International Journal of Psychosocial Rehabilitation, Vol. 24, Issue 05, 2020
ISSN: 1475-7192

User Interest Prediction Model for Hybrid Tag Recommender Automation Systems

T. Fahad Iqbal and R. Gnanajeyaraman

Abstract--- *The problem of web search has been discussed in variety of situations and there are many approaches recommended by different researchers earlier. Personalized recommender systems aim to help users access and retrieve relevant information or items from large collections, by automatically finding and suggesting products or services of likely interest based on observed evidence of the users" preferences. For many reasons, user preferences are difficult to guess, and therefore re-commender systems have a considerable variance in their success ratio in estimating the users tastes and interests. In such a scenario, self-predicting the chances that a recommendation is accurate before actually submitting it to a user becomes an interesting capability from many perspectives. Performance prediction has been studied in the context of search engines in the Information Retrieval field, but there is little if any prior research of this problem in the recommendation domain. Based on computed state support measure, the method computes the interest probability and finally generates recommendations to the prediction model. The proposed method increases the efficiency of the web search and reduces the overall search time complexity.*

Keywords--- *Web Search, Web Mining, Web Search State Graph, User Interest Prediction, Hybrid Tag Recommender.*

I. INTRODUCTION

Information Retrieval (IR) technologies have gained outstanding prevalence in the last two decades with the explosion of massive online information repositories, and much in particular the World Wide Web. IR systems are researched and designed in ways that seek to maximise the degree of satisfaction of certain objective conditions, typically though not necessarily only – user satisfaction. IR research and development have revolved around the definition of models and algorithms that best achieve this goal, methodologies and metrics that let assess how well the goal is achieved by different systems, and sound theories providing a solid ground and orientation in the development of IR algorithms and their consistent evaluation. Among many new trends stemming from this main stream of research and developments, a new re-search goal started to be considered by the early 2000˝s: is it possible to predict how good a result returned by an IR system is going to be, before presenting it to the user, or even, before running the IR system at all (Cronen-Townsend et al., 2002)? This question has given rise to a fertile strand of research on so-called performance prediction in IR.

Performance prediction has many potential uses in IR. From the user's perspective it may provide valuable feedback that can be used to direct a search, from the systems perspective it may help to distinguish poorly performing queries, and from the system administrators perspective it may let identify queries related to a specific

T. Fahad Iqbal, Research Scholar, Department of Innovative Informatics, Institute of Computer Science and Engineering, Saveetha School of Engineering, Chennai, Saveetha Institute of Medical and Technical Services. E-mail: fahad32in@gmail.com
R. Gnanajeyaraman, Asso. Prof., CSE, SBM College of Engg & Tech, Dindigul. E-mail: r.gnanajeyaraman@gmail.com

DOI: 10.37200/IJPR/V24I5/PR2020131
Received: 07 Mar 2020 | Revised: 26 Mar 2020 | Accepted: 04 Apr 2020

subject that are difficult for the search engine. Performance prediction approaches are based on the analysis and characterisation of the evidence used by an IR system to assess the relevance (utility, value, etc.) of retrieval objects (documents, goods, etc.) at execution time (Cronen-Townsend et al., 2002). The most classic and basic retrieval scenario involves a user query and a collection of documents as the basic input to form a ranked list of search results, but other additional elements can be taken into account to select and rank results (Baeza-Yates and Ribeiro-Neto, 2011).

Any information the retrieval system takes as input can be taken as input for the performance prediction as well, and often the prediction methods use additional information beyond that. The user context (current tasks, query logs, preferences, etc.), global properties of the document collection, comparisons with respect to other reference elements such as historic data, and the output from other systems, among others, are some examples of the different sources of information that a predictor may draw evidence from.

Predicting the performance of a subsystem, module, function, or input by contrasting the performance estimation for a query for each component, enables an array of dynamic optimisation strategies that select at runtime the option which is pre-dicted to work best or, when larger systems or hybrid approaches are used, allows for adjusting on the fly the participation of each module. The IR field is pervaded with cases where information relevance, retrieval systems, models, and criteria are based on a fusion or combination of sub-models. Personalised retrieval systems (including techniques such as personalised search, recommender systems, collaborative filtering, and retrieval in context) are clear examples where performance prediction can be applied since such systems combine several sources of evidence for relevance assessment, such as explicit queries, search history, explicit user ratings, social information, user feedback, and context models.

Performance prediction finds additional motivation in personalised recommendation, inasmuch these applications may decide to produce recommendations or hold them back, delivering only the sufficiently reliable ones. Furher more, current Recommender Systems (RS) are characterised by an increasing diversification of the types and sources of data, content, evidence and methods, available to make decisions and build their output. In such context, predicting the performance of a specific recommendation approach or component becomes an appealing problem, as it lets properly combine the available alternatives, and make the most of them by dynamically adapting the recommendation strategy to the situation at hand. The question gains increasing relevance today, with the proliferation of hybrid recommendation techniques to improve the accuracy of the methods – the Netflix prize was a paradigmatic example of the use of this, where all the top ranked participants used combinations of large sets of recommendation methods.

This calls for the research of hybrid approaches with a level of dynamic self-adjustment mechanisms, in order to optimise the resulting effectiveness of the recommendation systems, by opportunistically taking advantage of high-quality data when available, but avoiding sticking to fixed strategies when they can be predicted to yield poor results under certain conditions.

DOI: 10.37200/IJPR/V24I5/PR2020131
Received: 07 Mar 2020 | Revised: 26 Mar 2020 | Accepted: 04 Apr 2020

Performance prediction in IR is typically assessed in terms of the correlation be-tween a predictor‟s scores and a system‟s performance values on a per-query basis. This requires reliable performance evaluation metrics and methodologies, which have been thoroughly analysed, and are currently well established in the IR field, mostly oriented to ad-hoc search. In contrast, evaluation in the RS field is more open, and the variability in evaluation approaches and experimental configurations is significant. How to measure the performance of a recommender system is a key issue in our re-search since the system quality measurements may be influenced by statistical proper-ties of the measurement approach and/or the experimental design. Throughout this thesis we shall focus on the accuracy of the system, where we have to avoid that if a metric – i.e., precision – is biased towards some form of noise along with the recommender's quality, then a predictor capturing only that noise would appear as an (equivocal) effective performance predictor. Hence, statistical biases (noises) of the evaluation methodologies should be well understood in order to enable a meaningful assessment of performance predictors. The modern search engine clubs set of features with the search results like generating advertisements in the pages of result and generates many information stamp at the web users result page. The search engine does not estimate how the generated recommendations are reputed by the users.

The users search history and their activities are monitored and generated as web log which is huge in size. The web log data set has number of information about the search history of large number of peoples which are generated at different time window. The time window can be from hour to years. There are methods to identify the user interest from the search history of particular time window of any size. For example, the concept based approaches identifies the search concept from the page content of visited web pages. Similarly the topic and concept of visited web pages are identified and using that the interest of user at any particular time window could be identified based on other features also.

But the user interest prediction is another dimension of web search which is required to improve not only the efficiency of web search but also can be used to develop the business. For example, an web user may be searching for cars with power steering at one time window, but when we look at the search history of the same user, we can identify that he would be search for the cars with other facilities like ABS, Bluetooth and other technologies. This represent the interest change and the change is happening at series of time window. The same has to be applied for the web search, where the user is searching for different contents or topics at different time window. So in order to produce efficient web search result, the search engine has to identify the transition of user interest.

At each time window or session of web search, each query can be stated as a state and all the topic of web page being visited by the user can be formed as state graph. In the search state graph Sg, there exist $K \times 2^n$ number of states and there will be a transition of state only if the user has visited the concern topic in sequence. From the web search state graph, the user interest can be identified and using them the approach can predict the future interest to improve the efficiency of web search.

DOI: 10.37200/IJPR/V24I5/PR2020131
Received: 07 Mar 2020 | Revised: 26 Mar 2020 | Accepted: 04 Apr 2020

II. RELATED WORKS

The main objective of the research presented here is to find predictive methods for the performance of specific components in recommender systems, and to improve the performance of combined recommendation methods, based on the dynamic, automatic analysis and prediction of the expected performance of the constituents of the composite methods, whereupon the relative participation of each constituent is adjusted, in accordance to its predicted effectiveness. To address these problems, this work has the following specific research objectives:

RG1: Analysis and formalisation of how retrieval performance is defined and evaluated in recommender systems. We need to develop an in-depth study on how recommender systems can be reliably evaluated in terms of numeric metric values, since we aim to predict their performance. Moreover, we have to investigate whether there is any bias on the way the systems are evaluated – either by the evaluation methodologies or metrics, since any bias in the evaluation process would lead to inconclusive or misleading results about the predictive power of the performance prediction methods proposed. If these biases do exist, we aim to precisely under-stand them and develop methodologies to isolate them; then, we shall check the effectiveness of the predictors against well-known baselines and whether it changes when unbiased methodologies are used.

RG2: Adaptation and definition of performance prediction techniques for recommender systems. We aim to study the potential of performance prediction in specific problems and settings in the area of Recommender Systems. We shall investigate the definition of a formal framework where performance predictors can be integrated. As a starting point, we aim to explore the adaptation of specific effective predictors from Information Retrieval such as query clarity (Cronen-Townsend et al., 2002) to recommender systems. Complementarily to the adaptation of known techniques, we aim to research the definition of new predictors based on models from Information Theory and Social Graphs, besides other heuristic, domain-specific approaches. Once we have defined some recommendation performance predictors, we shall assess the effectiveness of such predictors in terms of their correlation to performance metrics to estimate the predictive power of the performance predictors.

RG3: Application of performance predictors to hybrid and compound re-commender systems. We aim to identify and integrate the proposed predictors into combined recommendation methods, in order to achieve an actual improvement in the performance of the combined methods. With this goal in mind, we shall consider problems where an aggregation of recommendation methods is needed, and shall analyse how to apply the performance predictors mentioned above in such problems. Besides, a methodological study for the experimental approach, setup, and metrics should be performed in such a way that appropriate baseline methods and experimental designs are used. Finally, we shall assess the improvements and benefits of the combined methods when the performance predictors are applied.

III. METHODOLOGY

We proposed a novel user interest prediction system to generate recommendation and to support efficient web search. The method has number of stages namely Topical Detection, Web Search State Graph Generation, User Interest Prediction, Recommendation Generation. We discuss each of the functional components in detail in this section.

DOI: 10.37200/IJPR/V24I5/PR2020131
Received: 07 Mar 2020 | Revised: 26 Mar 2020 | Accepted: 04 Apr 2020

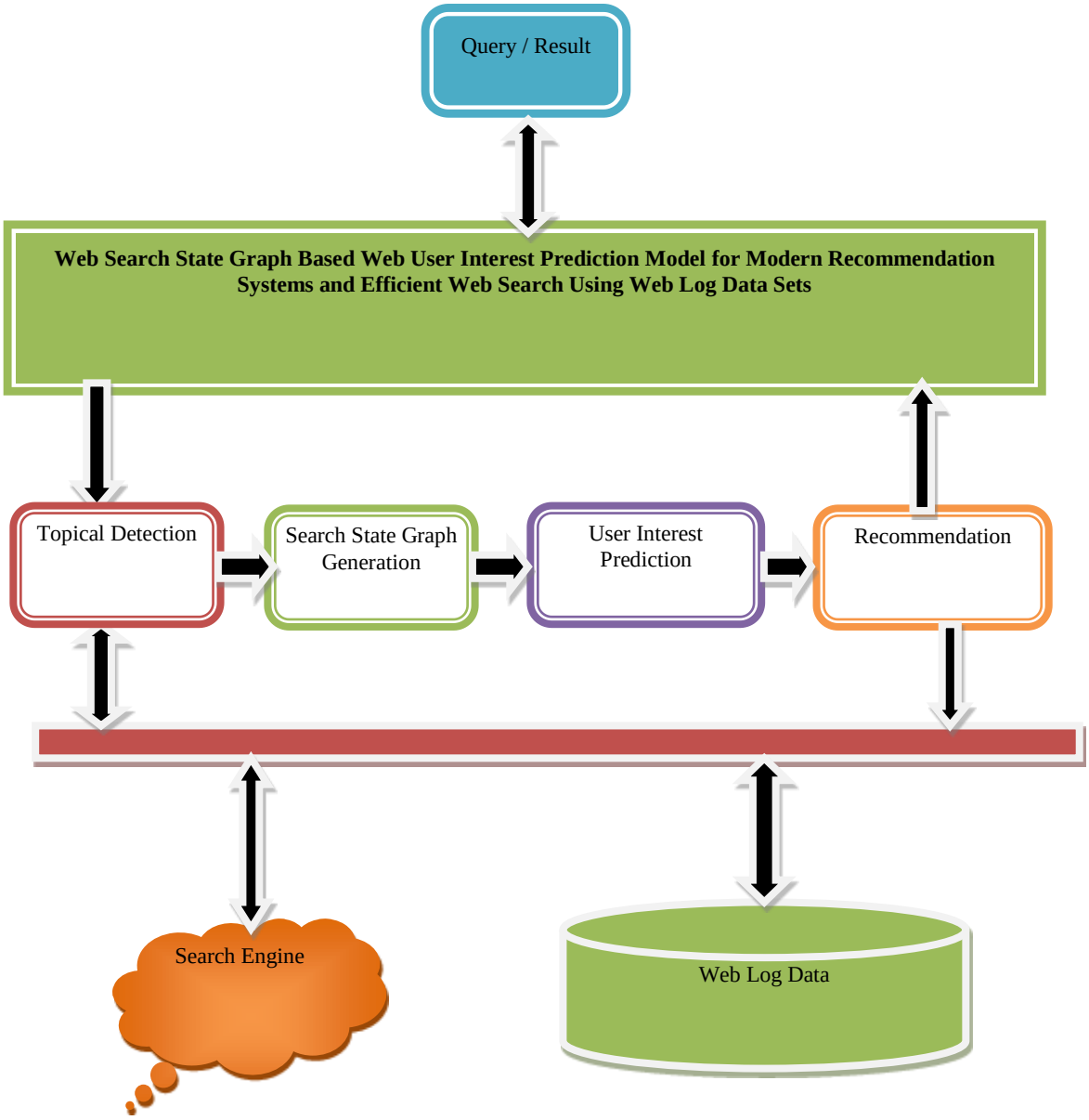


Figure 1: Proposed System Architecture

The Figure 1, shows the architecture of the proposed user interest prediction model and its functional components.

3.1 Topical Detection

At this stage, the method takes the query as the input text and removes the stop words from the text. The stop word removed content is applied with stemming process and extracted pure noun used to compute the topical similarity measure. The topical similarity measure shows how depth the term set is relevant to the topic considered. For each topic considered, the method computes the topical similarity measure and using that a single topic is identified and selected.

DOI: 10.37200/IJPR/V24I5/PR2020131
Received: 07 Mar 2020 | Revised: 26 Mar 2020 | Accepted: 04 Apr 2020

Input: Search Query/Page Content sq, Topic List Tl.

Output: Topic Tp.

Step1: start

Step2: Initialize Term set Ts.

Step2: read input text/page content

if text==query text

Ts = split terms into term set.

Ts = $\int Text \cap spaces$

else

page content pc = remove html tags.

Initialize html tag set Ht.

Pc = $\int Pc \cap \Sigma Ht$

End

Step3: for each term T_i from Ts

remove stop words.

Ts= $\sum_{i=1}^{size(Sl)} Ti \cap Sl(i)$

end

Step4: for each term T_i from Ts

perform stemming.

Ts= Stemming(T_i).

Tag the term Ti = PosTag(Ti)

if Ti==Noun then

else

Remove term from term set.

Ts = Ts$\cap Ti$

end

end

Step5: for each topic Tp_i from Topic List Tl

compute topical similarity measure Tsm.

Tsm = $\frac{\sum_{i=1}^{size(Tpi)} Tpi(k) \in Ts}{size(Tpi)}$

end

Step6: choose the topic with more Tsm.

Step7: stop.

The topical detection algorithm identifies set of terms which are pure noun and based on that the method computes the topical similarity measure. Based on topical similarity measure, the method chooses the topic with maximum similarity and identifies the topic of the query or the web page being considered.

DOI: 10.37200/IJPR/V24I5/PR2020131
Received: 07 Mar 2020 | Revised: 26 Mar 2020 | Accepted: 04 Apr 2020

3.2 Search State Graph Generation

The method initializes the state graph with the web log data set and for each user at each time window the method generates the graph. First the method splits the web log into number of time window and initializes a state graph with all the topics. From the logs splitted, the method identifies the topic of the page and generates link to the subsequent topic. Finally we will get a state transition graph with number of topics.

Algorithm:

Input: Web log Wl, Topic Taxonomy TT.

Output: Search Graph Set Sgs.

Initialize Sgs.

$$Sgs = \int \sum_{i=1}^{size(TT)} CreateGraph(root, TT(i))$$

Initialize Time window Tw.

$$Tw = \frac{TotalTimeT}{Numberofwindows}$$

for each Time window Twi from Tw

Collect web log generated at the time window Twi.

$$Wl_i = \sum_{i=1}^{size(Wl)} Wl(i)@Twi$$

end

for each log l from Wli

Topic Tp = Topical-Detection (Wli.PageContent)

Add state to the state graph Sg_i.

$$Sg_i = \sum(states \in Sgi) \cup Tp$$

generate link to the newly identified state.

end

Add to the graph set Sgs.

$$Sgs = \sum(Sgi \in Sgs) \cup Sgi$$

The above discussed algorithm generates the search state graph and using the graph being generated the next stage of the process.

3.3 User Interest Prediction

The user interest prediction is the cardinal section of the proposed approach. The method takes the search graph as the input and for each graph, the method identifies the set of all states being turned and linked. From the identified states, the method computes the state support measure for each of the state. The state support measure is computed using the web log and the time spent, actions performed on the web page and so on. Based on computed state support measure a top support state or topic is selected which represents the interest of the web user on the particular time window.

DOI: 10.37200/IJPR/V24I5/PR2020131
Received: 07 Mar 2020 | Revised: 26 Mar 2020 | Accepted: 04 Apr 2020

Algorithm
Input: Search State Graph Sgs, Web Log Wl
Output: User interest set Uis, State Support Measure SSM.
Start
Initialize states set Ss.
for each graph Sgi from Sgs
Identify set of all states.
Ss = $\sum States(ss) U \sum States(sgi)$
end
for each state Si from Ss
Compute state support measure SSM.
Compute total number of visits Tv= $\sum_{i=1}^{size(sgs)} Sgs(i) \in Si$
Compute total actions performed Ta = $\sum Actions \propto Si$
SSM = $\frac{Tv \times (Ta \times \beta)}{size(sgs)}$
end
Choose the most support state or interest Int = State(max(SSM))
Stop.

The above discussed algorithm computes the state support measure for each of the state or interest identified in different time window search state graph. Based onc computed measure a single interest with more support is identified as the interest of the user at the specific time window.

3.4 Recommendation

The method identifies the user interest at different time window and based on the identified user interest set, the method identifies the persistent interest, which is common in all the time window. Based on the support measures computed in the previous stage, we compute the cumulative search weight for each of the topic. Finally a single interest is selected and the search history of the particular state is returned as recommendations.

Algorithm:
Input: State Support Measure SSM, User interest set Uis
Output: Recommendation Rc.
Start
Identify set of all persistent interests Pint = $\sum Interest \in (\forall Tw)$
for each interest Int from Pint
compute cumulative search weight sw = $\frac{\sum_{i=1}^{size(Tw)} \sum SSM(Int)@Ti}{size(Tw)}$
end
Choose most weighted interest Mint = Int(Pint)@Max(sw)
generate recommendations.
Stop.

DOI: 10.37200/IJPR/V24I5/PR2020131
Received: 07 Mar 2020 | Revised: 26 Mar 2020 | Accepted: 04 Apr 2020

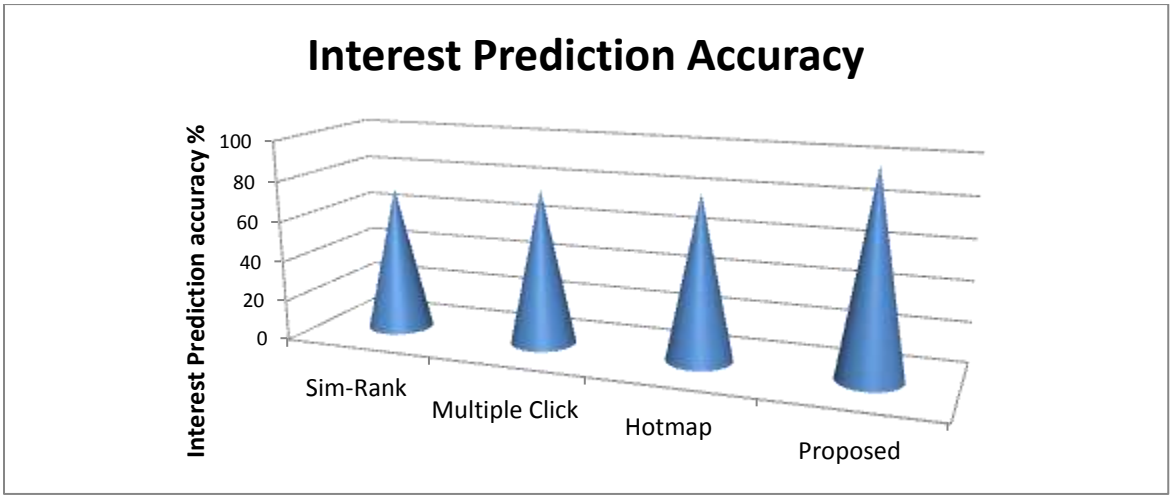
IV. Results And Discussion

The proposed search state graph based user interest prediction and recommendation system has been implemented and tested for its effectiveness. The proposed method has produced efficient results in all the factors of web mining.

Table: The Details of Implementation Parameters

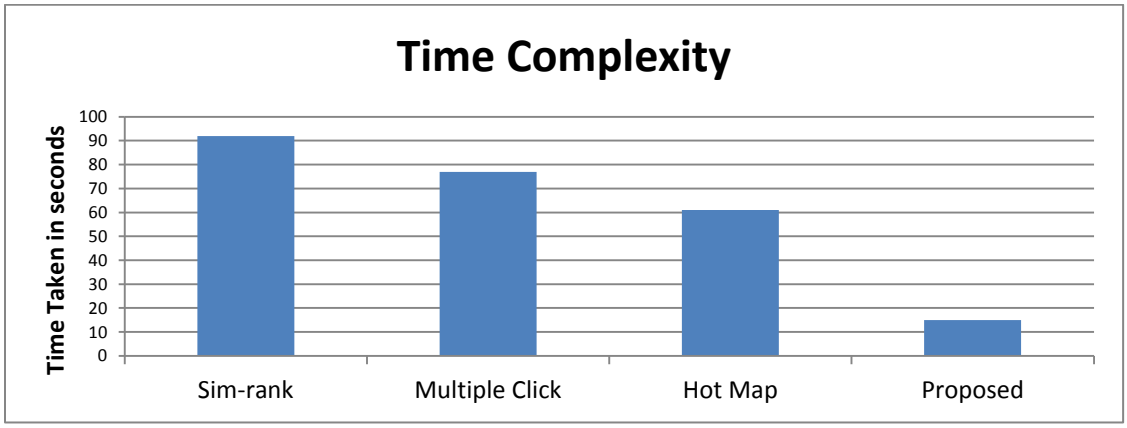
Parameter	Value
Size of web log	5 Million
Number of users	1500
Number of Interest	100
Time window considered	6 Months

The Table 1, shows the details of implementation has been used to evaluate the proposed method. The method has used 6 months log collected by monitoring the search history of 1500 users and interest into 100 numbers in overall the size of log becomes 5 million.



Graph 1: Comparison of Interest Prediction Accuracy

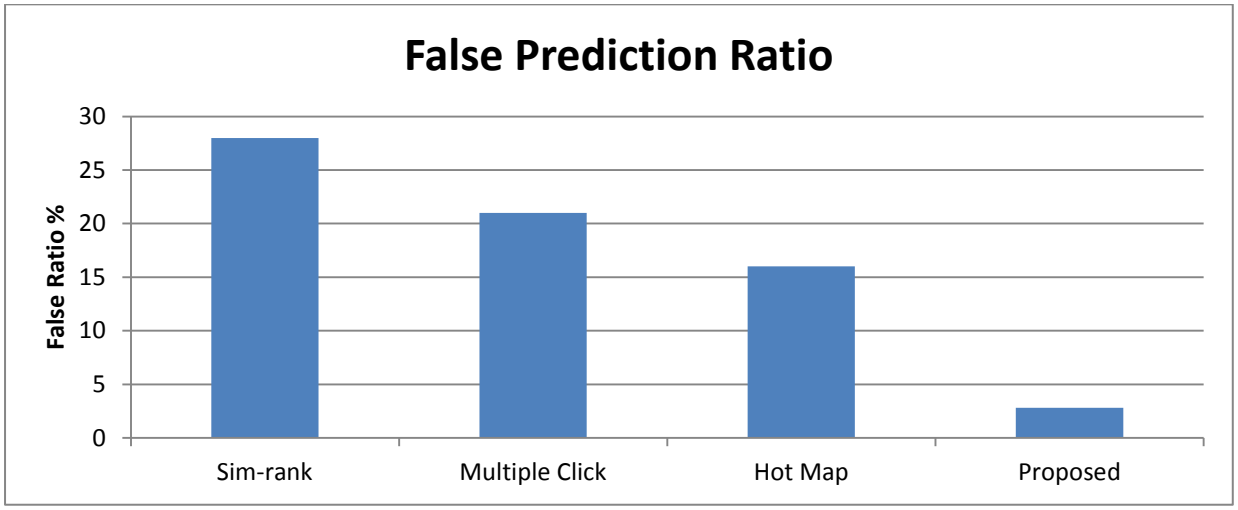
The Graph1 shows the comparison of interest prediction accuracy produced by different methods and it shows clearly that the method has produced higher accuracy in interest prediction.



Graph 2: Comparison of Time Complexity of Different Methods

DOI: 10.37200/IJPR/V24I5/PR2020131
Received: 07 Mar 2020 | Revised: 26 Mar 2020 | Accepted: 04 Apr 2020

The Graph2. Shows the comparison of time complexity produced by different methods and it shows clearly that the proposed method has produced less time complexity than others.



Graph 3: Comparison of False Prediction Ratio

The Graph 3, shows the result of comparative analysis on false prediction ratio produced by different methods. It shows clearly that the proposed method has produced less false ratio than other methods.

V. CONCLUSION

We proposed web search state graph based interest prediction to improve the performance of web search. The proposed method receives the user query and submit to the standard search engine and retrieves the result to return to the user. The pages visited and the actions performed by them and the time spent and number of clicks made and etc are traced and produced as log in the web log data set. The method identifies the topic of the web page by computing the topical similarity measure and generates the search state graph for each time window. Using identified topic and graph, the method compute the search support measure. Based on computed measure, the interest of the user at different time window is identified. Then, the method computes the cumulative search weight for each of the topic or interests using which final recommendations are produced. The method produces efficient results in web search and reduces the search time complexity and produces more efficient recommendations.

REFERENCES

[1] O. Nasraoui and R. Krishnapuram, "A New Evolutionary Approach to Web Usage and Context Sensitive Associations Mining," *Int'l J. Computational Intelligence and Applications,,* Sept. 2002.

[2] O. Nasraoui, C. Cardona, C. Rojas, and F. Gonzalez, "Mining Evolving User Profiles in Noisy Web Clickstream Data with a Scalable Immune System Clustering Algorithm," Aug. 2003

[3] Liang Deng, Martin D. F. Wong, An Exact Algorithm for the Statistical Shortest Path Problem, *ACM conference on Asia South Pacific design automation*, pages 965-970, *2006*.

[4] Orland Hoeber and Xue Dong Yang, Exploring Web Search Results Using Coordinated Views, *Fourth IEEE International Conference on Coordinated & Multiple Views in Exploratory Visualization*, pages 3-13, 2006.

[5] S.Sendhilkumar and T.V. Geetha, An Evaluation of Personalized Web Search for Individual User, *International Conference on Artificial Intelligence and Pattern Recognition (AIPR07)*, FL, USA, pages 484 -490, 2007.

DOI: 10.37200/IJPR/V24I5/PR2020131
Received: 07 Mar 2020 | Revised: 26 Mar 2020 | Accepted: 04 Apr 2020

[6] Alina Beygelzimer, John Langford, and Pradeep Ravikumar, Multiclass classification with filter trees. *Gynecologic Oncology* 105, 2, 312–320, (2007).

[7] Fan Guo X. Lou, Efficient Multiple-Click Models in Web Search, *ACM international conference on web search and data mining*, 2008

[8] Wang Xiao-gang, Web mining based on user access patterns for web personalization Computing, Communication, Control, and Management, CCCM, *ISECS International Colloquium on*, Vol 1, page:194 – 197, 2009.

[9] Ryen W. White. Predicting User Interests from Contextual Information" *Microsoft Research, ACM*-2009.

[10] Shaojie Qiao, Sim Rank: A Page Rank approach based on similarity measure, *IEEE international conference on Intelligent Systems and Knowledge Engineering (ISKE),* P**age(s):**390 – 395, 2010.

[11] Huajing Li, Personalized Feed Recommendation Service for Social Networks, *Social Com,* PP:96-103, 2010.

[12] Chunyang Liang User profile for personalized web search, *International conference on fuzzy systems and knowledge discovery,* Vol:3, pp:1847-1850, 2011.

[13] Pablo Castells, SaÃžl Vargas, and Jun Wang. Novelty and Diversity Metrics for Recommender Systems: Choice, Discovery and Relevance. *In Proceedings of International Workshop on Diversity in Document Retrieval (DDR)*, 29–37, 2011.

[14] Zeno Gantner, Steffen Rendle, Christoph Freudenthaler, and Lars Schmidt Thieme. My Media Lite: A free recommender system library. *In Proceedings of the fifth ACM conference on Recommender systems. ACM*, 305–308, 2011.

[15] Amr Ahmed, Yucheng Low " Scalable Distributed Inference of Dynamic User Interests for Behavioral Targeting ", *ACM*-(2011)

[16] Akther.A, Social network and user context assisted personalization for recommender systems, *IEEE, Innovations in Information Technology*, pp:95-100, 2012.

[17] Heng-Tze Cheng, Levent Koc, Jeremiah Harmsen, Tal Shaked, Tushar Chandra, Hrishi Aradhye, Glen Anderson, Greg Corrado, Wei Chai, Mustafa Ispir, et al. Wide & deep learning for recommender systems. *In Proceedings of the 1st Workshop on Deep Learning for Recommender Systems. ACM*, 7–10, (2016).

[18] Paul Covington, Jay Adams, and Emre Sargin, Deep Neural Networks for YouTube Recommendations. *In ACM Conference on Recommender Systems.* 191–198, 2016.

[19] Robin Devooght and Hugues Bersini., Collaborative Filtering wih Recurrent Neural Networks. (2016).

[20] Xiangnan He, Hanwang Zhang, Min-Yen Kan, and Tat-Seng Chua. Fast matrix factorization for online recommendation with implicit feedback. *In Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval. ACM*, 549–558, 2016.

[21] T.P.Anithaashri, R. Baskaran, Enhancing Multi-user Network using sagacity dismissal of conquered movements, *International Journal of American Scientific Publishers* pp:69-78, 2016

[22] Xiangnan He, Lizi Liao, Hanwang Zhang, Liqiang Nie, Xia Hu, and Tat-Seng Chua. Neural collaborative filtering. In Proceedings of the 26th International Conference on World Wide Web. *International World Wide Web Conferences Steering Committee*, 173–182, 2017.

[23] Shuai Zhang, Lina Yao, and Aixin Sun. 2017. Deep Learning based Recommender System: A Survey and New Perspectives. (2017).

[24] Guorui Zhou, Chengru Song, Xiaoqiang Zhu, Xiao Ma, Yanghui Yan, Xingya Dai, Han Zhu, Junqi Jin, Han Li, and Kun Gai. 2017. *Deep Interest Network for Click-Through Rate Prediction.* (2017).

[25] T.P. Anithaashri, G. Ravichandran, et.al. Secure Data Access Through Electronic Devices Using Artificial Intelligence, *ICCES*, 2018.

[26] T.P.Anithaashri, G. Ravichandran, R.Baskaran, Software Defined Network Security enhancement using Game Theory, *Elsivere COMNET,* vol157, pp:112-121, 2019

[27] C. Sindhu and G.Vadivu. Effects of Adjective Verb Adverb on Sentiment Classification Using Support Vector Machine for Green Communication. *Journal of green engineering*, volume.10, issue.1, pp.91–102, 2020

[28] M. Gayathri, C. Malathy, Siddharth Singh.MARK42: The Secured Personal Assistant using Biometric Traits Integrated with Green IOT, *Journal of green engineering*, volume.10, issue.1, pp. 255–267, 2020

DOI: 10.37200/IJPR/V24I5/PR2020131
Received: 07 Mar 2020 | Revised: 26 Mar 2020 | Accepted: 04 Apr 2020