Supporting Privacy Protection in Personalized Web Search with Secured User Profile

Archana R.Ukande[1], Nitin Shivale[2]

[1]Department of Computer Science, BSITR, Wagholi, Pune, India

[2]Assistant Professor Department of Computer Science, BSITR, Wagholi, Pune, India

Abstract: ***Web search engines (e.g. Google, Yahoo, Microsoft Live Search, etc.) are widely used to find certain data among a huge amount of information in a minimal amount of time. These useful tools also pose a privacy threat to the users: web search engines profile their users by storing and analyzing past searches submitted by them. For improving better search quality the String Similarity Match Algorithm (SSM Algorithm) can be implemented with the proposed. Current solutions propose new mechanisms that introduce a high cost in terms of computation and communication, to address this privacy threat. Personalized search is a promising way to improve the accuracy of web search, also it is attracting much attention recently. Effective personalized search requires collecting and aggregating user information, which often raises serious concerns of privacy infringement for many users. These concerns have become one of the main barriers for deploying personalized search applications, and privacy-preserving personalization is a great challenge. Adversaries are tried to resist in proposed system with the help of broader background knowledge (i.e. 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 can hide the user search results. With the help of this mechanism, privacy can be achieved.***

Keywords: Privacy protection; risk; profile; personalized web search; utility

1. Introduction

Novel protocol is proposed specially designed to protect the users' privacy in front of web search profiling. Adversaries are tried to resist in proposed system with the help of broader background knowledge i.e. knowing 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 can hide the user search results. In the existing System, Greedy DP and, Greedy IL algorithm it takes large computational and communication time.

For generalize the retrieved data by using the background knowledge [1], [5], [3], [7] through this adversaries can be avoided. Privacy protection in publishing transaction data is an important problem. A key feature of data transaction is the extreme scarcity, which renders any single technique ineffective in anonymizing such data. Among recent works, some suffer from performance drawbacks, some incur high information loss and some result in data hard to interpret. This approach proposes to integrate generalization and compression to reduce information loss. However, the integration is non-trivial. Novel techniques are proposed to address the efficiency and scalability challenges. A few previous studies [8], [9] suggest that people are willing to compromise privacy if the personalization by supplying user profile to the search engine yields better search quality

2. Literature Review

A. Privacy Protection In Personalized Search

In privacy protection, analytically observe the concern of privacy preservation in personalized search [10]. Here discriminate and describe four levels of privacy protection, and analyze numerous software architectures for personalized search. It shows that client-side personalization has advantages over the existing server-side personalized search services in preserving privacy in this situation; personalized web search cannot be done at the individual user level, but is possible at the group level. This may reduce the effectiveness of personalization because a group's information need explanation is used to model an individual user's information need.

However, if the group is appropriately constructed so that people with similar interests are grouped together, it has much richer user information to offset the sparse explanation of individual user information requirements. Thus the search performance may essentially be improved because of the availability of more information from the group profile [11] and [12]. In this circumstance, personalized web search cannot be done at the distinct user level, but is possible at the group level. This may reduce the effectiveness of personalization because a group's information need description is used to model an individual user's information need. However, if the group is properly constructed so that people with comparable interests are grouped together, it may have much richer user information to offset the sparse explanation of distinct user information needs. Thus the search performance may really be better because of the accessibility of more information from the group profile

a. Advantages

1. The architecture has an advantage of allowing for the use of a search engine's internal resources.
2. It improves the accuracy of web search.

b. Disadvantages

1. It does not fully protect user privacy.
2. They were not discussed different levels of privacy protection provided by search engines depending on a user's preference for the tradeoff between the privacy concern and the improved search service quality.

B. Implicit User Modeling For Personalized Search

In implicit user modeling for personalized search [2], explicated how to infer a user's interest from the user's search context and practice the conditional implied user model for personalized search. A decision speculative basis and develop methods for implicit user exhibiting in information retrieval. They developed an intelligent client-side web search agent (UCAIR) that can achieve eager implicit feedback, e.g., query development established on prior queries and instant result re-ranking established on search show that search agent can progress search accuracy over the popular Google search engine. In this paper, described how to make and update a user model based on the instant search context and implicit feedback information and use the model to improve the accuracy of ad hoc retrieval.

In order to extremely benefit the user of a retrieval system through implicit user modeling, offered to perform "eager implicit feedback". Those is, as soon as experimental any new piece of evidence from the user, and update the system's certainty about the user's information need and respond with improved retrieval outcomes based on the updated user model. A decision-theoretic basis for enhancing interactive information retrieval based on eager user model updating, in which the system replies to each achievement of the user by choosing a system exploit to enhance an efficacy function.

In a traditional retrieval model, the retrieval problem is often to match a query with documents and rank documents giving to their relevance values. As a result, the whole retrieval progression is a simple independent cycle of "query" and "result display". In the planned new recovery model, the user's search circumstance shows a significant role and the conditional implicit user typical is exploited directly to benefit the user. The novel retrieval model is thus essentially diverse from the traditional pattern, and is inherently more general.

a. Advantages

1. It expands search accuracy over the popular Google search engine.
2. The developed search cause thus can advance existing web search performance without any additional effort from the user.

b. Disadvantages

1. The search agent does not have control of the retrieval algorithm.
2. It should displayed summaries, but the document content is actually not.

C. IR Evaluation Method

IR evaluation method [4] is used for retrieving highly relevant documents. This paper proposes assessment approaches established on the use of non-dichotomous relevance judgments in IR investigates. It is maintained that evaluation methods should credit IR methods for their ability to retrieve highly relevant documents. This is desirable from the user point of view in modem large IR environments. The proposed methods are a novel application of P-R curves and average precision computations based on separate recall bases for documents of different degrees of relevance, and two novel measures cumulative computing gain the user obtains by examining the retrieval result up to a given ranked position. Then demonstrate the use of these evaluation methods in a case study on the effectiveness of query types, based on combination of query structures and expansion, in retrieving documents of various degrees of relevance. Test was run with a best match retrieval system (In- Query I) in a text database consisting of newspaper articles. Results indicate that the tested strong query structures are most effective in retrieving relevant documents. The differences between query types are statistically significant and practically essential. More generally, the novel evaluation methods and the case demonstrate that non-dichotomous relevance assessments are applicable in IR experiments and allow harder testing of IR methods.

a) Advantages

- The P-R curves demonstrate that the good performance of the expanded structured query types.
- The best performance overall was achieved with expanded, facet structured queries.

b) Disadvantages

- The DCV-based precision recall curves are better but still do not make the value gained by ranked position explicit.
- The RHL alone is not sufficient as a performance measure.

D. Automatic Identification of User Interest

Automatic identification of user interest is done for personalized search [6]. Here a framework is proposed to investigate the problem of personalizing web search based on user s' past search histories without user efforts. Proposed a user model to formalize user's interests on web -pages and correlate them with user's clicks on search results .Based on this described correlation an intuitive algorithm to actually learn user's interests. Two different methods are proposed, based on different assumptions on user behaviors, to rank search results based on the user's interests learned.

The both theoretical and real-life experiments to evaluate our approach, In the theoretical experiment, found that for a reasonably small user search trace, the user interests estimated by our learning algorithm can be used to pretty accurately predict view based on importance of web pages, which is expressed by Personalized PageRank, showing that our method is effective and easily applicable to real-life search engines. In the real-life, we applied our method to learn the interests of 10 subjects contacted. The results showed that, on average, our method per formed between 25%–33% better than Topic-Sensitive PageRank, which turned out to be much better than PageRank.

a. Advantages

1. The experiments show that user's preferences can be learned accurately even from personalized search based on user preference and small history data yields significant improvements over the best existing ranking mechanism in the literature.
2. PageRank is more Relevant than the global PageRank.

b. Disadvantages

1. It is not more users -specific information into consideration. The difficulties in doing this include integration of different information sources, modeling of the correlation between various information and the user's search behaviors, and efficiency concerns.
2. It does not design more sophisticated learning and ranking algorithms to further improve the performance of our system.

3. System Architecture

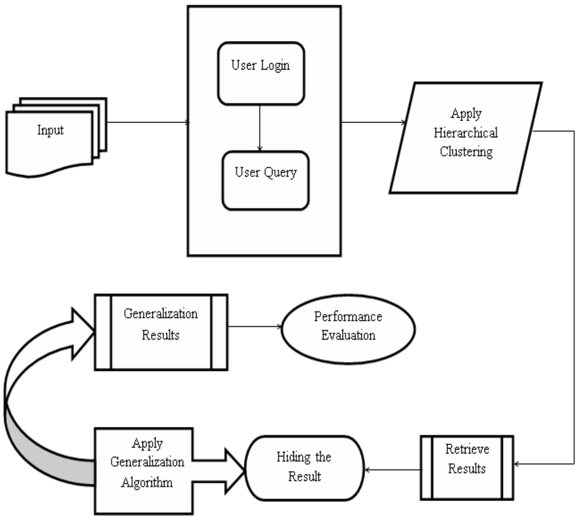


Figure 1: System Architecture

4. Modules

A. Dataset preprocessing

Most commonly a data set corresponds to the contents of a single statistical data matrix, or a single database table, where every column of the table represents a particular variable, and each row corresponds to a given member of the data set in question. The data set lists values for each of the variables, such as height and weight of an object, for each member of the data set. Each value is known as a datum. The data set may comprise data for one or more members, corresponding to the number of rows. This module, choose input dataset. Chosen dataset has been loaded into the database. After loading the dataset into the database, we can view the dataset. By using the string matching algorithm we filter out unwanted values in the dataset and it has been preprocessed and store into the database.

B. User Login

This is for user login page. In this module, users are entered by using the unique id and password. In this module, users are entered after registering. After registering each user has unique id. After login, user posts some queries which are based on our dataset which is loaded into the database.

C. Query Searching and Search Results Retrieval

In this module, user submits query. Based on the query, relevant results has been displayed and also based on the submitted query some history results are displayed. Based on the query and already posted queries, we can calculate the similarity values between them. In that three types of similarity values has been estimated. From that, the result is retrieved which is based on the high relevant results by using the minimum range of similar values.

D. Estimate Relevant Results

In this module, user posts query and sub query also. Based on the query and sub query, estimate the results based n string matching. Based on the relevant results and total number of data's in the dataset, we can estimate the support values.

E. Retrieve user profile in privacy manner

In this module, adversaries to mine the history results means, only query time has been displayed. In this, other information such as query, query results, username are not displayed by using the background knowledge. First we generalize the table, and then suppress the values based on the generalized table. Generalized values are stored in the history results. When the adversaries' views the history result means, they can only view the generalized results. Finally, the performance can be evaluated by using the parameter such as time, cost and communicational and computational cost.

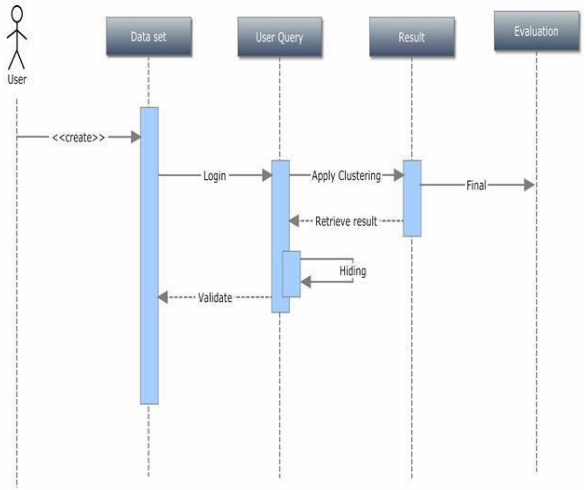


Figure 2: Sequence Diagram

5. Existing System

A. Methodology of Existing System

In the Existing Work, a client-side privacy protection framework called UPS for personalized web search was proposed. UPS could theoretically be adopted by any PWS that captures user profiles in a hierarchical taxonomy. The context allowed users to stipulate customized privacy requirements via the hierarchical profiles. In addition, UPS also performed online generalization on user profiles to protect the personal privacy without compromising the search quality. In this they proposed two greedy algorithms, namely GreedyDP and GreedyIL, for the online generalization. In this for query mapping process it has various steps to compute the relevant items.

Most works on anonymization focus on relational data where every record has the same number of sensitive attributes. There are a few works taking the first step towards anonymizing set-valued or transactional data where sensitive items or values are not clearly defined. While they could be potentially applied to user profiles, one main limitation is that they either assume a predefined set of sensitive items that need to be protected, which are hard to done in the web context in practice, or only guarantee the anonymity of a user but do not prevent the linking attack between a user and a potentially sensitive item.

Another approach to provide privacy in web searches is the use of a general purpose anonymous web browsing mechanism. Simple mechanisms to achieve a certain level of anonymity in web browsing include: (i) the use of proxies; or (ii) the use of dynamic IP addresses.

B. Disadvantages:

It has demonstrated the ineffectiveness or privacy risks of naive anonymization schemes. The utility data is limited to statistical information and it is not clear how it can be used for personalized web search. For retrieving the user query results, it takes high computational and communication time and also cost. Proxies don't solve the privacy problem. This solution moves the privacy threat from the web search engine to the proxies themselves. A proxy will pre-vent the web search engine from profiling the users, but the proxy will be able to profile them instead. The

Renewal policy of the dynamic IP address is not controlled by the user but the network operator.

6. Proposed System

Web search engines are widely used to find data from huge amount of information in a minimal amount of time. However, these tools also pose a privacy threat to the users, web search engines profiles their users by storing and analyzing past searches submitted by them. In the proposed system, we can implement the clustering algorithms for improving the better search quality results. It is retrieved by using the String Similarity Match (SSM) algorithm. To address this privacy threat, current solutions propose new mechanisms that introduce a low cost in terms of computation and communication. In this paper we present a novel protocol specially designed to protect the users' privacy in front of web search profiling. In this we try to resist adversaries with broader background knowledge i.e. knowing relationship among topics. Richer relationship, means generalize the user profile results by using the background knowledge which is going to store in history. Through this user search result can be abstracted. In the Existing System, Greedy DP and Greedy IL algorithm, it takes large computational and communication time.

A. Advantages:

1. It achieves better search results.
2. It achieves the privacy results when applying the background knowledge to the user profiling results.
3. It has less computational time and communicational time.
4. It achieves better accuracy when compared with the Existing Works.

B. Methodology of Proposed System

In the proposed system, propose Approximate String Similarity Match Algorithm (SSM Algorithm). For Client post query q to the server, server retrieves the query results QT to the client. Results have been extracted by using the clustering algorithm String matching algorithm.

One possible definition of the approximate string matching problem is the following:

1. Given a pattern string $P = p_1 p_2 ... p_m$
2. Text string $T = t_1 t_2 \ldots t_n$,
3. Find a substring $T_{j',j} = t_{j'} \ldots t_j$ in T, which, of all substrings of T, has the smallest edit distance to the pattern P.

C. Explanation

Step1: Detecting & removal of unwanted symbols

Step2 Compute similarity calculation for user given word and word in database

Step3: In that similarity calculation, extract the features in the dataset.

Step4: Then estimate the ASCII difference for user given word and words in database

Step5: The estimate the similarity values.

Step6: Then retrieve the most relevant documents based on the similar values

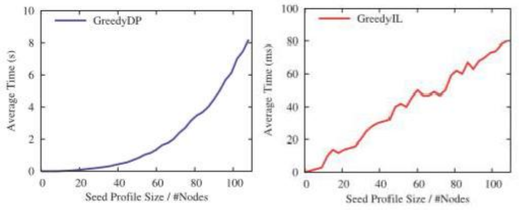
7. Results

A. The GreedyIL Algorithm

The GreedyIL algorithm improves the efficiency of the generalization. Prune-leaf operation reduces the discriminating power of the profile. In other words, the DP displays monotonicity by prune-leaf

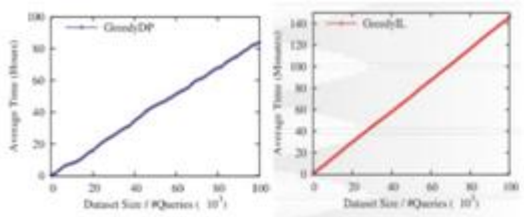
B. The GreedyDP Algorithm

The GreedyDP works in a bottom up manner. It removes all the personalized information from the profile. This causes significant memory requirements and computational cost.



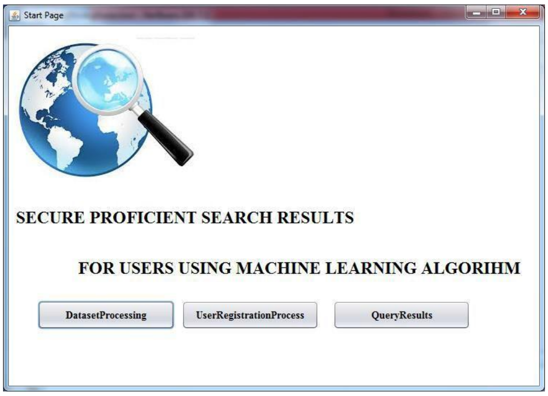
(a) Results of GreedyDP (b) Results of GreedyIL

Scalability by varying profile size



(a) Results of GreedyDP (b) Results of GreedyIL

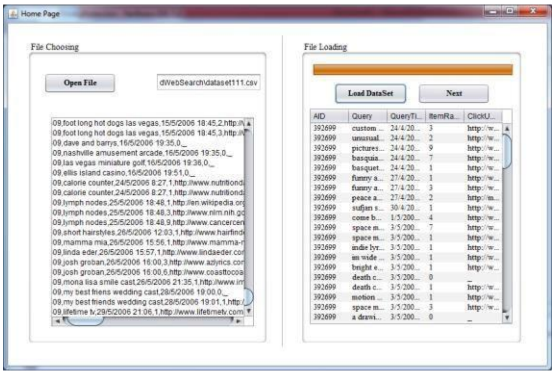
Scalability by varying data set size.



SECURE PROFICIENT SEARCH RESULTS

FOR USERS USING MACHINE LEARNING ALGORIHM

Figure 3: Home Page

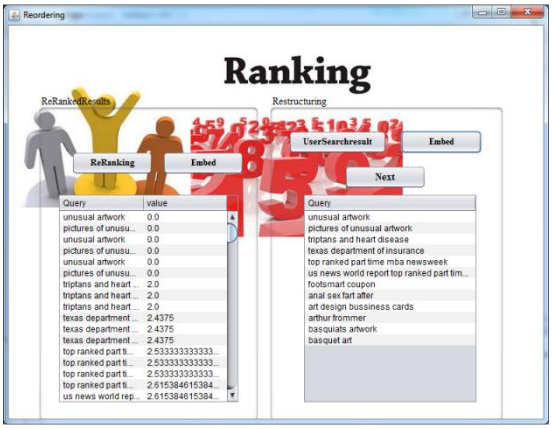


dWebSearch\dataset111.csv

```
09,foot long hot dogs las vegas,15/5/2006 18:45,2,http://\
09,foot long hot dogs las vegas,15/5/2006 18:45,3,http://\
09,dave and barrys,16/5/2006 19:35,0,_
09,nashville amusement arcade,16/5/2006 19:35,0,_
09,las vegas miniature golf,16/5/2006 19:36,0,_
09,ellis island casino,16/5/2006 19:51,0,_
09,calorie counter,24/5/2006 8:27,1,http://www.nutritiond
09,calorie counter,24/5/2006 8:27,1,http://www.nutritiond
09,lymph nodes,25/5/2006 18:48,1,http://en.wikipedia.org
09,lymph nodes,25/5/2006 18:48,3,http://www.nlm.nih.go
09,lymph nodes,25/5/2006 18:48,9,http://www.cancercen
09,short hairstyles,26/5/2006 12:03,1,http://www.hairfinde
09,mamma mia,26/5/2006 15:56,1,http://www.mamma-r
09,linda eder,26/5/2006 15:57,1,http://www.lindaeder.com
09,josh groban,26/5/2006 16:00,3,http://www.azlyrics.con
09,josh groban,26/5/2006 16:00,6,http://www.coasttocoa
09,mona lisa smile cast,26/5/2006 21:35,1,http://www.im
09,my best friens wedding cast,28/5/2006 19:00,0,_
09,my best friends wedding cast,28/5/2006 19:01,1,http:/
09,lifetime tv,29/5/2006 21:06,1,http://www.lifetimetv.com
```

AID	Query	QueryTi...	ItemRa...	ClickU...
392699	custom ...	24/4/20...	3	http://w...
392699	unusual...	24/4/20...	2	http://w...
392699	pictures...	24/4/20...	9	http://w...
392699	basquia...	24/4/20...	7	http://w...
392699	basquet...	24/4/20...	1	http://w...
392699	funny a...	27/4/20...	1	http://w...
392699	funny a...	27/4/20...	3	http://w...
392699	peace a...	27/4/20...	2	http://m...
392699	sufjan s...	30/4/20...	1	http://w...
392699	come b...	1/5/200...	4	http://w...
392699	space m...	3/5/200...	7	http://w...
392699	space m...	3/5/200...	1	http://w...
392699	indie lyr...	3/5/200...	1	http://w...
392699	im wide ...	3/5/200...	1	http://w...
392699	bright e...	3/5/200...	1	http://w...
392699	death c...	3/5/200...	0	_
392699	death c...	3/5/200...	1	http://w...
392699	motion ...	3/5/200...	1	http://w...
392699	space m...	3/5/200...	3	http://w...
392699	a drawi...	3/5/200...	0	_

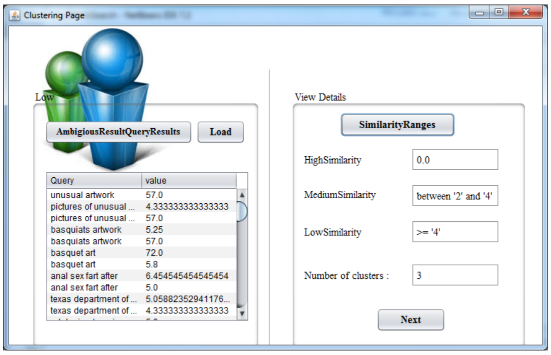
Figure 4: After loading data set



Query	value
unusual artwork	0.0
pictures of unusu...	0.0
unusual artwork	0.0
pictures of unusu...	0.0
unusual artwork	0.0
pictures of unusu...	0.0
triptans and heart ...	2.0
triptans and heart ...	2.0
triptans and heart ...	2.0
texas department ...	2.4375
texas department ...	2.4375
texas department ...	2.4375
top ranked part ti...	2.533333333333...
top ranked part ti...	2.533333333333...
top ranked part ti...	2.533333333333...
top ranked part ti...	2.615384615384...
us news world rep...	2.615384615384...

Query
unusual artwork
pictures of unusual artwork
triptans and heart disease
texas department of insurance
top ranked part time mba newsweek
us news world report top ranked part tim...
footsmart coupon
anal sex fart after
art design bussiness cards
arthur frommer
basquiats artwork
basquet art

Figure 4: Ranking of result



Query	value
unusual artwork	57.0
pictures of unusual ...	4.333333333333333
pictures of unusual ...	57.0
basquiats artwork	5.25
basquiats artwork	57.0
basquet art	72.0
basquet art	5.8
anal sex fart after	6.454545454545454
anal sex fart after	5.0
texas department of ...	5.05882352941176...
texas department of ...	4.333333333333333

HighSimilarity: 0.0

MediumSimilarity: between '2' and '4'

LowSimilarity: >= '4'

Number of clusters : 3

Figure 4: Similarity of suearch result

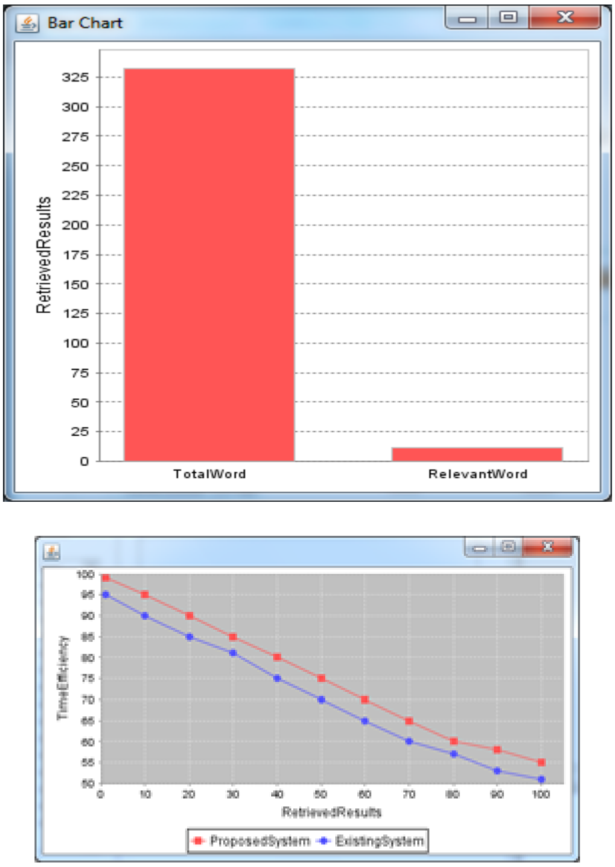


Figure 5: Result

8. Conclusion and Future Enhancement

For generalizing retrieve data by using the background knowledge. Through this we can resist the adversaries. Privacy protection in publishing transaction data is an important problem. A key feature of transaction data is the extreme scarcity, which renders any single technique ineffective in anonymizing such data. Among recent works, some incur high information loss, some result in data hard to interpret, and some suffer from performance drawbacks. This paper proposes to integrate generalization and compression to reduce information loss. The integration is non-trivial. We propose novel techniques to address the efficiency and scalability challenges.

Our proposed system gives better quality results and gives more efficiency. Privacy is too good when compared with the Existing system. In the Existing System, only generalization technique is used. Our String matching algorithm gives more accuracy when compared with the Greedy IL algorithm. Generalization and suppression technique achieves better privacy when compared with the existing system

References

[1] Z. Dou, R. Song, and J.-R. Wen, “A Large-Scale Evaluation and Analysis of Personalized Search Strategies,” Proc. Int’l Conf. World Wide Web (WWW), pp. 581-590,2007.

[2] X. Shen, B. Tan, and C. Zhai, "Implicit User Modeling for Personalized Search," Proc. 14th ACM Int'l Conf. Information and Knowledge Management (CIKM), 2005

[3] M. Spertta and S. Gach, "Personalizing Search Based on User Search Histories," Proc. IEEE/WIC/ACM Int'l Conf. Web Intelligence (WI), 2005.

[4] K. Ja ¨rvelin and J. Keka ¨la ¨inen, "IR Evaluation Methods for Retrieving Highly Relevant Documents," Proc. 23rd Ann. Int'l ACM SIGIR Conf. Research and Development Information Retrieval (SIGIR), pp. 41-48, 2000

[5] P.A. Chirita, W. Nejdl, R. Paiu, and C. Kohlschutter, "Using ODP Metadata to Personalize Search," Proc. 28th Ann. Int'l ACM SIGIR Conf. Research and Development Information Retrieval (SIGIR), 2005.

[6] F. Qiu and J. Cho, "Automatic Identification of User Interest for Personalized Search," Proc. 15th Int'l Conf. World Wide Web (WWW), pp. 727-736, 2006

[7] A. Pretschner and S. Gauch, "Ontology-Based Personalized Search and Browsing," Proc. IEEE 11th Int'l Conf. Tools with Artificial Intelligence (ICTAI '99), 1999.

[8] Y. Xu, K. Wang, B. Zhang, and Z. Chen, "Privacy-Enhancing Personalized Web Search," Proc. 16th Int'l Conf. World Wide Web (WWW), pp. 591-600, 2007.pp. 1497-1500, 2009

[9] A. Krause and E. Horvitz, "A Utility-Theoretic Approach to Privacy in Online Services," J. Artificial Intelligence Research, vol. 39, pp. 633-662, 2010.

[10] X. Shen, B. Tan, and C. Zhai, "Privacy Protection in Personalized Search," SIGIR Forum, vol. 41, no. 1, pp. 4-17, 2007

[11] Y. Xu, K. Wang, G. Yang, and A.W.-C. Fu, "Online Anonymity for Personalized Web Services," Proc. 18th ACM Conf. Information and Knowledge Management (CIKM), pp. 1497-1500, 2009.

[12] Y. Zhu, L. Xiong, and C. Verdery, "Anonymizing User Profiles for Personalized Web Search," Proc. 19th Int'l Conf. World Wide Web (WWW), pp. 1225-1226, 2010.