Web Page Personalization Techniques in the purview of Page Ranking Methodology using Artificial Intelligence Approach

Mohammad Suaib
Dept. of Computer Sc.
Integral University, Lucknow
Email: suaib [AT] iul.ac.in

Mohammadi Akheela Khanum
Dept. of Computer Sc.
Integral University, Lucknow
Email: akheela [AT] iul.ac.in

Abstract: **The advent of fast speed internet and increase in internet users of smart phones has led an unimaginable increase in internet data on Web. Since there is no centralized monitoring of data to be stored, indexed and retrieved on web, it throws a smart challenge to Search Engines to retrieve queried information from the Web not just in time but also to the exact and close precision of user interest and mean. So, in view of exponential increase in size of Web information, the Web Search Engines must be smart & able enough to obtain the queried information as per the need and preferences of internet users.**

There are various schemes of personalized web ranking such Personalized Page Ranking (PPR), User Interest Score (UIS), Term Frequency and Inverse Document Frequency (TF-IDF), User Interest Hierarchy (UIH) which have been evolving alongside the advancement of internet technologies and contemporary growths in Web Information but with the passage of time these techniques either lack on efficiency or become obsolete to cater the need and interest of internet user to their satisfaction. This paper performs an extensive survey on a wide range of web personalization schemes using page ranking methodology in the purview of artificial intelligence approach. It does performs a comparative analysis among studied schemes and establishes a relation among them with respect to their effectiveness and novelty in conformity with user preferences and interests.

Keywords: Personalized web search, User-preference, Personalized Page Ranking (PPR), User Interest Score (UIS), Term Frequency and Inverse Document Frequency (TF-IDF), User Interest Hierarchy (UIH)

I. INTRODUCTION

Search engine are meant for retrieving web information in response to the query input by user. But understanding the intention of user behind his/her query is extremely important for search engine in order to be able to give the most accurate result of search to the utter satisfaction and preferences of user. Though there are various web search technologies but still they lacks of many grounds and conditions in which search engine users are not satisfied with the search results produced by them. Different users for the same query might have different preferences and interests for the Web search results and similarly different queries from same user but from a different location might have different emphases on web search results returned by search engines.

A particular query could mean different things in different context and the expected context can be apprehend by the user itself only. think of a specified query "silver", a user might be searching a ornament of silver, or a color or for a kind of fish. Traditional search engines used to provide alike set of results without apprehending the intent of user behind the query.

Therefore, the requirement of personalized web search mechanism in order to produce the most appropriate search results as highly ranked pages. A web search personalization is subject to different levels of efficiency for different queries, different users as well as search contexts.

As the size of internet based information grows exponentially, the voluminous raw data need to be fetched smartly in an efficient manner. There are several approaches and schemes to obtain, tailor and use digital data as per our requirement. To deal with customization and manipulation of data to draw useful information we need some lucid techniques for that and there come Data Mining in picture to play its role.

There exist many approaches, methods and goals for data mining. to quote a few of them are- Evolutionary algorithms (EA) [1]. This includes biology inspired algorithms such as Genetic Algorithms (GA)[2]. Differential Evolution Algorithms (DE)[3], and swarm based approaches like Ant Colonies[4], and Particle Swarm Optimizations (PSO) [5]. DM has also been used as classifier using Artificial Bee Colony (ABC) algorithm[6]. Some other most common classifiers for Data mining are PART, DPSO, SOM, Naive Bayes, Classification Tree and Nearest Neighbour (KNN).

ABC algorithm is a kind of new swarm intelligent algorithm, invented by Karabog in Erciyes University of Turkey in 2005 . This algorithm is simple to understand by concept, easy to apply, and has less control attributes and parameters. For this very reason this algorithm has been widely used in many optimization applications like in digital IIR filters [7][8] , Artificial Neural Networks [9][10][11] and many others applications[12][13][14].

Web systems utilize the User Relevance Feedback [15] to construe the information as per the user's need & choices. The vector space model computes the similarity between the query and the document and is based on the terminological overlap between them. Relevance Feedback requires the user to classify the documents into relevant and irrelevant groups.

Rocchio algorithm is used to expand the queries from the feedback thus obtained. Users are generally reluctant to provide information on whether they are interested in a particular document or not, so relevance feedback is not satisfying mechanism to fulfill the user needs.

II. ANALYSIS : PAST RELATED WORK

Though there are various web search technologies but still they lacks of many grounds and conditions in which search engine users are not satisfied with the search results produced by them. traditional Search Engine produces search results on the basis of keyword matches without apprehending the user's needs and preference. Ramadhan [16] proposed a heuristic based solution to differentiate the significance of various back links by assigning a different weight factor to them depending on their location in the directory tree of the Web space. This Rank computation completely relies on the link structure of a web page and hence it fails to consider the user's interest.

There is a novel approach to rank the page based on user search history and many other local preference of user. In such algorithms, Web pages are ranked at first and ordered according to ranks and then it is returned as search result for user. In order to measure the relative importance of web pages, Page-Ranking method is used for computing a ranking for every web page based on the graph of the web. PageRank has applications in search, browsing, and traffic estimation.

There are some other technique for search result optimization. In this context, the Artificial Bee Colony "ABC" algorithm produces good results in the optimization problem because ABC has many advantages like memory, local search and solution improvisation phenomenon [7][10][13] and[12][17].

However, in some cases, researchers found ABC gets stuck in local optimum that consequently leads to the convergence performance and get into uncertainties on the results obtained from the standard ABC algorithm [18] [19] & [20].

Aderhold et al. explored the influence of the population size of the ABC and suggested two variants of ABC which use new approaches for updating the position of artificial bees [21]. In a work of Stanarevic et al. a modified ABC was proposed which includes ''smart bee" concept that uses its historical memories of location and quality of the food sources [22][23]. Lei et al, invented that original ABC is suffering from low accuracy and effectiveness while solving optimization problems therefore they proposed a modification of the original ABC by adding a special kind of weight which was influenced by particle swarm optimization [24].

In addition to the approaches mentioned above, there are a number of Meta-Heuristics approaches also and some of them have been proven to be the most successful meta-heuristic algorithms include genetic algorithm (GA)[25], [26], [27], ant colony optimization [28],[29], particle swarm optimization (PSO)[30],[31], and artificial bee colony (ABC)[32] . Some of these meta-heuristics classes have been developed in the recent past, which include cuckoo search[33], [34], seven-spot ladybird optimization (SLO)[35] and bacteria foraging algorithm (BFA)[36]. These metaheuristic algorithms have been tested and are widely applicable in different fields of problems such as manufacturing, scheduling, services, transportation, geology, astronomy and what not[37].

Despite of seeing all these positive sides of algorithms they still suffer from the bottleneck and all these algorithms do stand fairly good equally across all sorts of problems, few perform better in a specific application domain whereas the same may not perform in another class of problem.

Web personalization could be achieved by organizing the user profile as User Interest Hierarchy (UIH) [38].

UIH tracks the user interest implicitly and DHC algorithm is used for the same in order to classify the results. Different characteristics of a term are derived and accordingly the terms are scored. This approach does not present any consideration for merging the current term which is similar to the existing term in the hierarchy. UIH could be refined by specifying two new characteristics namely term and node specificity [39].

Using these features the top results can be re-ranked. But the same approach fails to handle some new queries that are provided by users. A weighted URL ranking algorithm is used to rank the web search results based on the features extracted from hyperlinks, anchor terms and user interested domains.

The retrieved results from the search engines are weighed according to the occurrence of tokens and are again weighed in accordance with the user interested domain and the same are retained for re-ordering the results according to the match with the query weight. For personalization [40] some client side algorithms are developed keeping in view of user sentiments, usage behavior and search tendency. The different algorithms [41] used for link analysis like Page Rank (PR), Weighted Page Rank (WPR) and Hyperlink-Induced Topic Search (HITS) algorithms are discussed and compared.

III. COMPARISON: PAGE RANKING TECHNIQUES

Out of various techniques we studied above the "page ranking approach" is found good to be implemented for personalized web search. A major application of PageRank is searching. There are two search engines which use PageRank. The first one is a simple title-based search engine. The second one is a full text search engine such as Google[49]. Google utilizes a number of factors to rank search results including standard IR measures, proximity, anchor text (text of links pointing to web pages), and PageRank. While a comprehensive user study of the benefits of PageRank is beyond the scope of this paper, we have performed some comparative experiments and provide some sample results in this paper. The benefits of PageRank are the greatest for under specified queries. For example, a query for \Stanford University" may return any number of web pages which mention Stanford (such as publication lists) on a conventional search engine, but using PageRank, the university home page is listed first.

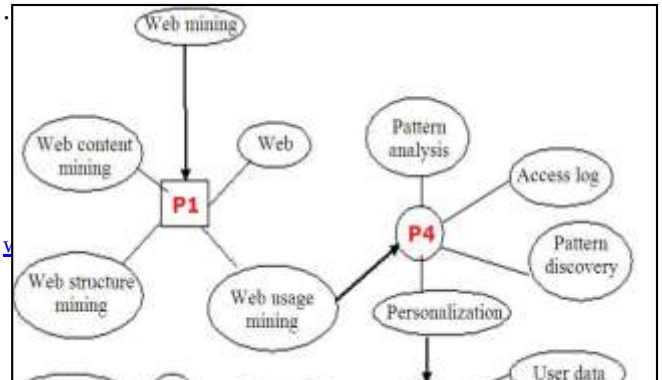

Fig: Tracking user interest through preferred network

A weighted URL ranking algorithm is used to rank the web search results based on the features extracted from hyperlinks, anchor terms and user interested domains The retrieved results from the search engines are weighed according to the occurrence of tokens and are again weighed in accordance with the user interested domain and the same are retained for re-ordering the results according to the match with the query weight. For personalization purpose some client side algorithms are developed keeping in view of user sentiments, usage behavior and search tendency. Various algorithms [41] that have been used for link analysis like Page Rank (PR), Weighted Page Rank (WPR) and Hyperlink-Induced Topic Search (HITS) algorithms have been discussed and compared above in our analysis section.

The rank of the relevant results is computed in accordance with the user interest. The ranking of a result considers both TF-IDF measure and user interest score (UIS). Comparing various ranking schemes, the hybrid (TF-IDF + UIS) scheme poses a good reflection in our analysis as shown in table-1 below.

Table-1: Comparative calculation of Page Ranking approaches

Preferred terms	TF-IDF	UIS	PPR	Hybrid (US+TF-IDF)
Web	Excellent	_	Fair	Excellent
Web usage mining	Good	Adequate	Good	Fairly Good
Web structure mining	Good	Fair	-	Good
Web content mining	Good	Neutral	Fair	Good
Personalization	Excellent	Good	Fair	Excellent
Pattern analysis	Good	Good	Poor	Excellent
Usage History	_	Good	Fair	Good

Personalized page rank is computed on various parameters in accordance with the user preferences. While computing the rank, the weight of the UIS and TF-IDF are kept varying according to the nature of the query and the user preferences.

In contrast to the traditional Web page ranking schemes, the Hybrid scheme uses a different set of parameter for ranking calculations for Keyword Indexing as shown below in table-2 –

Table-2: Query-Term Preference List Keyword Indexing

	Traditional	Hybrid
1	Web	Personalization
2	Web Mining	Usage data
3	Web structure mining	Profile
4	Web content mining	User data
5	Internet	Access log
6	Data mining	Web usage mining
7	Web Usage History	Pattern analysis

To implement the page Re-ranking following steps were best to be followed in the implementation of hybrid page ranking scheme.

- ✓ A set of documents that matches the user query is fetched from the search engine (top K documents)
- ✓ The terms in the initial set of documents are weighed using TF-IDF measure and by using the same the user preferred network of concepts is framed
- ✓ The network is tracked for UIS and the proposed feature weights are computed
- ✓ The result set is ranked based on computed UIS and TF-IDF value

In addition to these PR techniques we also have performed an extensive comparative analysis among latest page ranking approaches that have been used in the recent years from (2013 to 2019).

As result of our analysis we contemplated the following table-3 that focuses on the strength, weaknesses and the approaches used within each paper. Some of the techniques are found good in specific context and some fall short on the same. But overall each one of them possesses its own merits and virtues with itself.

Table-3: Comparative analysis among Page ranking approaches used in researches from 2013 to 2017

Year	Author	Approach	Pros	Cons
2013	Derhami V[42]	Simple Collaborative filtering approach has been used to decide on ranks of the pages.	Considers the similarity of users preferences to estimate ranks of web pages.	Newly introduced web get ranked improperly under this scheme despite having been strongly concerned with the web query.
2014	Roobam and Vallimayli[43]	CF is used over memory usages	CF based result is used while preprocessing on web pages	New web Pages go ranked

		Optimization to construe page ranking.	to achieve fair page rank for the relevant query.	miserably under this scheme even though they are highly concerned with the web query.
2015	Kaviarasanet al[44]	Collaborative filtering approach has been applied using K-Mean clustering to decide on ranking of pages.	Recommendation techniques to enhance page rank in context to web search query.	While ranking the page, it has been proven ineffective while dealing new web pages for the same query
2015	Moreno et al.[45]	Collaborative filtering approach is used in classification and association rules for deciding the page rank.	Hybrid Recommendation system and Ontology approach is used decide on page ranks w.r.t. user's query	Does not yields promising results on varied and complex query by different users with new queries.
2015	Bairagade et al.[46]	Web Crawler approach is referenced for improvising page rank.	Documents are preprocessed while crawling in web.	Unable to improve page ranking modules.
2016	Sharma and Lodhi [47]	ML (Machine Learning) and DT (Decision Tree) approaches are used to rank the pages.	User's Logs are referenced and Traditional machine Learning method used in page rank decision.	Fall short off on other applicable machine learning techniques
2017	Aqlan et al. [41]	NN (Neural Network) approach & regression technique has been used to decide page ranking	Latest AI techniques incorporated for improving page rank algorithm	Does not enquire about other traditional AI techniques befitting in page ranking

The table speaks up the advents and disadvantages of each approaches used in page ranking algorithm. The above table-3 depicts the strength, weaknesses and the approaches used within each paper. Some of the techniques are found good in specific context and some fall short on the same.

Among all approaches UIS shows better performances and construe promising results on personalized web search. Hence, based on the interest score of user the Page Rank is calculated from the preferred profile based on network. User preferences & choices are categorized and traced without user involvement. Finally the corresponding results are mapped and search result is presented to the user.

IV. CONCLUSION

Page-Rank is a global ranking of all available web pages regardless of its content it is based only on its location in the Web graph structure. Using PageRank, one can order search results in a way that important and most concerned Web pages can be given preference in search results. PageRank can be used to sort out a small set of frequently used documents which can answer most of the queries from user end. If the small database falls short to answer a search query then only full database on web needs to be consulted. Hence, PageRank could be a good way to fetch highly concerned and representative pages to display , for search result, for a cluster center. Also, the structure of the Web graph is very

useful for a variety of information retrieval tasks while referring a Page-Rank approach.

Among all available page ranking approaches, the hybrid page ranking algorithm (TF-IDF + UIS) stands tall and gives out promising result to the satisfaction of users up-to a major extent.

In future, to further improve the ranking of the search results, the profile convergence features could be analyzed thoroughly to give out even more précised and accurate personalized web search result.

REFERENCES

[1] Tan, Teoh, Yu, & Goh, "Evolutionary algorithms ", 2009; Whitley, 2001.
[2] Sumida, Houston, McNamara, & Hamilton, "Genetic Algorithms ", 1990.
[3] Storn & Price, "Differential Evolution Algorithms", 1997.
[4] Colorni, Dorigo, & Maniezzo, 1991.
[5] Kennedy J, "Particle Swarm Optimizations", 2006
[6] D. Karaboga , "Artificial Bee Colony (ABC) algorithm", 2005.
[7] N. Karaboga , "Digital IIR filters", 2009.
[8] Karaboga, N. " A new design method based on artificial bee colony algorithm for digital IIR filters. Journal of the Franklin Institute, 2009, 346(4), 328-348.
[9] D. Karaboga & Akay, "Artificial Neural Networks", 2005
[10] Karaboga, D., Akay, B., & Ozturk, C. " Artificial Bee Colony (ABC) Optimization Algorithm for Training Feed-Forward Neural Networks Modeling Decisions for Artificial Intelligence, 2007 (pp. 318-329).
[11] Karaboga, D., & Akay, B. B. "An artificial bee colony (abc) algorithm on training artificial neural networks (Technical Report TR06): Erciyes University, Engineering Faculty, Computer Engineering Department, 2005.
[12] Karaboga, D., & Basturk, B. (2007). A powerful and efficient algorithm for numerical function optimization: Aartificial bee colony (ABC) algorithm. Journal of Global Optimization, 39(3), 459-471.
[13] Karaboga, D., & Basturk, B. (2007). Artificial Bee Colony (ABC) Optimization Algorithm for Solving Constrained Optimization Problems Foundations of Fuzzy Logic and Soft Computing (pp. 789-798).
[14] Karaboga, D., & Basturk, B. (2008). On the performance of artificial bee colony (ABC) algorithm. Applied Soft Computing, 8(1), 687-697.
[15] Algarni, A., Y. Li and X. Tao, "Mining specific and general features in both positive and negative relevance feedback". Proceedings of the 19th Text REtrieval Conference: Relevance Feedback Track, Nov. 2010. pg. 16-19,
[16] Ramadhan, H.A., K. Shihab and J.H. Ali,. Improving the ranking capability of the hyperlink based search engines using heuristic approach. J. Comput. Sci., 2006 : 638-645. DOI: 10.3844/jcssp. 2006.638.645
[17] Zhao, H., Pei, Z., Jiang, J., Guan, R., Wang, C., Shi, X.,. A hybrid swarm intelligent method based on genetic algorithm an artificial bee colony. In: Advances in Swarm Intelligence. springer, 2010 pp. 558–565.
[18] Luo, J., Wang, Q., Xiao, X., "A modified artificial bee colony algorithm based on converge-onlookers approach for global optimization". Appl. Math. Computat. 2013, 219 (20), 10253–10262.
[19] Xiang, W.-L., An, M.-Q., 2013. An efficient and robust artificial bee colony algorithm for numerical optimization. Comput. Operat. Res. 40 (5), 1256–1265.
[20] Kong, X., Liu, S., Wang, Z., 2013. An improved artificial bee colony algorithm and its application. Int. J. Signal Process. Image process. Pattern Recognit. 2013 , 6 (6).
[21] Aderhold, A., Diwold, K., Scheidler, A., Middendorf, M., 2010. Artificial bee colony optimization: a new selection scheme and its performance. In: Nature Inspired Cooperative Strategies for Optimization (NICSO 2010). Springer, pp. 283–294.
[22] Stanarevic, N., Tuba, M., Bacanin, N., 2010. Enhanced artificial bee colony algorithm performance. In: Proceedings of the 14th WSEAS International Conference on Computers: Part of the 14th WSEAS CSCC Multiconference, vol. 2, pp. 440–445.
[23] Stanarevic, N., 2011. Comparison of different mutation strategies applied to artificial bee colony algorithm. In: Proceedings of the European computing conference (ECC11), pp. 257–262.
[24] Lei, X., Huang, X., Zhang, A., 2010. Improved artificial bee colony algorithm and its application in data clustering. In: Bio-Inspired Computing: Theories and Applications (BIC-TA), 2010 IEEE fifth International Conference on. IEEE, pp. 514–521.
[25] R. Haupt and S. Haupt, "The binary genetic algorithm," Pract. Genet. Algorithms, Second, pp. 27–50, 1998.
[26] M. Dorigo and G. Di Caro, "The Ant Colony Optimization Meta-Heuristic," pp. 11–32, 1999.
[27] M. Dorigo and T. Stützle, Ant Colony Optimization. 2004.
[28] J. Kennedy and R. Eberhart, "Particle swarm optimization," Neural Networks, Proceedings., IEEE Int. Conf., 1995.
[29] J. Kennedy and R. Eberhart, "Particle swarm optimization," Neural Networks,. Proceedings., IEEE Int. Conf., vol. 4, pp. 1942–1948 vol.4, 1995.

[30] R. Poli, J. Kennedy, and T. Blackwell, "Particle swarm optimization," Swarm Intell., vol. 1, no. 1, pp. 33–57, 2007.
[31] D. T. Pham, D. T. Pham, a. Ghanbarzadeh, a. Ghanbarzadeh, E. Koc, E. Koc, S. Otri, S. Otri, S. Rahim, S. Rahim, M. Zaidi, and M. Zaidi, 2008.
[32] "The bees algorithm–a novel tool for complex optimisation problems," Intell. Prod. Mach. Syst. 2nd I• PROMVirtual Conf. 3-14 July 2006, p. 454, 2006
[33] I. Fister, X. S. Yang, and D. Fister, "Cuckoo search: A brief literature review," Stud. Comput. Intell., vol. 516, pp. 49–62, 2014.
[34] X. S. Yang and S. Deb, "Cuckoo search via Lévy flights," 2009 World Congr. Nat. Biol. Inspired Comput. NABIC 2009 - Proc., pp. 210–214, 2009.
[35] K. M. Passino, "Bacterial foraging optimization," Int. J. Swarm Intell. Res., vol. 1, no. 1, pp. 1–16, 2010.
[36] P. Wang, Z. Zhu, and S. Huang, "A New Meta- Heuristic Technique for Engineering Design Optimization : Seven-Spot Ladybird Algorithm," in 2nd International Symposium on Computer, Communication, Control and Automation, 2013, pp. 387–392.
[37] M. J. Lehane, "The Biology of Blood-Sucking in Insects," Liverpool Sch. Trop. Med., p. 321, 2005.
[38] Kim, H.R. and P.K. Chan,. Personalized ranking of search results with learned user interest hierarchies from bookmarks. Proceedings of the 11th SIGKDD International Conference on Knowledge Discovery and Data Mining, Aug. 21-21, Chicago, Illinois, USA. 2005
[39] Hu, J. and P.K. Chan, 2008. Personalized web search by using learned user profiles in re-ranking. Florida Institute of Technology, USA.
[40] Teevan, J., S.T. Dumais and E. Horvitz,. Personalizing search via automated analysis of interests and activities. Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Aug. 15-19, ACM Press, Salvador, Brazil, pp: 449-456. DOI: 10.1145/1076034.1076111, 2005.
[41] Kumar, P.R. and A.K. Singh,. Web Structur Mining: Exploring Hyperlinks and Algorithms for Information Retrieval. Am. J. Applied Sci., 7: 840-845. DOI: 10.3844/ajassp. 2010.840.845, 2010.
[42] Derhami V, Khodadadian E, Ghasemzadeh M, ZarehBidoki AM (2013) Applying reinforcement learning for web pages ranking algorithms. Appl Soft Comput 13(4):1686–1692, 2013.
[43] Roobam R, Vallimayli V (2014) Survey on ontology based semantic web usage mining for enhanced recommendation model. Int J Sci Eng Res 5(5):1164–1169, 2014.
[44] Kaviarasan S, Hemapriya K, Gopinath K (2015) Semantic web usage mining techniques for predicting users' navigation requests. Int J Innov Res Comput Commun Eng 3(5):4261–4270,2015.
[45] Moreno MN, Segrera S et al (2015) Web mining based framework for solving usual problems in recommender systems: a case study for movies recommendation. Neurocomputing J 176(C):73–80, 2015.
[46] Bairagade R, Singh N, Afre N, Bhamare D (2015) A survey on SmartCrawler: a deep-web harvesting approach. Int J Adv Res Comput Eng Technol 4(9):3611–3614, 2015.
[47] Sharma S, Lodhi SS (2016) Development of decision tree algorithm for mining web data stream. Int J Comput Appl 138(2):34–43, 2016.
[48] Aqlan HAA, Ahmed S, Danti A (2017) Death prediction and analysis using web mining techniques. In: ICACCS, Coimbatore, India, 978-1-5090-4559-4/17/$31-EEE,2017.
[49] www.google.com, 18:00 IST, Oct, 2019.