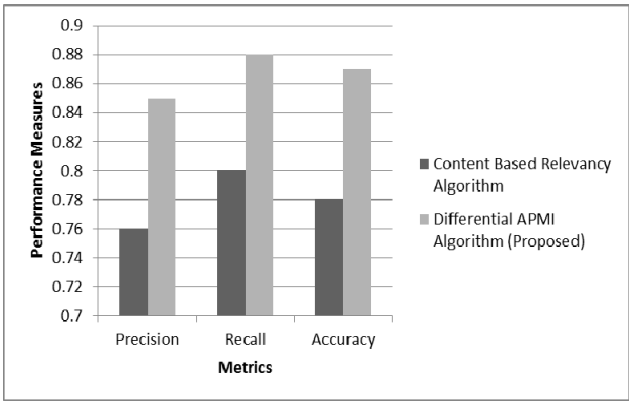

Fig. 4. Comparing the Average Performance of Differential APMI and Content Based Relevancy Algorithm

The Accuracy of individual queries for Differential APMI Algorithm as well as Content Based Relevancy Algorithm is depicted in Fig. 3. It is clearly evident that the proposed Differential APMI Algorithm yields a higher accuracy for all the queries in comparison to the Content Based Relevancy Algorithm [2]. A higher accuracy measure is achieved by the proposed algorithm because it follows a semantic strategy for eliminating the irrelevant data elements. This further increase the relevancy of the results yielded. Besides, the strategy of using differential thresholds makes the accuracy much higher making the proposed methodology reliable and robust for Web Page Recommendation. Figure 4 indicates the average Precision, Recall and Accuracy of Differential APMI and Content Based Relevancy Algorithm [2]. It is clearly evident that the Proposed Differential APMI Algorithm outperforms the Content Based Relevancy Algorithm [2].

VI. CONCLUSIONS

A novel and an efficient strategy for Web Page Recommendation that uses a triadic differential strategy is

proposed. The proposed approach has improvised the Content Based Relevancy Algorithm [2] which matches the keywords and the content words. A Differential Adaptive PMI algorithm that computes the semantic similarity between the query words, keywords and content words differentially with heterogeneous thresholds is implemented. The keywords and the content words are obtained from the URLs in the URL Base which contains Web Usage Information. A new strategy of Adaptive PMI is proposed for semantic similarity computation. The approach computes the semantic heterogeneity at varied thresholds to ensure that the Web Pages recommended are highly relevant to the query. Also, the strategy that is formulated is a semantic methodology for Web Page Recommendation. The proposed algorithm yields an overall precision of 0.85, an overall recall of 0.88. An overall accuracy of 0.87 is achieved which is much better than the previous strategies.

ACKNOWLEDGEMENT

I thank God the Almighty and Eternal Father for having given me the strength and wisdom to carry out this research work. I thank my parents for having supported and encouraged me during the course of this research.

References

[1] Antoniou, G. and Van Harmelen, F., "A Semantic Web Primer," in MIT press, 2014.

[2] Harish Kumar B T, Vibha Lakshmikantha and Venugopal K R "Content Based Web Page Re-ranking Using Relevancy Algorithm," in Quest Journals Journal of Electronics and Communication Engineering Research, vol 2, no. 7, 2014.

[3] Maratea, A. and Petrosino, A., November "An Heuristic Approach to Page Recommendation in Web Usage Mining," in the Proceedings of Ninth IEEE International Conference on Intelligent Systems Design and Applications, pp. 1043-1048, 2009.

[4] Niwa, S., Doi, T. and Honiden, S. " Web Page Recommender System Based on Folksonomy Mining for ITNG'06 Submission," in the Third International Conference on Information Technology: New Generations (ITNG'06), pp. 388-393, 2006.

[5] Nguyen, T.T.S., Lu, H.Y. and Lu, J., "Web-page Recommendation Based on Web Usage and Domain Knowledge," In IEEE Transactions on Knowledge and Data Engineering, vol. 26, no. 10, pp.2574-2587, 2014.

[6] Gündüz, Ş. and Özsu, M.T., August "A Web Page Prediction Model Based on Click-Stream Tree Representation of User Behavior," in Proceedings of the ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining ,pp. 535-540, 2003.

[7] Rose, J.D., Komala, J. and Krithiga, M., "Efficient Webpage Retrieval Using WEGA," in Procedia Computer Science, 87, pp.281-287, 2016.

[8] Karunkuzhali, D., Sangavee, R., Sndiya, K., Vidhya, S. and Nivetha, R., "An Approach for Webpage Ranking Using SEO Suggestor and SEO Analysis Tool," in Imperial Journal of Interdisciplinary Research, vol. 2, no.5, 2016.

[9] Neelima, G. and Rodda, S., "Predicting User Behavior Through Sessions Using the Web Log Mining," in the Proceedings of 2016 IEEE International Conference on Advances in Human Machine Interaction (HMI), pp. 1-5, 2016.

[10] Forsati, R., Moayedikia, A. and Shamsfard, M., "An Effective Web Page Recommender Using Binary Data Clustering," in Information Retrieval Journal, vol. 18, no.3, pp.167-214, 2015.

[11] Mishra, R., Kumar, P. and Bhasker, B., "A Web Recommendation System Considering Sequential Information," in Decision Support Systems, vol. 75, pp.1-10,2015

[12] Mary, P.S. and Baburaj, E.,. "Constraint Informative Rules For Genetic Algorithm-Based Web Page Recommendation System," Journal of Computer Science, vol.9, no.11, pp.1589, 2013.

[13] Romsaiyud, W. and Premchaiswadi, W., 2011, "Exploring Web Search Behavior Patterns to Personalize the Search Results," in the Proceedings of Third International IEEE Conference on Intelligent Networking and Collaborative Systems (INCoS), pp. 313-319, 2011.

[14] Agrawal, R. and Soni, R., "Rank Fusion of Results from Multiple Search Engines: An Implementation," in IEEE International Conference on Machine Intelligence and Research Advancement (ICMIRA), pp. 224-229), 2013.