Journal of
Data Acquisition and Processing

ISSN: 1004-9037
https://sjcjycl.cn/
DOI: 10.5281/zenodo.754708

AN EFFICIENT PERSONALIZED WEB SEARCH PROTECTION MECHANISM USING IMPROVED GREEDY ALGORITHM

[1]Dr. N. Shanmuga Priya, [2]Mr. A. Abdul Azeez, [3]Mr. A. Dhanuswar

[1]Associate Professor and Head, Department of Computer Applications, Dr. SNS Rajalakshmi College of Arts & Science, Coimbatore.

[2]PG Student, II MCA, Department of Computer Applications, Dr. SNS Rajalakshmi College of Arts & Science, Coimbatore.

[3]PG Student, II MCA, Department of Computer Applications, Dr. SNS Rajalakshmi College of Arts & Science, Coimbatore.

ABSTRACT: - Personalization is a push to reveal most applicable reports utilizing data about user's objective, area of enthusiasm, scanning history, question connection and so on that gives higher worth to the user from the substantial arrangement of results. As the web substance is developing exponentially, more refined strategies are required to convey the applicable substance to the individual user. The most widely recognized challenges experienced while looking the Web are: i) Problems with the information itself ii) Problems confronted by the users attempting to recover the information they need iii) Problems in comprehension the connection of pursuit solicitations and iv) Problems with distinguishing the adjustments in user's data need. The principal explanation for every one of the issues is size of the web that restrains its utility. Here, introduce two eager calculations, in particular GreedyDP and GreedyIL, for runtime speculation. We likewise give an online forecast instrument to choosing whether customizing an inquiry is advantageous. Broad trials exhibit the adequacy of our system. The test comes about likewise uncover that GreedyIL altogether beats GreedyDP regarding productivity. Consequently, this paper exhibits another plan that creates twisted user inquiries from a semantic perspective with a specific end goal to save the value of user profiles. Plus, phonetic examination strategies are utilized to appropriately translate complex inquiries performed by users and create new semantically-related ones in like manner. The execution of the new plan is assessed as far as semantic protection of new inquiries, security level and runtime.

Keywords: - Opinion mining, opinion targets extraction, opinion words extraction, expectation maximization.

1. INTRODUCTION

Data mining is a powerful new technology with great potential to help companies focus on the most important information in their data warehouses. It has been defined as, the automated analysis of large or complex data sets in order to discover significant patterns or trends that would otherwise go unrecognized.

Data mining tools can also automate the process of finding predictive information in large databases. Questions that traditionally required extensive hands-on analysis can now be answered directly from the data — quickly [1]. A typical example of a predictive problem is targeted marketing. Data mining uses data on past promotional mailings to identify the targets most likely to maximize return on investment in future mailings. Other predictive problems include forecasting bankruptcy and other forms of default, and identifying segments of a population likely to respond similarly to given events.

Data mining techniques can yield the benefits of automation on existing software and hardware platforms to enhance the value of existing information resources, and can be implemented on new products and systems as they are brought on-line.

Web personalization is a strategy, a marketing tool, and an art. Personalization requires implicitly or explicitly collecting visitor information and leveraging that knowledge in your content delivery framework to manipulate what information you present to your users and how you present it. Correctly executed, personalization of the visitor's experience makes his time on your site, or in your application, more productive and engaging. Personalization can also be valuable to you and your organization, because it drives desired business results such as increasing visitor response or promoting customer retention [2]. Unfortunately, personalization for its own sake has the potential to increase the complexity of your site interface and drive inefficiency into your architecture. It might even compromise the effectiveness of your marketing message or, worse, impair the user's experience.

Web personalization can be seen as an interdisciplinary field that includes several research domains from user modeling, social networks, web data mining, human-machine interactions to Web usage mining; Web usage mining is an example of approach to extract log files containing information on user navigation in order to classify users. Other techniques of information retrieval are based on documents categories' selection. Contextual information extraction on the user and/or materials (for adaptation systems) is a technique fairly used also includes, in addition to user contextual information, contextual information of real-time interactions with the Web. Proposed a multi-agent system based on three layers: a user layer containing user's profiles and a personalization module, an information layer and an intermediate layer. They perform an information filtering process that reorganizes Web documents. Propose reformulation query by adding implicit user information. This helps to remove any ambiguity that may exist in query: when a user asks for the term "conception", the query should be different if he is an architect or a computer science designer [3].

Requests can also be enriched with predefined terms derived from user's profile develop a similar approach based on user categories and profiles inference. User profiles can be also used to enrich queries and to sort results at the user interface level. Other approaches also consider social-based filtering and collaborative filtering. These techniques are based on relationships inferred from users' profile. Implicit filtering is a method that observes user's behavior and activities in order to categorize classes of profile.

2. LITERATURE REVIEW

Search personalization is based on the fact that individual users tend to have different preferences and that knowing the user's preference can be used to improve the relevance of the results the search engine returns. There have been many attempts to personalize web search.

Deepak, G., et.al, (2020), proposed Personalization of the recommendation of web pages is certainly necessary to estimate the user interests for suggesting web pages as per their choices [4]. A framework for personalized web page recommendation based on a hybridized strategy is proposed. Web Pages are recommended based on the user query by analyzing the Web Usage

Data of the users. An array of strategies is intelligently integrated together to achieve an efficient Web Page Recommendation system. Latent Semantic Analysis is applied to the User-Term Matrix and the Term-Frequency Matrix that are built from the Web Usage Information to form a Term Prioritization Vector. Further, techniques like Latent Dirichlet Allocation for Topic-based Segregation of the URLs and Normalized Pointwise Mutual Information strategies are used for recommending web pages based on users' queries. The Personalization is achieved by prioritizing the Web pages based on the Prioritization Vector. Also, a unique methodology is incorporated into the system to retrieve trustworthy websites. An overall Accuracy of 0.84 is achieved which is better than the existing strategies.

Sajjan, R. S., & Veer, S. A. (2019), proposed Web search engines (e.g. Google, Yahoo, Microsoft Live Search, Bing, etc.) are mostly used to search certain information from a large amount of data in a very few amount of time [5]. Aforementioned engines are built for all kind of people and not for any particular client that is, it gives generalized result for input query and not user specific result, to address this problem personalized web search is best way to increase the accuracy of web search in terms of giving user specific results. However, effective personalized web search requires gathering and aggregating user information (e.g. user name, contact no, etc), which often raises serious concerns of privacy infringement for many users. In fact, these privacy concerns have become one of the major reasons for deploying personalized web search applications and how to do privacy-preserving personalization is a great challenge. In this proposed system, we propose and try to resist adversaries with broader background knowledge, such as richer relationship among topics. Richer relationship means we generalize the user profile results by using the background knowledge which is going to store in history. Through this we are able to hide the user search results.

Alweshah, M., et al., (2023), addressed the complexities of GS are especially noticeable in the context of microarray expression data analysis, owing to the inherent data imbalance [6]. The main goal of this study is to offer a simplified and computationally effective approach to dealing with the conundrum of attribute selection in microarray gene expression data. We use the Black Widow Optimization algorithm (BWO) in the context of GS to achieve this, using two unique methodologies: the unaltered BWO variation and the hybridized BWO variant combined with the Iterated Greedy algorithm (BWO-IG). By improving the local search capabilities of BWO, this hybridization attempts to promote more efficient gene selection. A series of tests was carried out using nine benchmark datasets that were obtained from the gene expression data repository in the pursuit of empirical validation. The results of these tests conclusively show that the BWO-IG technique performs better than the traditional BWO algorithm.

Rakesh, G., & Pratibha, M. (2016), suggested that Internet now a day's growing widely and continuously, as it is increasing in its size the users who use the services of it demanding search results for their queries accurately and quickly [7]. This raises concerns of privacy issues however as users are typically uncomfortable to provide their personal information to the service provider on the Internet. The aim is to deal with the privacy issues surrounding personalized search and discusses ways that privacy can be improved so that users can become more comfortable with the release of their personal data in order to receive more accurate search results.

Bhandare, M. S. K., & Kapse, A. S. (2021), proposed personalized web search (PWS) has demonstrated its effectiveness in improving the quality of various search services on the Internet [8]. However, evidences show that users' reluctance to disclose their private information during search has become a major barrier for the wide proliferation of PWS. We study privacy protection in PWS applications that model user preferences as hierarchical user profiles. We propose a PWS framework called UPS that can adaptively generalize profiles by queries while respecting user- specified privacy requirements. Our runtime generalization aims at striking a balance between two predictive metrics that evaluate the utility of personalization and the privacy risk of exposing the generalized profile. We present two greedy algorithms, namely GreedyDP and GreedyIL, for runtime generalization. We also provide an online prediction mechanism for deciding whether personalizing a query is beneficial. Extensive experiments demonstrate the effectiveness of our framework.

Kumar, K., & Bright Keswani, D. V. J. (2020), define the problem of Web search engines are our daily needs now days to find desired information on the internet [9]. The search engines build user profiles from the user search history. Accurate and rich user profile serve better personalized search results but pose high risk to user privacy. Not only this, sensitive information infringement and unsolicited advertisement is also a big issue. So, it is necessary to stop collection of sensitive data of user by obfuscating at client side so that user identity can be safeguarded. So, this work is carried out keeping in mind privacy protection of user and personalization. In this work, actual queries are obfuscated with a number of dummy queries which are semantically related to actual query. Numbers of dummy queries are added with a semantic distance which is controlled by the user.

3. EXISTING METHODOLOGY & ITS STRUCTURE

The existing profile-based Personalized Web Search does not support runtime profiling. A user profile is typically generalized for only once offline, and used to personalize all queries from a same user indiscriminatingly. Such "one profile fits all" strategy certainly has drawbacks given the variety of queries. One evidence reported in is that profile-based personalization may not even help to improve the search quality for some ad hoc queries, though exposing user profile to a server has put the user's privacy at risk [10].

The existing methods do not take into account the customization of privacy requirements. This probably makes some user privacy to be overprotected while others insufficiently protected. For example, in, all the sensitive topics are detected using an absolute metric called surprisal based on the information theory, assuming that the interests with less user document support are more sensitive. However, this assumption can be doubted with a simple counterexample: If a user has a large number of documents about "sex," the surprisal of this topic may lead to a conclusion that "sex" is very general and not sensitive, despite the truth which is opposite. Unfortunately, little prior work can effectively address individual privacy needs during the generalization.

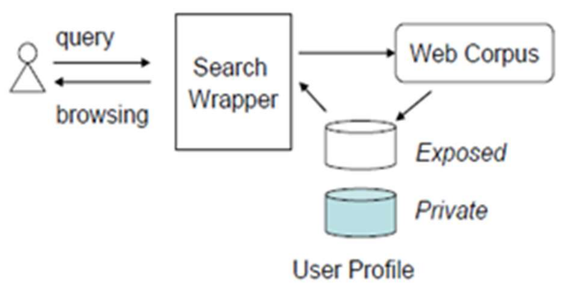

Fig 3.1: - Existing System Structure

Figure 3.1 provides an overview of the whole system. An algorithm is provided for the user to automatically build a hierarchical user profile that represents the user's implicit personal interests. General interests are put on a higher level; specific interests are put on a lower level. Only portions of the user profile will be exposed to the search engine in accordance with a user's own privacy settings [11]. A search engine wrapper is developed on the server side to incorporate a partial user profile with the results returned from a search engine. Rankings from both partial user profiles and search engine results are combined. The Personalized results are delivered to the user by the wrapper.

The solution has three parts: First, a scalable algorithm automatically builds a hierarchical user profile from available source data. Then, privacy parameters are offered to the user to determine the content and amount of personal information that will be revealed. Third, a search engine wrapper personalizes the search results with the help of the partial user profile.

Unfortunately, the previous works of privacy preserving PWS are far from optimal [14]. The problems with the existing methods are explained in the following observations:

1. The existing profile-based PWS do not support runtime profiling. A user profile is typically generalized for only once offline, and used to personalize all queries from a same user indiscriminatingly. Such "one profile fits all" strategy certainly has drawbacks given the variety of queries.
2. It is proved that Profile-based personalization may not even help to improve the search quality for some ad hoc queries, though exposing user profile to a server has put the user's privacy at risk.
3. The existing methods do not take into account the customization of privacy requirements. This probably makes some user privacy to be overprotected while others insufficiently protected. For example, in all the sensitive topics are detected using an absolute metric called surprised based on the information theory, assuming that the interests with less user document support are more sensitive.
4. Many personalization techniques require iterative user interactions when creating personalized search results. They usually refine the search results with some metrics which require multiple user interactions, such as rank scoring, average rank, and so on. This paradigm is, however, infeasible for runtime profiling, as it will not only pose too much risk of privacy breach, but also demand prohibitive processing time for profiling. Thus, we need predictive metrics to measure the search quality and breach risk after personalization, without incurring iterative user interaction.

Any personal documents such as browsing history and emails on a user's computer could be the data source for user profiles. Our hypothesis is that terms that frequently appear in such documents represent topics that interest users [13]. This focus on frequent terms limits the dimensionality of the document set, which further provides a clear description of users' interest. This approach proposes to build a hierarchical user profile based on frequent terms.

4. PROPOSED METHODOLOGY & ITS CONTRIBUTIONS

The proposed work introduced a security saving personalization web seek structure UPS, which can sum up profiles for every question as indicated by user determined protection necessities. Depending on the meaning of two clashing measurements, in particular personalization utility and security hazard, for progressive user profile, figure the issue of protection saving personaization seeks as Risk Profile Generalization, with its NP-hardness demonstrated [12].
It creates two basic yet compelling speculation calculations, GreedyDP and GreedyIL, to bolster runtime profiling. While the previous tries to amplify the discriminating power (DP), the last endeavors to minimize the information loss (IL). By abusing various heuristics, GreedyIL beats GreedyDP fundamentally. It gives a modest system to the user to choose whether to customize a question in UPS. This choice can be made before each runtime profiling to improve the solidness of the indexed lists while maintain a strategic distance from the pointless presentation of the profile.

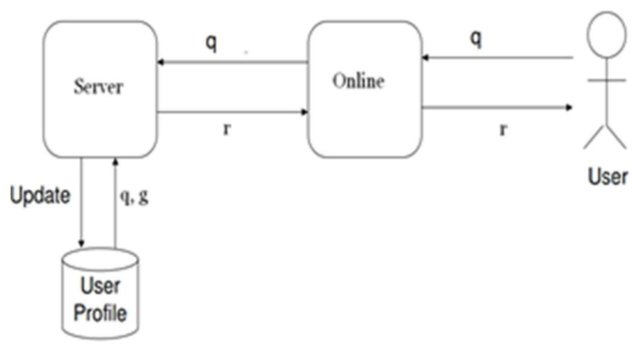

Fig 4.1: - General Architecture of Proposed Scheme

A greedy algorithm is a numerical procedure that recursively builds a set Recursion of articles from the littlest conceivable constituent parts. is a way to deal with critical thinking in which the answer for a specific issue relies on upon answers for littler occasions of the same issue.
Greedy algorithms search for straightforward, simple to-actualize answers for complex, multi-step issues by choosing which next stride will give the most evident advantage. Such calculations are called covetous on the grounds that while the ideal answer for each littler occasion will give a quick yield, the calculation doesn't consider the bigger issue all in all. Once a choice has been made, it is never reevaluated [15].
The preferred standpoint to utilizing a ravenous calculation is that answers for littler examples of the issue can be clear and straightforward. The detriment is that it is totally conceivable that the most ideal fleeting arrangements may prompt the most noticeably bad long haul result.
Greedy algorithms are regularly utilized as a part of bundles with the least number of jumps machine learning, business intelligence (BI), artificial intelligence (AI) and programming.

4.1. GreedyDP Algorithm

Given the complexity of our problem, a more practical solution would be a near-optimal greedy algorithm. The first greedy algorithm GreedyDP works in a bottom up manner. Starting from G0, in every ith iteration, GreedyDP chooses a leaf topic t € TGi (q) for pruning, trying to maximize the utility of the output of the current iteration, namely Gi+1. During the iterations, we also maintain a best profile so-far, which indicates the Gi+1 having the highest discriminating power while satisfying the ϑ risk constraint.

The iterative process terminates when the profile is generalized to a root-topic. The best-profile-so-far will be the final result (G*) of the algorithm. The main problem of GreedyDP is that it requires re-computation of all candidate profiles (together with their discriminating power and privacy risk) generated from attempts of prune-leaf on all t € TGi (q). This causes significant memory requirements and computational cost.

4.2. GreedyIL Algorithm

The GreedyIL algorithm improves the efficiency of the generalization using heuristics based on several findings. One important finding is that any prune-leaf operation reduces the discriminating power of the profile. In other words, the DP displays monotonicity by prune-leaf.

The benefits of making the above runtime decision are twofold:

1. It enhances the stability of the search quality;
2. It avoids the unnecessary exposure of the user profile.

However, this is extremely rare in practice. Therefore, GreedyIL is expected to significantly outperform GreedyDP.

Algorithm 1: GreedyIL($\mathcal{H}$, q, δ)

Input : Seed Profile $\mathcal{G}_0$; Query q; Privacy threshold δ
Output: Generalized profile $\mathcal{G}*$ satisfying δ-Risk

```
let Q be the IL-priority queue of prune-leaf decisions;
    i be the iteration index, initialized to 0;
// Online decision whether personalize q or not
if DP(q, R) < μ then
    Obtain the seed profile G_0 from Online-1;
    Insert ⟨t, IL(t)⟩ into Q for all t ∈ T_H(q);
    while risk(q, G_i) > δ do
        Pop a prune-leaf operation on t from Q;
        Set s ← par(t, G_i);
        Process prune-leaf G_i --(-t)--> G_{i+1};
        if t has no siblings then                 // Case C1
            Insert ⟨s, IL(s)⟩ to Q;
        else if t has siblings then               // Case C2
            Merge t into shadow-sibling;
            if No operations on t's siblings in Q then
                Insert ⟨s, IL(s)⟩ to Q;
            else
                Update the IL-values for all operations on
                t's siblings in Q;
        Update i ← i + 1;
    return G_i as G*;
return root(R) as G*;
```

Most web search engines consider almost no data about the user and their security strategies, thus the consequence of an inquiry is excessively summed up and not protection ensured. Keeping in mind the end goal to make more particular inquiries as indicated by the user profile, the user intrigues based web indexes have created. UPS, short for Users Personalized security safeguarded Search is personalization hunt and protection insurance framework. The profile coordinator is a system that permits the user to manufacture a progression with the as of now slithered reports. This progressive system can be constructed utilizing a bunching calculation or a characterization calculation. These offer two unique calculations to give two diverse approaches to arrange the archives. For the grouping calculation, the user just needs to give data about the quantity of bunch, requiring little exertion, thus time, for the user. The order calculation requires more support from the user, as he/she needs to give data about every class. The upside of the grouping calculation versus the bunching calculation is that the subsequent arrangement will be more personalization by the user, thus it will fit his interests superior to the bunching calculation.

The Privacy segment plays out the capacity of arrangement coordinating of the user seek profiles that can be of enthusiasm for the user with the thought of personalized policy [16]. To accomplish this, a chain of importance for the user profile and the route history of the user are vital. The Query is a web application. This application is specifically used to enter the inquiry questions and recover their outcomes with dependability.

5. EXPERIMENTAL RESULTS

Expanding use of individual and conduct data to profile its users, this is normally assembled certainly from inquiry history, scanning history; navigate information bookmarks, user archives, et cetera. It gives a modest system to the user to choose whether to customize an inquiry in UPS. This choice can be made before each runtime profiling to improve the soundness of the query items while evade the superfluous presentation of the profile [17].

Our broad tests exhibit the productivity and viability of our UPS structure. The system permitted users to determine personalized privacy necessities by means of the various levelled profiles. Likewise, UPS additionally performed online speculation on user profiles to secure the individual protection without trading off the pursuit quality [19]. It gives runtime profiling, which in actuality advances the personalization utility while regarding user's security necessities; takes into account customization of protection needs; and does not require iterative user association.

5.1. Performance Evaluation Parameters

The accompanying execution parameters are usually utilized as a part of security assurance system assessment. The current methodology is contrasted and proposed plan utilizing these assessment parameters [18]. The execution of the TC procedure can be measured by one or a greater amount of the accompanying techniques:

i) Recall

Category	No. of User Profiles	Recall	
		Existing	Proposed

20NG	412	60 Sec	50 Sec
Sports	300	80 Sec	70 Sec
Health	669	83 Sec	81 Sec
Society	442	84 Sec	83 Sec
Local News	254	91 Sec	89 Sec

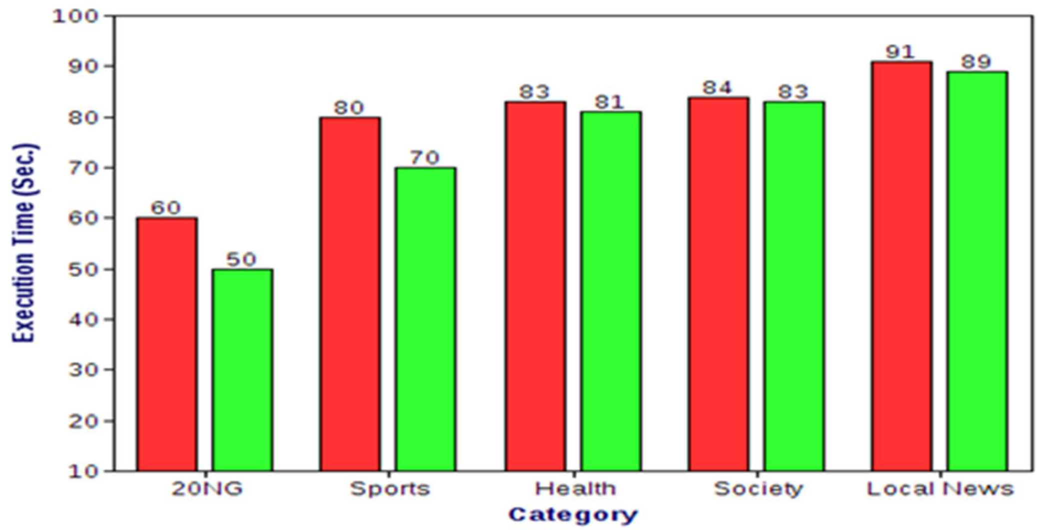

Fig 5.1: - Evaluation of Recall using GreedyIL Algorithm

In the above plotting, the red line represents the existing approach and the green line represents the GreedyIL for executing the user profile of the various users with various categories. The existing hierarchical link approach takes more time for extract the result from the dataset.

ii) Precision

Category	No. of User Profiles	Precision	
		Existing	Proposed
20NG	412	75%	98%
Sports	300	61%	96%
Health	669	58%	90%
Society	442	68%	91%
Local News	254	68%	73%

The precision represents the accuracy of retrieval or categorizing the data. In the above result, the red line represents the existing approach and the green line represents the GreedyIL for executing the user profile. Existing approach accuracy level is poor compare with the GreedyIL Approach of the proposed one.

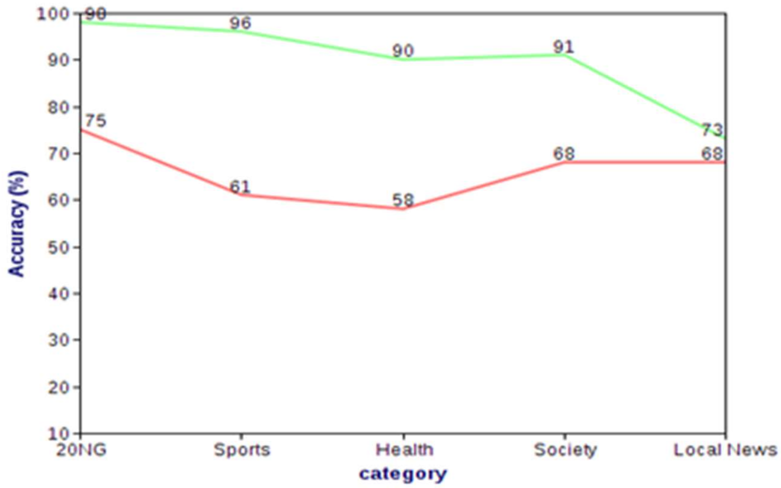

Fig 5.2: - Evaluation of Precision using GreedyIL Algorithm

6. CONCLUSION

The remarkable development of information on the Web has forced new challenges for the construction of effective search engines. This research work provides information on User customizable Privacy preserving Search framework-UPS for Personalized Web Search. UPS could potentially be adopted by any PWS that captures user profiles in a hierarchical taxonomy. The framework allowed users to specify personalized privacy requirements via the hierarchical profiles.

Another important conclusion we revealed in this research work is that personalization does not work equally well under various situations. We defined the click entropy to measure variation in information needs of users under a query. Experimental results showed that personalized Web search yields significant improvements over generic Web search for queries with a high click entropy. For the queries with a low click entropy, personalization methods performed similarly or even worse than generic search.

7. REFERENCES

[1] Wu, Z., Lu, C., Zhao, Y., Xie, J., Zou, D., & Su, X. (2021). The protection of user preference privacy in personalized information retrieval: challenges and overviews. Libri, 71(3), 227-237.

[2] Rodriguez-Garcia, M., Batet, M., Sánchez, D., & Viejo, A. (2021). Privacy protection of user profiles in online search via semantic randomization. Knowledge and Information Systems, 63, 2455-2477.

[3] El-Ansari, A., Beni-Hssane, A., Saadi, M., & El Fissaoui, M. (2021). PAPIR: privacy-aware personalized information retrieval. Journal of Ambient Intelligence and Humanized Computing, 12(10), 9891-9907.

[4] Deepak, G., Shwetha, B. N., Pushpa, C. N., Thriveni, J., & Venugopal, K. R. (2020). A hybridized semantic trust-based framework for personalized web page recommendation. International Journal of Computers and Applications, 42(8), 729-739.

[5] Sajjan, R. S., & Veer, S. A. (2019). Providing Privacy in Profile Based Personalized Web Search. concern, 12, 16.

[6] Alweshah, M., Aldabbas, Y., Abu-Salih, B., Oqeil, S., Hasan, H. S., Alkhalaileh, S., & Kassaymeh, S. (2023). Hybrid black widow optimization with iterated greedy algorithm for gene selection problems. Heliyon, 9(9).

[7] RAKESH, G., & Pratibha, M. (2016). Customized Personalized Web Search With Privacy Protection. International Research Journal of Engineering and Technology (IRJET), 3.

[8] Bhandare, M. S. K., & Kapse, A. S. (2021). Designing Improvisation In Privacy Protection For Personalized Web Search Using Deep Learning Approach. International Journal Of Engineering And Computer Science, 10(6), 25336-25346.

[9] Kumar, K., & Bright Keswani, D. V. J. (2020). Privacy Protection in Personalized Web Search using Obfuscation. International Journal, 8(4).

[10] Guo, G., Yang, T., & Liu, Y. (2018). Search engine based proper privacy protection scheme. IEEE Access, 6, 78551-78558.

[11] Sangamithra, B., Manjunath Swamy, B. E., & Sunil Kumar, M. (2022, November). Personalized Ranking Mechanism Using Yandex Dataset on Machine Learning Approaches. In Proceedings of the International Conference on Cognitive and Intelligent Computing: ICCIC 2021, Volume 1 (pp. 629-639). Singapore: Springer Nature Singapore.

[12] Zhu, R., Wang, M., Zhang, X., & Peng, X. (2023). Investigation of personal data protection mechanism based on blockchain technology. Scientific Reports, 13(1), 21918.

[13] Singh, N., & Singh, A. K. (2018). Data privacy protection mechanisms in cloud. Data Science and Engineering, 3(1), 24-39.

[14] Pramod, D. (2023). Privacy-preserving techniques in recommender systems: state-of-the-art review and future research agenda. Data Technologies and Applications, 57(1), 32-55.

[15] Khan, R. U., Ullah, M., Shafi, B., & Ihsan, I. (2023). A Survey on Performance Evaluation Mechanisms for Privacy-Aware Web Search Schemes. In Protecting User Privacy in Web Search Utilization (pp. 26-45). IGI Global.

[16] Petit, A. (2017). Introducing privacy in current web search engines (Doctoral dissertation, Université de Lyon; Universität Passau (Deutscheland)).

[17] RAYIDI, M., & CHOUDRI, S. (2016). Supporting Privacy Protection in Personalized Web Search.

[18] SRIKANTH, Y., & SHESHIKALA, M. (2016). Improvement of Privacy Protection in Personalized Web Search: using UPS Frame Work.

[19] Kumar, K., Kumar Gupta, M., Saxena, N., & Jaglan, V. (2022). Homomorphic Encryption Based Privacy Protection for Personalised Web Search. Indian Journal of Science and Technology, 15(8), 318-325.