JES
Journal of Engineering Sciences

Vol 10, Issue 10, OCT/2019
ISSN NO:0377-9254

RELEVANCE FEEDBACK BASED INTELLIGENT RETRIEVAL FOR CAPTURING USER CONTEXT IN PERSONALIZED WEBSITES

Salma Shafiuddin[1]

[1]PG Scholar, MTech, Dept of Computer Science Engineering, Osmania University, Hyderabad, India.

Abstract—*With the rapid growth of content on the Internet, it has become a si gnificant challenge to obtain useful data to satisfy accurate search demands, particularly in personalized websites. General search engines face difficulties in addressing the challenges brought by this exploding amount of information. A novel web usage mining approach is proposed for supporting effective periodic* web search personalization *with real-time location and relevant feedback technology based on click-through data analysis to design and implement an efficient, configurable, and intelligent retrieval framework for personalized websites. To improve the retrieval results, this project also proposes a strategy of personal web re-visitation by context which is based on the user's current access context to obtain the relationship between the user query conditions and retrieval results. Finally, this project designs a personalized PageRank algorithm including modified parameters to improve the ranking quality of the retrieval results using the relevant feedback and access context from different users in different interest groups. The proposed intelligent retrieval framework improves the user experience.*

1. INTRODUCTION

The goal of Intelligent Retrieval based on Relevance Feedback is is to collect valid information that isperso nalized to user interests and is captured to analyze the behavior of the user. There has been extensive use of various retrieval approaches in the past to retrieve massive amounts of information on the internet. For instance, people can u se search engines to crawl datafrom the Web easily, such as via Google and Bing.

The page rank concept of general search engines basically focuses on linkage relations between webpages, keywords, dwell-time and concept-based search results that provide value information and related services for a particular field, a particular person and a particular demand.

However, the retrieval results often contain substantial amounts of unnecessary information, and some required results can be hidden in the back of a webpage; thus, users have to spend a lot of time finding the relevant results. Different users have different search needs on account of their different ages, interests and occupations.

To overcome these problems, our main contributions are as follows -To propose an accurate and intelligent retrieval frame-work with real-time location and relevant feedback technology for personalized websites.

To predict user retrieval intentions by analyzing the user's real-time location to determine a personalized search range. To improve the retrieval results, a strategy of implicit relevant feedback based on click-through data analysis is proposed, which can obtain the relationship between the user query conditions and the retrieval results.

The concept of 'Search Context' which captures the user's current 'Access Context' has been introduced. This can determine user's current browsing context

from the user's concurrent activities to give better enhanced search results.

To design a personalized PageRank algorithm including modified parameters to improve the ranking quality of the search results using the relevant feedback from users in different interest groups. The solution ensures that different users obtain different search results that are closer to the user's access requirements, even with the same keywords search.

2. RELATED WORK

Researchers have made great effort to improve the efficiency of information retrieval. In the literature, A personalized re-ranking algorithm through mining user dwell times is proposed by deriving from a user's previously online reading or browsing activities. The system acquires document level user dwell times via a customized web browser, from which it infers concept word level user dwell times in order to understand a user's personal interest [2].

In addition, with the increase in the demands on user satisfaction, vertical search engines have provided certain value information and related services for a particular field, a particular person and a particular demand (e.g., travel searches and educational resource searches) [3]. However, detailed and accurate information is still not able to be obtained by vertical or general search engines. For instance, if we are at a particular University, we need to search today's news, find the location of the science library, etc. A general search engine will not provide satisfactory results.

Reference [5] proposed a topic-sensitive PageRank algorithm to avoid the theme drift problem of the algorithm. Reference [6] Personalizes information retrieval for multi-session tasks by examining the roles of task stage, task type, and topic knowledge on the interpretation of dwell time as an indicator of document usefulness. Reference [7] proposed a new web search personalization approach that captured the user's interests and preferences in the form of concepts by mining search results and their click-through.

Reference [8] discussed an an Ontology-Based, Multi-Facet (OMF) personalization framework for automatically extracting and learning a user's content and location preferences based on the user's click-through.

Although these research works have reduced the amount of noise in results, many of them still cannot effectively capture the intentions of users. To address this issue, current solutions are applied to personalized search. A number of techniques and tools like bookmarks, history tools, search engines, metadata annotation and exploitation, and contextual recall systems have been developed to support personal web re-visitation. The most closely related work of this study is *Memento* system [12], which unifies context and content to aid web re-visitation. It defined the context of a web page as other pages in the browsing session that immediately precede or follow the current page, and then extracted topic-phrases from these browsed pages based on the Wikipedia topic list. In comparison, the context information considered in this work includes access time, location and concurrent activities automatically inferred. Reference [25] presented an algorithm in the personalization of web searches, called a Decision Making Algorithm, to classify the content in the user history.

There have been studies on click-through data of relevant feedback being introduced into retrieval systems. Reference [10] presented a context-based information refinding system called ReFinder which leverages human's natural recall characteristics and allows users to refind files and Web pages according to

the previous access context. ReFinder refinds information based on a query-by-context model over a context memory snapshot, linking to the accessed information contents. Context instances in the memory snapshot are organized in a clustered and associated manner, and dynamically evolve in life cycles to mimic brain memory's decay and reinforcement phenomena.

Generally, although most approaches on personalization have achieved good performance, some difficulties, such as real-time performance and user experience, prevent them from being widely applied. Our method provides an accurate and intelligent retrieval framework. Our framework captures certain aspects and the experimental results demonstrate the method's effectiveness.

3. INTELLIGENT RETRIEVAL FRAMEWORK

In this project, a personalized page rank algorithm is proposed, which ranks webpages based on user's interests' mechanisms using parameters such as keywords extraction, relevance feedback using click-through data and user's access context. The architecture of Intelligent Retrieval Framework is demonstrated below.

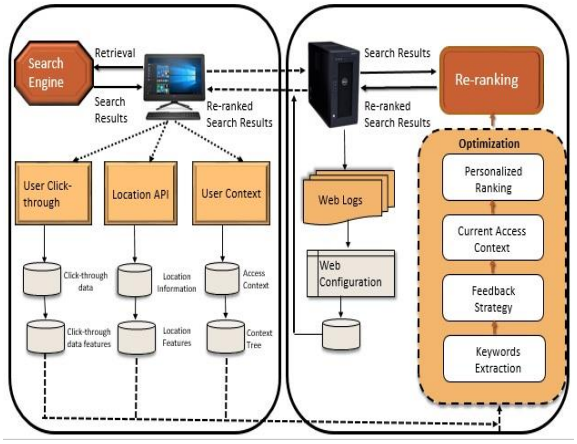

Fig 3.1 An Overview of Intelligent Retrieval Framework

The personalized page rank is calculated using the methods demonstrated below.

3.1 KEYWORD EXTRACTION

The optimized algorithm for the key-words extraction algorithm is based on a statistical model where the text is first broken into clauses, which go with one another by permutation and combination. Then, we use the optimized public substring extraction algorithm to address the clause set. Finally, we extract the keywords of the text according to the weights of the candidate keywords, which depend on the word frequency and word length.

The optimized keywords extraction algorithm presents improvements in terms of its space and time complexities.

3.2 RELEVANT FEEDBACK

A strategy to obtain users click-through data via implicit feedback is presented, which can improve the performance of search engines and the user's satisfaction. This aids in finding the information effectively.

Click set (CS)

Given a query keywords with accessible links and the CS satisfying

CS=(ID,Q,R,C); where

ID is number of the user's interest group and used to distinguish users in different groups;

Q is query keywords, which shows the query conditions of the retrieval;

R denotes a collection of all links returned from the search engine, in which the order of the links in the set is the display order on the webpage; and

C denotes a collection of all links clicked by the user.

Feedback Set (FS)

The FS is used to indicate the relevant feedback information obtained from the click data analysis, and the FS satisfies

FS=(ID,map); where

'map' is a relational table that stores relative degrees of correlation between two webpages.

The algorithm [1] works by extracting the links clicked by the user from the original set of links returned by the search engine. The more the number of clicks, the higher is the relevant degree of that link.

The relationship (l_i,l_j) is obtained which indicates that the relevant degree of link i is higher than that of link j for the keywords used in the query.

3.3 SEARCH CONTEXT

The 'search context' module captures the current access context (activities inferred from the currently running computer programs) into a probabilistic context tree.

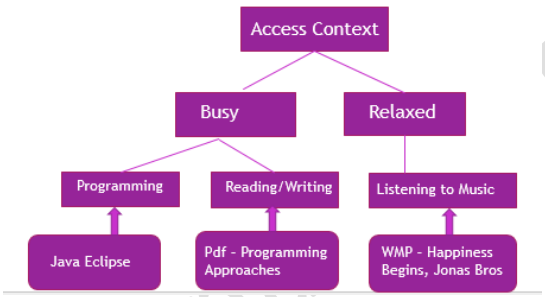

Fig. 3.4 An overview of Probabilistic Context tree for webpage access

The figure above gives a leveled probabilistic context tree that represents different activities of the user.

The obtained probabilistic context trees will evolve dynamically in life cycles to reflect the gradual degradation of human's episodic memorization as well as the context keywords that users will use for recall. That is, for each node in the probabilistic context tree, its association score will progressively decay with time. For different hierarchical values in the probabilistic context tree, as specific values at lower levels usually de-grade faster than general ones at upper levels in human's memory, different decay rates $_{leveli}$ (i = 1; 2; 3; · · ·) are assigned in line with the Ebbinghaus Forgetting Curve [10] , a graph illustrating how we forget information over time. Ebbinghaus took himself as a test subject to examine his own capacity to recollect information by creating a set of 2,300 three-letter, mean-ingless words to memorize. He studied multiples lists of these words and tested his recall of them at different time intervals over a period of one year. Ebbinghaus discovered that 58.2% was remembered after 20 minutes, 44.2% after 1 hour, 35.8% after 8-9 hours, 33.7% after 1 day, 27.8% after 2 days, and 25.4% after 6 days. Fitting formula $v = \sum_{n=1}^{v} e^{-\lambda\sqrt{t}}_{n}$ with these experimental values, we can calculate and obtain seven different decay rates, and the average decay rate approximates to 0.05. Based on these findings, we initialize the decay rates at different hierarchical levels by $\lambda_{level_1=0.05}$, $\lambda_{level_2=\frac{1}{2*1}\lambda_{level_1}}$, $\lambda_{level_3=\frac{1}{3*2}\lambda_{level_2}}$. Overall $\lambda_{level_{1+1}=\frac{1}{(i+1)*i}\lambda_{level_i}}$, (i=1,2,3…), whose values will be dynamically adjusted according to user's revisit queries and relevance feedback.

3.4 PERSONALIZED RANKING METHOD

The traditional PageRank algorithm is implemented based on linkage relations between webpages, but it ignores the importance of the webpages for different users. The proposed system presents a method whereby the relevant feedback information obtained from the click-through data is introduced into the PageRank algorithm.

According to the proposed relevant feedback information extraction strategy, we obtain the map table for the relationship of relevant degrees between links. However, the personalized PageRank value is influenced by not only the relationships among links but also the user click behavior. Thus, we regularly update the map table to more accurately reflect the current retrieval intention for the same group user.

The improvement in the traditional PageRank algorithm consists in adding a vector q, which represents the modification of the PageRank value using the relevant feedback information obtained from the click-through data. During the traversal of the map table, if the relevancy of link l_i for the same keywords is greater than link l_j and the webpage weight of link l_i is less than link l_j, we modify the weight stored in the database by the vector q. The calculation is as follows

$$q[l_i] = \frac{\sum l_i,l_j\ (\text{Rank}(l_i)-\text{Rank}(l_j)}{2} / N(l_i, l_j) \quad (1)$$

$$q[l_j] = -q[l_i]$$

Rank(l_i) represents the current weight of the link l_i in the database, and N(l_i,l_j) represents the number of relationships in the relevancy table. The click status of a user cannot represent other users; thus, we need to analyze and merge the click-through data of different users, which gradually makes the vector q perfect. Formula (3) represents the accumulation process of the modified vector q.

$$qold[l_i] = k1qold[l_i] + k2qnew[l_j] \quad (2)$$

$q_{old}[l_i]$ represents the original value of the modified vector for link l_i. $q_{new}[l_j]$ indicates the modified value calculated based on the relevancy of the newly acquired click-through data. Introducing the modified vector q into the traditional PageRank equation, the following formula (4) is obtained:

$$\forall l_i \quad Rank_{n+1} \quad (l_i) = \sum_{lj\in B_{li}} Rank_n(l_i) / Nl_j + q[l_i] \quad (3)$$

B_{li} represents the collection of all links in, and N_{lj} rep-resents the total number of chain links to the webpage. For formula (4), a variable d is added to control the coefficient of the modified vector q and the traditional PageRank value. The calculation is as follows:

$$\forall l_i\ Rank_{n+1}(l_i) = d * \sum_{l_j\in B_{l_i}} Rank_n(l_i) / Nl_j + (1-d)q[l_i] \quad (4)$$

Formula (4) and formula (5) add the modified vector q to the traditional PageRank. The corresponding formula including the webpage access probability C is as follows:

$$\forall l_i\ Rank_{n+1}\ (l_i) = \frac{d*[(1-C)+C*\sum_{l_j\in B_{l_i}} Rank_n(l_i)]}{N_{l_j}} + (1-d)q[l_i] \quad (5)$$

The relevant feedback information provided by different users is different, and the value of the modified vector q is different; therefore, the calculated personalized PageRank value also shows significant differences.

Therefore, even if users of different groups use the same retrieval keywords, the retrieved results will be reordered based on the value of the personalized PageRank.

The personalized page rank is further enhanced and the improvement is in adding vector 'v'.

$$\text{Context } v = \sum_{n=1}^{v} e^{-\frac{\lambda\sqrt{t}}{n}} \quad (6)$$

'λ' represents the Context tree order.

The final Context Score can be calculated as the product of personalized page rank value and context 'v'

Page Rank Score = Rank l(i) * Context (v) (7)

The calculation process of this personalized page rank algorithm is shown in the algorithm below

3.5 ALGORITHM

input - the relation of the link;

the relevant feedback information

output - personalized PageRank value

while *the PageRank value converges* do

Calculate PageRank value of webpage;

Calculate the value of related feedback vector according to formulas(1); (2); (3);

Calculate PageRank value according to formula(5);

Calculate context vector according to formula(6);

Calculate personalized page rank value according to formula(7);

end

4. PERFORMANCE METRICS

The web re-visitation performance metrics include pages' finding rate, average precision, average recall and aver-age rank error for a set of re-finding requests.

Assume a user's web re-visitation re-quest Q returns a ranked list of n result pages, from which the user aims to re-find target pages, and confirms m relevant result pages $\{w_1; \cdots; w_m\}$.

1) The finding of re-visitation Q is - Find(Q) = 1 if the user confirms one or more relevant result pages (i.e., m > 0), and 0 otherwise.

2) The precision of re-visitation Q is - $\text{Precision}(Q) = \frac{m}{n}$

3) The recall of re-visitation Q is - $\text{Recall}(Q) = \frac{m}{\Psi}$

4) The rank error of re-visitation Q is - $\text{RankError}(Q) = \sum_{j=1}^{m} \frac{\text{Pos}(Q,w_{i)-j}}{\text{Pos}(Q,w_i)} \Big/ m$

where function $\text{Pos}(Q, w_i)$ returns the position of the i-th confirmed page w_i in the result page list.

Let Q be a set of user's web re-visitation requests. The finding rate, average precision, average recall and average rank error of Q are thus defined as follows –

1) $\text{FindRate}(Q) = \frac{\sum Q \in Q \text{ Find } (Q)}{|Q|}$
2) $\text{Average Precision}(Q) = \frac{\sum Q \in Q \text{ Precision } (Q)}{|Q|}$
3) $\text{Average Recall}(Q) = \frac{\sum Q \in Q \text{ Recall } (Q)}{|Q|}$
4) $\text{Average Rank Error}(Q) = \text{Average } \frac{\sum Q \in Q \text{ Rank Error}(Q)}{|Q|}$
5) $\text{Average F1 Measure}(Q) = 2 * \frac{\text{Avg Precision} * \text{Avg Recall}}{\text{Avg Precision} + \text{Avg Recall}}$

5. EXPERIMENTAL RESULTS

The proposed approach proved to show better enhanced results that were more preferable than the existing methods and has outlined to overcome certain aspects with regard to personalized search.

The figure below compares the finding rate of the proposed approach with the existing approach.

Table 1. Performance comparison of average finding rate

Re-finding performance	Average Finding rate
Intelligent Retrieval approach	0.9242
Existing System	0.8235

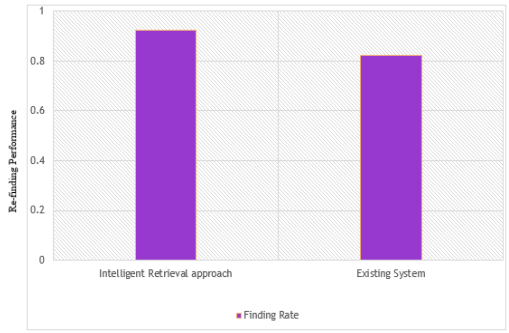

Fig.5.1 Performance comparison of finding rate

The results demonstrate accuracy in finding the desired webpage with minimum time. The finding rate of intelligent retrieval framework v is 92.10% compared to existing method's 81.11%.

The figure below compares the average rank error of the proposed approach with the existing approach.

Table 2. Performance comparison of Average rank error

Re-finding performance	Average rank error
Intelligent Retrieval approach	0.3145
Existing System	0.6105

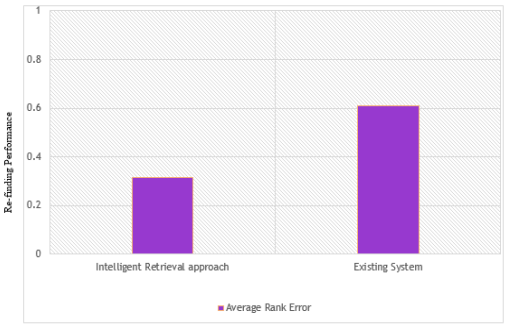

Fig. 5.2 Performance comparison of Average Rank Error

The results demonstrate accuracy in reducing average rank error of the webpage, thus providing an enhanced user experience.

The figure below compares the average F1 measure of the proposed approach with the existing approach.

Table 3. Performance comparison of Average F1 measure

Re-finding performance	Average F1 measure
Intelligent Retrieval approach	0.4152
Existing System	0.1951

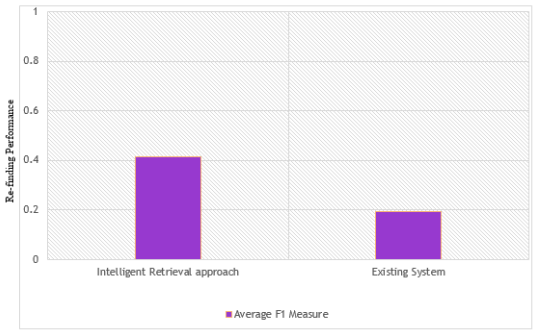

Fig. 5.3 Performance comparison of Average F1 measure

The results demonstrate accuracy in delivering a greater average F1 measure.

6. CONCLUSION

The proposed approach presents an intelligent retrieval model and a strategy for implicit correlation feedback based on click-through data analysis is considered, which obtains the relationship between the user query conditions and retrieval results. Finally, a personalized PageRank algorithm including modified parameters is designed to improve the ranking quality of the retrieval results using the relevant feedback from other users in the interest group.

This has been extended by adding the concept of 'Personal Web Re-visitation by Context' which captures the user's current access context while searching. The proposed method demonstrates that this

context and content based re-finding obtains remarkable retrieval performances with minimum effort and provides a superior user experience.

Finally, the proposed framework not only provides an efficient, intelligent, feedback based personalized retrieval approach but also provides the access context to reflect the user's behavior.

REFERENCES

[1] Yayuan Tang, Hao Wang, Kehua Guo, "Relevant Feedback Based Accurate and Intelligent Retrieval on Capturing User Intention for Personalized Websites", IEEE Access, 2018, vol.6, pp.2169-3536.A.

[2] Xu, S., Jiang, H., & Lau, F.C.M, "Mining user dwell time for personalized web search re-ranking", AAAI Press, 2011, pp. 23672372.

[3] Wu, Y., Shou, L., Hu, T., & Chen, G, "Query Triggered Crawling Strategy: Build a Time Sensitive Vertical Search Engine", in Proc. of IEEE CW, 2008, pp. 422-427.

[4] Liu, X.,& Turtle, H., "Real-time user interest modeling for real-time ranking", ASIS&T, 2013, pp. 64(8), 15571576.

[5] Haveliwala, T. H., "Topic-Sensitive PageRank: A Context-Sensitive Ranking Algorithm for Web Search", IEEE Transactions on Knowledge & Data Engineering, 2003, pp. 15(4), 784-796.

[6] Liu, J., & Belkin, N.J., Personalizing information retrieval for multi-session tasks: Examining the roles of task stage, task type, and topic knowledge on the interpretation of dwell time as an indicator of document usefulness, ASIS&T, 2015, pp. 66(1), 5881.

[7] Leung, W. T., Ng, W., & Lee, D. L., "Personalized Concept-Based Clustering of Search Engine Queries", IEEE Transactions on Knowledge & Data Engineering, 2011, pp. 20(11), 1505-1518.

[8] Leung, W. T., Lee, D. L., & Lee, W. C., "Personalized Web search with location preferences", in Proc. of IEEE ICDE, 2010, Vol.41, pp.701-712.

[9] L. Tauscher and S. Greenberg, "Re-visitation Patterns in World Wide Web Navigation", In CHI, 1997, pp. 399–406.

[10] T. Deng, L. Zhao, H. Wang, Q. Liu, and L. Feng., "ReFinder: A Context-Based Information Re-finding System", IEEE TKDE, 2013, 25(9):2119–2132.

[11] Agichtein, E., Brill, E., & Dumais, S., "Improving web search ranking by incorporating user behavior information", In Proc. of ACM SIGIR, 2006, pp. 1926.

[12] C. E. Kulkarni, S. Raju, and R. Udupa. "Memento: unifying content and context to aid webpage re-visitation." In *UIST*, pp. 435–436, 2010.

[13] Chuklin, A., Markov, I.,& de Rijke, M., "Click Models for Web Search. Synthesis Lectures on Information Concepts, Retrieval, and Services", Morgan & Claypool Publishers, 2015.

[14] Guo, F., Liu, C., Kannan, A., Minka, T., Taylor, M., "Click chain model in web search", in Proc. of ACM, 2009, pp (11-20).

[15] Jing, Y., & Baluja, S, "Visualrank: applying pagerank to large-scale image search", IEEE Transactions on Pattern Analysis & Machine Intelligence, 2008, pp. 30(11), 1877.

[16] Pavel Berkhin, "A survey on pagerank computing", Internet Mathematics, 2005, pp. 2(1), 73-120.

[17] 2(1), 73-120, "Pmse: a personalized mobile search engine", IEEE Transactions on Knowledge & Data Engineering, 2013, pp. 25(4), 820-834.

[18] Divya, R., & Robin, C. R. R., "Onto-search: An ontology based personalized

mobile search engine", in Proc. of ICGCCEE, 2014, pp.(1-4).

[19] Gardarin, G., Kou, H., Zeitouni, K., Meng, X., & Wang, H, "SEWISE: An Ontology-based Web Information Search Engine. Natural Language Processing and Information Systems", International Conference on Applications of Natural Language To Information Systems, 2008, pp. (106-119).

[20] Zhang, Y., Yang, X., & Mei, T., "Image search reranking with query-dependent click-based relevance feedback", IEEE Transactions on Image Processing, IEEE Transactions on Image Processing, 2014, pp. 23(10), 4448.

[21] Cui, H., Wen, J. R., Nie, J. Y., & Ma, W. Y., "Probabilistic query expansion using query logs.", 2002, pp 325-332.

[22] Smyth, B., Balfe, E., Freyne, J., Briggs, P., Coyle, M., & Boydell, O, "Exploiting query repetition and regularity in an adaptive community-based web search engine", User Modeling and User-Adapted Interaction, pp. 14(5), 383-423.

[23] Burke, R., & Ramezani, M., "Matching recommendation tech-nologies and domains. Recommender Systems Handbook", Recommender Systems Handbook, 2011, pp. 367-386.

[24] Abdullah, N. Y., Husin, H. S., Ramadhani, H., & Nadarajan, "Pre-processing of query logs in web usage mining", 2012, pp. 11(1), 82-86.

[25] Choong, C. Y., Mikami, Y., & Nagano, R. L., "Language identification of web pages based on improved n-gram algorithm", International Journal of Computer Science Issues, 2011, pp. 8(3).

[26] Yi, X., Hong, L., Zhong, E., Liu, N.N., & Rajan, S. "Beyond clicks: Dwell time for