Available Online at www.ijcsmc.com

International Journal of Computer Science and Mobile Computing

A Monthly Journal of Computer Science and Information Technology

ISSN 2320–088X

IMPACT FACTOR: 6.017

IJCSMC, Vol. 6, Issue. 1, January 2017, pg.36 – 42

Addressing Low Precision in Web Log Mining for Personalized Information Retrieval

Fatima Jameel Kadhim

Department of Computer Science / Thiqar University, Thiqar, Iraq

Kjkadhim@yahoo.com

***Abstract:** The methodological approach to achieve the elicited research objectives is presented in this study. The study will employ an experimental methodology involving testing of algorithms to identify the suitable algorithm for an optimized process of information retrieval. The algorithm that presents a better precision in the personalized IR in web log mining represents the core value of the architecture that this study presents.*

***Keywords:** Algorithms, Information retrieval, Wearable computing, Query, Augmented reality*

I. INTRODUCTION

The massive explosion of information and its growing trend have been suggesting continuous researches in areas of information search and information retrieval (IR). There have been increases in the needs to provide users with working platforms to deal with the over-abundant information on the internet-especially to achieve appropriate information retrieval with less effort (Hong, Park, Lee, Shin, & Woo, 2005);(Insley, 2003). IR is concerned with finding appropriate documents from the massive database and libraries. And towards its optimal results, the inclusion of wearable computing, mobile augmented reality (AR), ubiquitous computing environment and personalized information retrieval have been significantly advocated for (Frakes & Baeza-Yates, 1992) .

Personalized information retrieval is a crucial approach to attend to the inevitable experience of information overabundance in the present information age. Its main goal is providing only the relevant information to users when they need it using an appropriate approach (Schneider at al., 2010). Personalizing information retrieval through web search can be done through Content-based Filtering, Demographic Personalization, and Collaborative filtering, Utility-based Information Retrieval and Knowledge-Based Recommendation (Mylonas, 2008). The popular search technologies like Yahoo and Google are applying personalized web search and browsing in their search engines (Schneider at al., 2010).

However, the information associated with query could not be determined due to query terms uncertainty and query short, queries keywords. As a result, many documents which are irrelevant with the input query are being retrieved and precision of the retrieval process is dishonored (Baeza-Yates, Hurtado, & Mendoza, 2005). This experience has posed great challenge to research works in web search and IR works. (Mylonas, 2008) addressed this query-associated problem in IR using context and ontological knowledge approach, (Choi, 2011) introduced Near Field Communication (NFC) for smart phone IR and (Suomalainen, Hyttinen, & Tarvainen, 2010) while addressing issues in personalised IR pointed to the need to move from system-centred approach to user-centred approach in personalised IR.

Notably, many studies that have been carried out for improving the process of information retrieval recommend similar queries sets as the input query response with ranks of some suggested queries done in accordance with the relevant prerequisites .It is however posited that there can be further recommendation of queries when the needed information of the previous sessions of the past queries are issued on the search engine(Baeza-Yates, et al., 2005).

However, from the understanding of the information foraging theory, the guidance of users in the information community is guided by the information scent approach. Users have the tendency of clicking the search results retrieved pages that conform to their needed information. Notably, these pages contain Information scent in association with it in accordance with their needed information. The extent of the satisfaction of the information needed by the user, the more to be the associated information scent (Chi, Pirolli, Chen, & Pitkow, 2001).

Web log mining is for different applications: ranging from web users' search for the web site's organization. Web users' search process are confronted with retrieval related problem majorly due to the employed IR approach since review from extant literatures has posited that the search result is highly dependent on the approach employed for the IR (Mylonas, 2008); (Suomalainen, et al., 2010). However, achieving precision in the result delivered during the search process in the IR has been the center of research concern. It is thus worthwhile to further work on precision in IR especially in the web log mining process.

II. RESEARCH METHODOLOGICAL FRAMEWORK

In achieving the elicited research objectives of this study, figure 3.1 depicts the iterative procedural steps to be taken. This marks the research methodological framework.

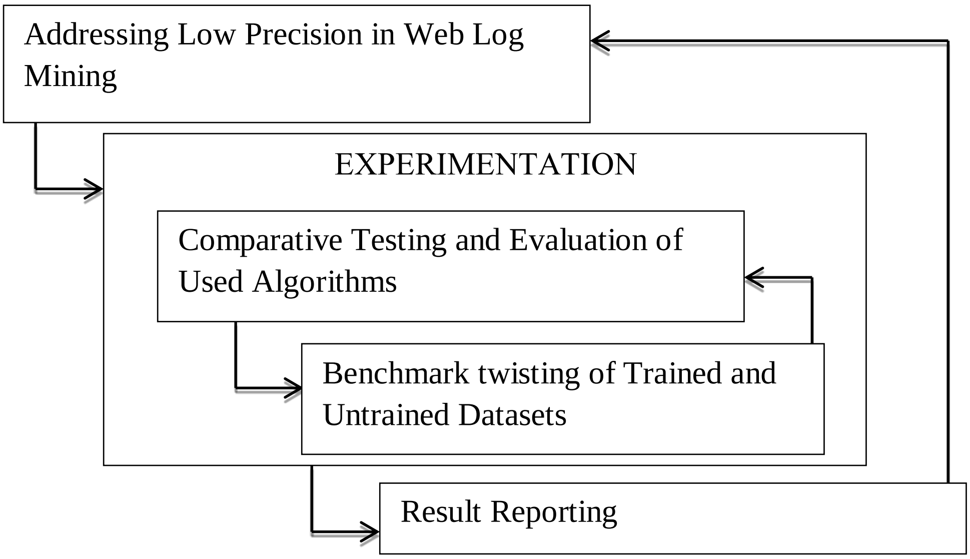
Addressing Low Precision in Web Log Mining
EXPERIMENTATION
Comparative Testing and Evaluation of Used Algorithms
Benchmark twisting of Trained and Untrained Datasets
Result Reporting

Figure 3.1: Research Methodological Framework.

III. Experiment Setting

The experiment will be carried out on the data set that contains the clicked documents connected with queries supplied to the Google search engine. The set of data would be collected from the Google search engine web history. The users will generate the data set in view of their web search interest. Notably, the web history contains the following fields.

1. Time of the Day
2. Query terms
3. Clicked URLs

Experiments will be carried out on the data set got from the history of the web and will be loaded into a database for further processing.

The result page of Google search engine returns consists of URLs with information about URLs On the submission of input query. The considered Query sessions entails query terms along with clicked URLs. The clicked URLs are those URLs which are clicked by the users before submitting another query.

Similarity of any two query sessions will be calculated using the cosine measure in this research work. The query sessions will be clustered using k-means algorithm and this will be carried out several times to get different values of k and for each value of k, criterion function will be computed. The maximum and threshold values will both be recorded in order to evaluate the criterion function.

The experiment will be carried out on test queries that are randomly selected which would have been classified into trained and queries set that is untrained. Trained queries are those input queries that have sessions related to them in the data set and untrained queries are those input queries that do not have sessions in data set that is related with them. Some of the test queries in each of the categories are given in Table 3.1

Table 3.1: Sample of Queries taken in each of the categories

Category	**Queries**
Untrained Set	Movies, Space food, novels ,magazine, Movies ,Numbness, Nature, family play Games, movie pictures, software download, online tutorial
Trained Set	Homeloan , distance education online, free pics, cgi perl tutorial, moons of neptune, how to play .vcd files, .vcd file, .api com, mpeg movies, drag onball ,intranet , help desk manager job description, free software

The Platforms that will be used in carrying out the experimentation are: Oracle database employing Java. The web Sphinx crawler will be employed in fetching the query session clicked documents in the data set. Each query session will then be changed into the vector representation employing Information Scent and the content of clicked the URLs. The algorithm of k-means will be effected for generating query session's clusters and representing each query session with a mean value of vector of terms.

The contributions of the queries that are recommended within the cluster are selected for the input query. It is decided when some unknown users that have knowledge in domain that input query should be. The importance will be judged through the analysis of the recommended queries that are answered from the set of result indicating top 10 correct responses. This is determined when the URL answers are similar to the input query.

The setup will then be carried out on 21 trained queries that are randomly selected and selected untrained queries. Therefore, the mean precision of the trained and untrained query set will be calculated for different number of recommended queries from the result set showing the top 10 answers.

IV. Experimentation Algorithm

The algorithm to be used for the experimentation process is provided below:

Algorithm

```
1. Offline Preprocessing phase at regular and periodical intervals
    1.1. Extract the queries and associated clicked URLs from the data set.
    1.2. Preprocess the Extracted Queries to find the query sessions.
```

1.3. Model the Information need associated with each query session using information scent and weighted vector of the content of pages in the session using (1)(2)(3).

1.4. Cluster the Query sessions using information need associated with each query session using k-means.

1.5. For each cluster Cj create a list of queries Qj in cluster Cj.

2. Online searches:

2.1. Find the Cj cluster to which input query q belongs.

2.2. If no cluster found them

2.2.1. Find the Cj cluster, which is most similar to the term weight vector of input query q as per the threshold value set for similarity measure.

2.3. Rank the list of queries Qj associated with selected cluster Cj in order of their relevance to input query q up to certain similarity threshold value.

2.4. Return the ranked set of queries.

The queries rank in set Qj is calculated using a similarity measure of each query vector x in Qj to input query vector q such that those queries with high value of similarity to input query q are ranked higher than those queries with low value of similarity to input query q where sin (x, q) is calculated using the cosine measure between vector x and q.

V. CONCLUSION

This research has made efforts in satisfying the users Information need and improving the precision of information retrieval through the recommendation of related queries which approximate the information need in association with the input query employing Information Scent. The information need associated with the query is modeled using information scent and content feature of clicked pages in the session. The suggested queries aid the retrieval of the documents relevant to the users' information need which he was unable to get through his initial query.

SUMMARY

This study presented the methodological steps to be taken in achieving the elicited research objectives and answer the research questions. Experimental methods used the Oracle database platform and randomly selecting test queries which would have been categorized into trained queries set and untrained queries set. The trained and untrained dataset were presented, and the search query algorithm was also illustrated in this article.

ACKNOWLEDGEMENT

I would like to acknowledge all staff in the Computer laboratory for their direct or indirect effort and contribution to the realization of this paper.

REFERENCES

[1] Baeza-Yates, R., Hurtado, C., & Mendoza, M. (2005). *Query recommendation using query logs in search engines.* Paper presented at the Current Trends in Database Technology-EDBT 2004 Workshops.

[2] Bedi, P., & Chawla, S. (2007). Improving information retrieval precision using query log mining and information scent. *Information Technology Journal, 6*(4), 584-588.

[3] Bhushan, R., & Nath, R. (2013). Web Crawler–A Review. *International Journal of Advanced Research in Computer Science and Software Engineering Volume-3, Issue-8, August-2013*, 54.

[4] Bhushan, R., & Vats, M. (2013). Information Retrieval Using Web Mining Techniques. *International Journal of Advanced Research in Computer Science and Software Engineering Volume-3, Issue-8, August-2013*, 169.

[5] Chang-bin, J. (2010). *Application of cloud model in personalized service recommendation of web log mining.* Paper presented at the Biomedical Engineering and Computer Science (ICBECS), 2010 International Conference on.

[6] Chawla, S. (2012). Semantic Query Expansion using Cluster Based Domain Ontologies. *International Journal of Information Retrieval Research (IJIRR), 2*(2), 13-28.

[7] Chawla, S., & Bedi, P. (2008). *Improving information retrieval precision by finding related queries with similar information need using information scent.* Paper presented at the Emerging Trends in Engineering and Technology, 2008. ICETET'08. First International Conference on.

[8] Chen, Z. (2007). *Web Log Mining Based On Fuzzy Immunity Clonal Selection Neural Network.* Paper presented at the Service Systems and Service Management, 2007 International Conference on.

[9] Chi, E. H., Pirolli, P., Chen, K., & Pitkow, J. (2001). *Using information scent to model user information needs and actions and the Web.* Paper presented at the Proceedings of the SIGCHI conference on Human factors in computing systems.

[10] Choi, J., Kim, M. and Raghavan,V. V. . (2011). Adaptive feedback methods in an extended boolean model. *In Proceedings of ACM SIGIR Workshop on Mathematical/Formal Methods in Information Retrieval*, *New Orleans*,LA,Sept.2001

[11] Cooley, R., Mobasher, B., & Srivastava, J. (1997). *Web mining: Information and pattern discovery on the world wide web.* Paper presented at the Tools with Artificial Intelligence, 1997. Proceedings., Ninth IEEE International Conference on.

[12] Duan, J., & Liu, S. (2012). *Research on web log mining analysis.* Paper presented at the Instrumentation & Measurement, Sensor Network and Automation (IMSNA), 2012 International Symposium on.

[13] Etzioni, O. (1996). The World-Wide Web: quagmire or gold mine? *Communications of the ACM, 39*(11), 65-68.

[14] Frakes, W., & Baeza-Yates, R. (1992). Information retrieval: Algorithms and data structures. *Chile: University Of Chile*.

[15] Gao, W.-H. (2010). *Research on client behavior pattern recognition system based on web log mining.* Paper presented at the Machine Learning and Cybernetics (ICMLC), 2010 International Conference on.

[16] Grace, L., Maheswari, V., & Nagamalai, D. (2011). Analysis of web logs and web user in web mining. *arXiv preprint arXiv:1101.5668*.

[17] Gudivada, V. N., Raghavan, V. V., Grosky, W. I., & Kasanagottu, R. (1997). Information retrieval on the world wide web. *Internet Computing, IEEE, 1*(5), 58-68.

[18] Hong, D., Park, Y.-K., Lee, J., Shin, V., & Woo, W. (2005). *Personalized Information Retrieval Framework.* Paper presented at the ubiComp workshop (ubiPCMM).

[19] Insley, S. (2003). Obstacles to general purpose augmented reality. *ECE 399H, Information Security & Cryptography, Oregon, EUA*.

[20] Jansen, B. J., Spink, A., Bateman, J., & Saracevic, T. (1998). *Real life information retrieval: a study of user queries on the Web.* Paper presented at the ACM SIGIR Forum.

[21] Lalithadevi, B., Ida, A. M., & Breen, W. A. (2013). A New Approach for Improving World Wide Web Techniques in Data Mining. *International Journal, 3*(1).

[22] Lu, Z., Yao, Y., & Zhong, N. (2003). *Web log mining.* Paper presented at the Web Intelligence.

[23] Mukherjee, I., Bhattacharya, V., Banerjee, S., Gupta, P., & Mahanti, P. (2012). *Efficient web information retrieval based on usage mining.* Paper presented at the Recent Advances in Information Technology (RAIT), 2012 1st International Conference on.

[24] Mylonas, P. V., D.; Castells, P.; Fernandez, M. and Avrithis, Y.. (2008). Personalized information retrieval based on context and ontological knowledge.. *The Knowledge Engineering Review, 23*(01), 73-100.

[25] Pirolli, P. (1997). *Computational models of information scent-following in a very large browsable text collection.* Paper presented at the Proceedings of the ACM SIGCHI Conference on Human factors in computing systems.

[26] Pirolli, P. (2004). The Use of Proximal Information Scent to Forage for Distal Content on the World Wide Web, Working with Technology in Mind: Brunswikian Resources for Cognitive Science and Engineering: Oxford Press.

[27] Saad, S. Z., & Kruschwitz, U. (2011). Applying web usage mining for adaptive intranet navigation *Multidisciplinary Information Retrieval* (pp. 118-133): Springer.

[28] Sharma, S., & Varshney, M. (2010). *An efficient approach for Web-log mining using ART.* Paper presented at the Education and Management Technology (ICEMT), 2010 International Conference on.

[29] Spiliopoulou, M. (1999). The laborious way from data mining to web log mining. *Computer Systems Science and Engineering, 14*(2), 113-126.

[30] Suomalainen, J., Hyttinen, P., & Tarvainen, P. (2010). *Secure information sharing between heterogeneous embedded devices.* Paper presented at the Proceedings of the Fourth European Conference on Software Architecture: Companion Volume.

[31] Tao, X., Li, Y., & Zhong, N. (2011). A personalized ontology model for web information gathering. *Knowledge and Data Engineering, IEEE Transactions on, 23*(4), 496-511.

[32] Xing, W., & Ghorbani, A. (2004). *Weighted pagerank algorithm.* Paper presented at the Communication Networks and Services Research, 2004. Proceedings. Second Annual Conference on.