American-Eurasian Journal of Scientific Research 11 (2): 112-118, 2016
ISSN 1818-6785

DOI: 10.5829/idosi.aejsr.2016.11.2.22834

Web Search State Graph Based Web User Interest Prediction Model for Modern Recommendation Systems and Efficient Web Search Using Web Log Data Sets

[1]P. Sasikumar and [2]M. Karthikeyan

[1]Department of Computer Science and Engineering, Selvam College of Technology
[2]Principal, Tamil Nadu College of Engineering

Abstract: The problem of web search has been discussed in variety of situations and there are many approaches recommended by different researchers earlier. The earlier approaches suffer with the problem of predicting the user's future interest which is highly required for recommend systems. We propose a novel web search state graph clustering approach to identify the user interest in web search and the same is applied to predict the future interest of the user. The proposed method splits the web log data into number of time window and for each time window an web search state graph has been generated. Using generated search state graph, the method identifies list of common states and computes state support measure for each of the interest identified. 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.

Key words: Web Search · Web Mining · Web Search State Graph · User Interest Prediction

INTRODUCTION

The problem of web search has been analyzed of the past 30 years by number of researchers, where the ultimate aim is to produce efficient search result to the web user. The user submits search query to the search engine and the search engine returns set of result to the user. The search engine does not consider about what the user is looking for and generates results accordingly. 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.

Corresponding Author: P. Sasikumar, Department of Computer Science and Engineering, Selvam College of Technology, India

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.

Related Works: There are many approaches has been discussed to identify the user interest in web search and interest prediction. We discuss about the methods being discussed earlier in this section.

An Evaluation of Personalized Web Search for Individual User [1] focused on the evaluation of the results of individual user's User Conceptual Index based search and introduces three measures for the purpose. Context includes factors like the nature of information available, the information currently being examined, when and what applications in use and so on. The Individual oriented search encompasses elements like the user's goals, prior and tacit knowledge, past information seeking behaviors, among others.

Exploring Web Search Results Using Coordinated Views [2], In HotMap, the frequencies of each of the query terms from the user's queries are depicted visually using color-coding. This allows the users to easily identify "hot" documents based on the frequent appearance of the query terms within the document surrogates. In addition to this visual representation, the search results can be dynamically resorted based on the query term frequencies, supporting an interactive exploration of the search results.

Application of user access pattern for web personalization has been discussed in [3]. The method uses sequential patterns to perform web personalization. The method incorporates the sequential pattern mining algorithm which identifies the frequent sequential Web access patterns. The generated sequential patterns are formulated in a tree structure and the same will be used to perform matching and to generate proposed results or recommendation.

Efficient Multiple-Click Models in Web Search [4], presents a click model which logs the clicks and then contain the submitted query, a ranked list of returned documents, whether each of them is clicked or not and other information that might be useful. Click models learn from user clicks to help understand and incorporate users' implicit feedback. And they follow a probabilistic approach which treats user clicks as random events and the goal is to design generative models which are able to approximate underlying probabilities of clicks with high accuracy.

SimRank: A Page Rank approach based on similarity measure [5], propose a new page rank algorithm based on similarity measure from the vector space model, called SimRank, to score web pages. Firstly, a new similarity measure used to compute the similarity of pages and apply it to partition a web database into several web social networks (WSNs). Secondly, they improve the traditional Page Rank algorithm by taking into account the relevance of page to a given query. Thirdly, we design an efficient web crawler to download the web data. And finally, experimental studies are performed to evaluate the time efficiency and scoring accuracy of SimRank with other approaches.

All the above discussed approach has the problem of false identification of interests and has low accuracy in identifying the user interest.

MATERIALS AND METHODS

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.

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

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.

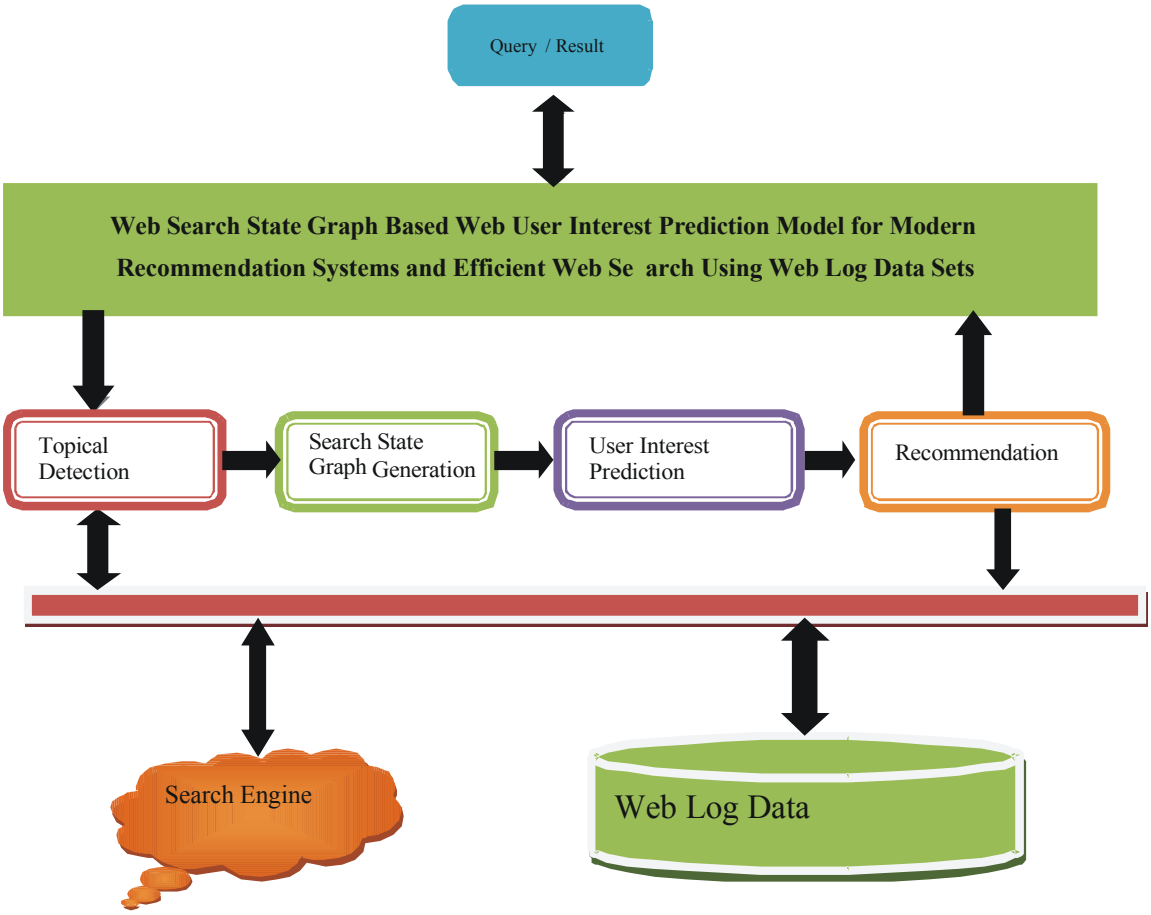


Fig. 1: Proposed System Architecture

Input: Search Query/Page Content sq, Topic List Tl.
Output: Topic Tp.
Step 1: start
Step 2: Initialize Term set Ts.
Step 2: 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
Step 3: for each term T_i from Ts
remove stop words.

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

end

Step 4: 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
Step 5: 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
Step 6: choose the topic with more Tsm.
Step 7: 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.

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.

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

Initialize Time window Tw.

$$\text{Tw} = \frac{Total\ Time\ T}{Number\ of\ windows}$$

for each Time window Twi from Tw
Collect web log generated at the time window Twi.

$$\text{Wl}_i = \sum_{i=1}^{size(wl)} Wl(i)@Twi$$

end
for each log l from Wli
Topic Tp = Topical-Detection (Wli.PageContent)
Add state to the state graph Sg_i.

$$\text{Sg}_i = \Sigma(states \in Sgi) \cup Tp$$

generate link to the newly identified state.
end
Add to the graph set Sgs.

$$\text{Sgs} = \Sigma(Sgi \in Sgs) \cup Sgi$$

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

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.

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.

$$\text{Ss} = \Sigma states\ (ss)\ U\ \Sigma\ States(sgi)$$

end
for each state Si from Ss
Compute state support measure SSM.
Compute total number of visits

$$\text{Tv} = \sum_{i=1}^{size(sgs)} Sgs(i) \in Si$$

Compute total actions performed $\text{Ta} = \Sigma\ Actions \propto Si$

$$\text{SSM} = \frac{Tv \times (Ta \times \beta)}{size(sgs)}$$

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.

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 = Σ *Interest* $\in$ ($\forall$ *Tw*)

for each interest Int from Pint

compute cumulative search weight sw = $\frac{\sum_{i=1}^{size(Tw)} \Sigma SSM(Int)@Ti}{size(Tw)}$

end

Choose most weighted interest Mint = Int(Pint)@Max(sw)

generate recommendations.

Stop.

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.

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.

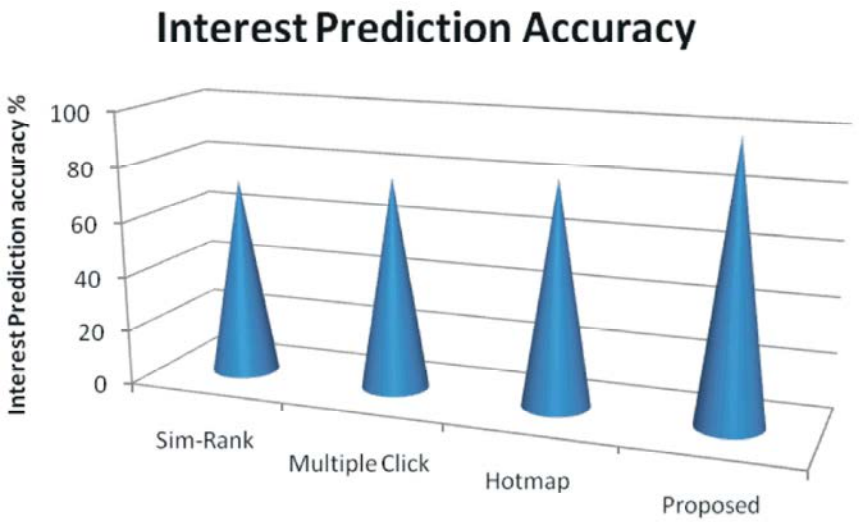
The Graph 1 shows the comparison of interest prediction accuracy produced by different methods and it shows clearly that the method has produced higher accuracy in interest prediction.

The Graph 2 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.

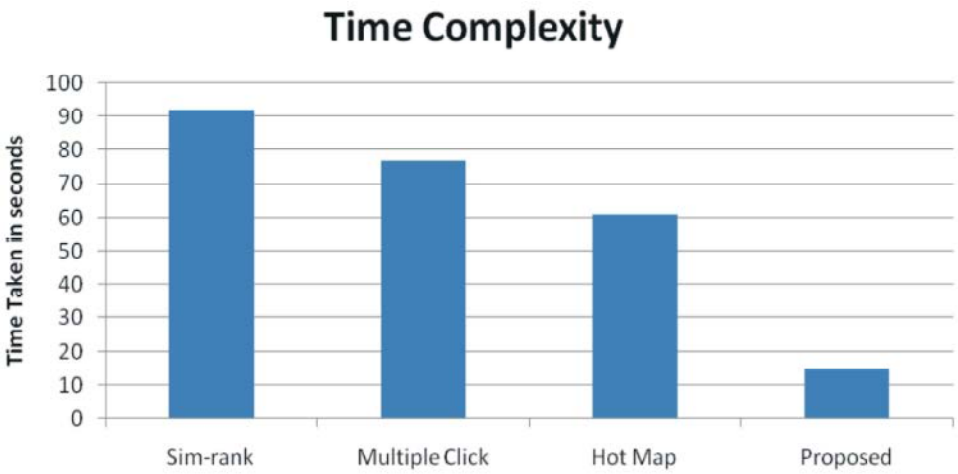
The Graph 3 shows the result of comparative analysis on false prediction ratio produced by different methods. It shows clearly that the proposed method has produced less false ratio than other methods.

Table 1: The details of implementation parameters

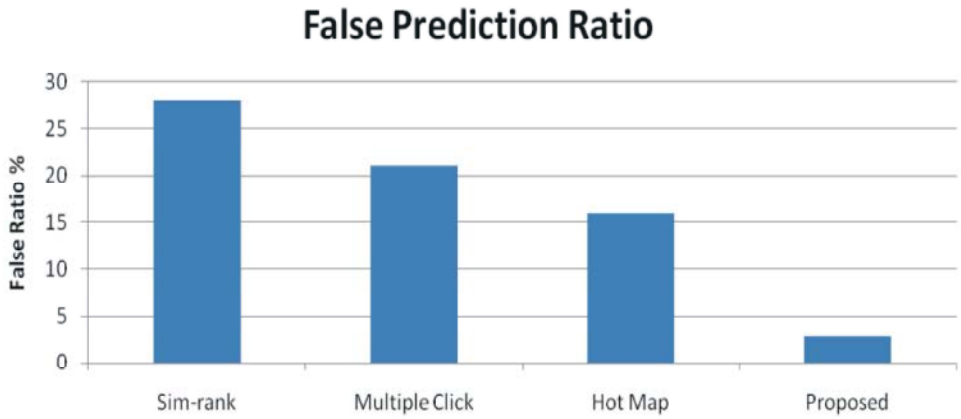
Parameter	Value
Size of web log	5 Million
Number of users	1500
Number of Interest	100
Time window considered	6 Months



Graph 1: Comparison of interest prediction accuracy



Graph 2: Comparison of time complexity of different methods



Graph 3: Comparison of false prediction ratio

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. Wang Xiao-gang, 2009. Web mining based on user access patterns for web personalization Computing, Communication, Control and Management, CCCM, ISECS International Colloquium On, 1: 194-197.
2. Fan Guo X. Lou, 2008. Efficient Multiple-Click Models in Web Search, ACM International Conference on Web Search and Data Mining.
3. Shaojie Qiao, 2010. SimRank: A Page Rank approach based on similarity measure, IEEE international conference on Intelligent Systems and Knowledge Engineering (ISKE), Page(s): 390-395.

4. Liang, Deng and Martin D.F. Wong, 2006. An Exact Algorithm for the Statistical Shortest Path Problem, ACM conference on Asia South Pacific Design Automation, pp: 965-970.
5. Sendhilkumar, S. and T.V. Geetha, 2007. An Evaluation of Personalized Web Search for Individual User, International Conference on Artificial Intelligence and Pattern Recognition (AIPR07), FL, USA, pp: 484-490.
6. Orland, Hoeber and Xue Dong Yang, 2006. Exploring Web Search Results Using Coordinated Views, Fourth IEEE International Conference on Coordinated & Multiple Views in Exploratory Visualization, pp: 3-13.
7. Chunyang Liang, 2011. User profile for personalized web search, International conference on fuzzy systems and Knowledge Discovery, 3: 1847-1850.
8. Amr Ahmed, Yucheng Low, 2011. Scalable Distributed Inference of Dynamic User Interests for Behavioral Targeting, ACM-(2011).
9. Ryen W. White, 2009. Predicting User Interests from Contextual Information" Microsoft Research, ACM.
10. Huajing Li, 2010. Personalized Feed Recommendation Service for Social Networks, Social Com., pp: 96-103.
11. Nasraoui, O. and R. Krishnapuram, 2002. A New Evolutionary Approach to Web Usage and Context Sensitive Associations Mining, Int'l J. Computational Intelligence and Applications.
12. Nasraoui, O., C. Cardona, C. Rojas and F. Gonzalez, 2003. Mining Evolving User Profiles in Noisy Web Clickstream Data with a Scalable Immune System Clustering Algorithm.
13. Akther, A., 2012. Social network and user context assisted personalization for recommender systems, IEEE, Innovations in Information Technology, pp: 95-100.
14. Nasraoui, O., C. Rojas and C. Cardona, 2006. A Framework for Mining Evolving Trends in Web Data Streams Using Dynamic Learning and Retrospective Validation.
15. Maloof, M.A. and R.S. Michalski, 1995. Learning Evolving Concepts Using Partial Memory Approach.
16. Maloof, M.A. and R.S. Michalski, 2000. Selecting Examples for Partial Memory Learning, Machine Learning, 41(11): 27-52.
17. Agarwal, D. and S. Merugu, 2007. Predictive discrete latent factor models for large scale dyadic data. KDD.