ONTOLOGICAL USER PROFILES FOR EFFECTIVE PRIVACY PROTECTION IN PERSONALIZED WEB SEARCH

*Dr.N. Umadevi, **V.Ramya

*Department of computer science and information technology,
Sri Jayendra Saraswathy Maha Vidyala College of Arts & Science
Coimbatore
** Department of computer science and information technology
Sri Jayendra Saraswathy Maha Vidyala College of Arts & Science
Coimbatore

ABSTRACT

Personalized web search retrieval is the most concerned application in the search engine based content retrieval scenarios where the users are required to get their most wanted results as per their requirements. Tracking user's preferences and retrieving contents based on their preference needs is the complex scenario. The main objective of this work is to retrieve the contents from the web servers based on the individual user preferences by constructing their personalized profile in the anonymized manner. The profile is represented by using the ontological concept where the different meaning of sentences can be extracted. Ontology based user profile construction is integrated with the existing concepts called the privacy profile construction in the existing work. This proposed research methodology attempts to construct the triplets of the user query, so that the exact meaning of their requirement can be identified easily. With the help of triplets, the original meanings of the sentences are identified and based on which profile is reconstructed and represented in the ontological format. This method will gather the user privacy requirements from them which would be anonymized in the future. The experimental tests conducted in java simulation environment which is integrated with the online server for retrieving the user preferred web sites. The experimental tests were conducted to prove the improvement over the proposed methodology in terms of performance measures called the precision and recall to identify which methodology can retrieve the contents in the more efficient manner. The comparison were plotted in the graphical representation and the finally concluded that the proposed methodology provides better result than the existing research works. Findings of these work demonstrates that the proposed methodology provides better result than the previous research works in terms improved accurate retrieval of contents.

Keywords:Personalized web search, Ontology construction, Triplet extraction

INTRODUCTION

Personalized web search is a process of retrieving the contents from the web servers based on the user requests submitted by them. Most of the application started to enable the search retrieval process in the task to satisfy the user needs. For example Google is one of the most popular search engines in the real world environment which enables the personalized web search with the help of gmail product.

1

The gmail product will track the every information about the corresponding user and will store it in the database. In case of another search related to previous topic gmail would highlight the contents that are accessed by the users previously. This application enables users to search in the personalized manner and the efficient constructions of users are done optimally. Most of the users prefer to retrieve the contents based on their interest which might require their personal identity information to be revealed to the web servers, so that they can keep track of all user activities.

The privacy becomes the biggest issue in the online web search retrieval application where the personal information of users might get violated in case of providing their details to third party server. There may be possibility of violation of the user's sensitive personal information which needs to be resolved in the efficient manner. Various research work has been proposed which concentrates on providing the anonymity for the users who intends to retrieve their most interested pages.

The main contribution of this work is to construct the ontological based user profiles in terms of the user browsing behaviour with the consideration of their privacy requirements. Initially user profiles would be constructed in terms of their browsing behaviour. The pages and links that are visited by the users for the corresponding key word would be retrieved and will calculate the relationship between those users submitted key word and the retrieved contents. This relationship is established semantically by understanding conceptual relationship between the keyword and the pages that are retrieved. This is done by extracting the triplets by using which formation of the sentences can be learned in terms of user submitted key words. The flow of the project is given as:

- Gathering user queries
- Construct the user profiles based on ontological concept
- Achieve privacy level of user profiles with the consideration of the user requirements
- Retrieve the matched documents to the users
- Update the ontological user profile

By following these steps, privacy preserved personalized web search is enabled in this proposed research which is discussed detailed in the following sub sections.

The organization of this research work is given as follows: In this section detailed description about the introduction of the research work is given. In section 2, various research works has been discussed in the detailed manner which was conducted in the previous work. In section 3, proposed work of this research methodology is discussed in detail. In section 4, performance evaluation tests were conducted with the consideration of the various performance measures. In section 5, overall research of this work is concluded.

RELATED WORKS

In this section various research works have been discussed in the detailed manner to know the merits and demerits of the existing research works and the problem scenario of our proposed research methodologies. The discussion previous works are given as follows:

Jaime Teevan [1] et al introduced the automated searching and building user profiles for the web search users in order to provide them convenient and the satisfiable environment which can retrieve the contents as per their requirements. This is done by enabling the monitoring process over the user browsing behaviour through which the user personal interests can be obtained. This work also follows

a user feedback system through which users personal interests can be extracted and processed well. The feedback based monitoring system provides a convenient way for the users to submit their sensitive information, thus the privacy of the users could be obtained efficiently.

Kazunari Sugiyama [2], Feng Qiu [3] and Shen [7] et al introduced another approach which is different than the previous work in which user profile construction is done in the adaptive manner. This methodology attempts to construct the personalized user profile without the involvement of users. To do so, collaborative filtering approach is introduced which attempts to monitor the user behaviour in terms of various user profile details. This work is based on the pure browsing behaviour of the users where the user transaction history would be gathered from the different users in terms of the unique id of the users. After gathering these details users can obtain the valid personal information about them in terms of the different construction profiles where their browsing histories are stored securely. Whenever the user requesting for the new content, the web server will compare the queries with the personalized profile based on which final result would be decided.

Yabo Xu [4] and Spertta [6] et al provides a new way for hiding the personal details of the users present in the environment in case of presence of the sensitive information in the document. This methodology provides meaningful application for the users who don't want to reveal their personal sensitive information. This method would organize the user's sensitive and personal information in the hierarchical manner. After representing the user's personal information in the hierarchical manner their sensitive information which they don't want to reveal to the outside world would be gathering. The specialization approach will be applied on the hierarchical user profiles thus the anonymity level of the users profile is increased.

Alexander Pretschner [5] et al introduced an novel way for construction of user profiles in terms of users various requests which is used to retrieve the real meaning of the user interest in the semantic way. The semantic behaviour of the user profiles are constructed in terms of the various interest score value of the users. Thisapproach provides a better way for construction of user profiles with the concern of their various submission details. The ontological user profiles provides an convenient way for the users to obtain the more similar web contents that matches with the similarity score values.

The working scenario of the various previous research works has been given and described detail in this section. The working scenario of the proposed methodology is described in the following section 3.

ONTOLOGICAL BASED PRIVACY PRESERVED PERSONALIZED USER PROFILE CONSTRCUTION

The proposed research methodology of this work concentrates on retrieving web search contents in terms of their user interest level. This work attempts to retrieve the contents in terms of the various search interest of the users which is present in the user profiles in the anonymized way. The anonymization of the user profiles are done with the consideration of the sensitive information that are represented by the users. User profiles are constructed by adapting the ontological behaviour based on which one can obtain the semantic meaning of the users that are present in the system.
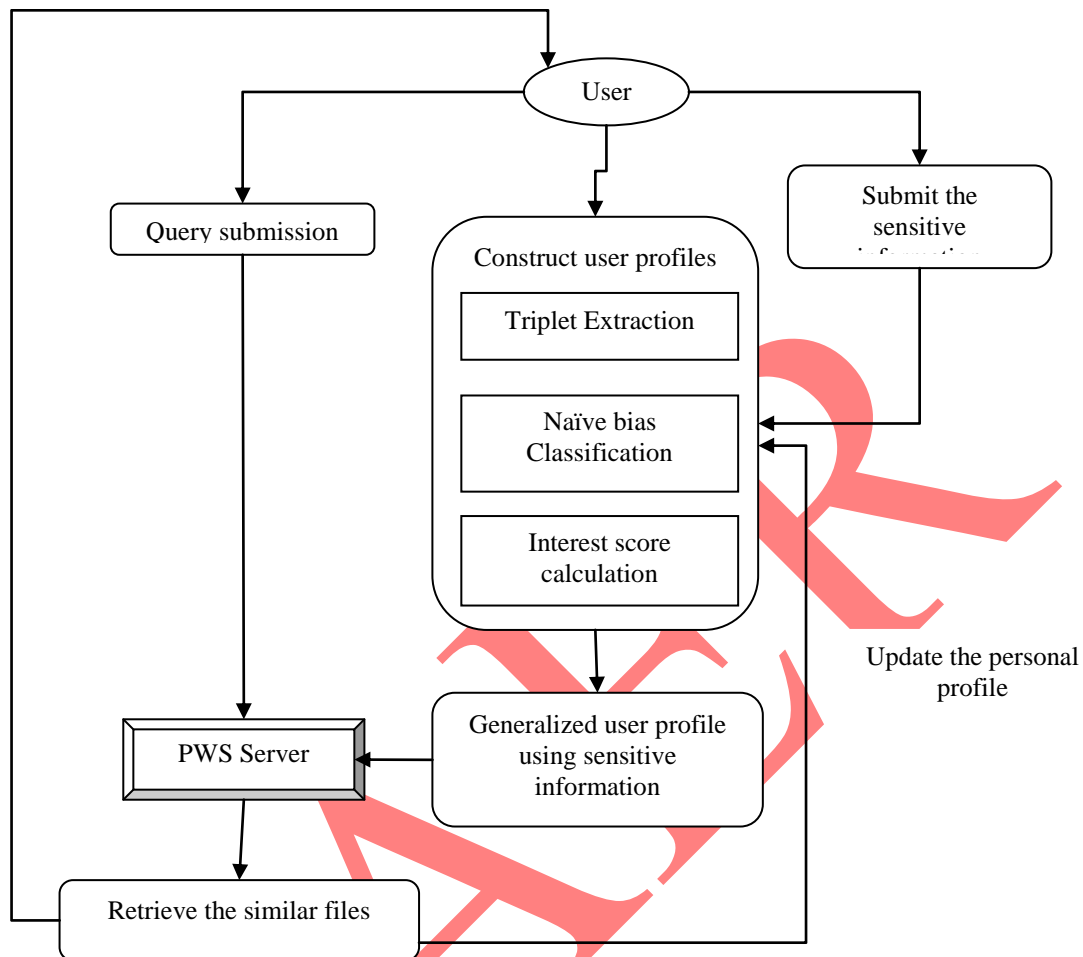
Figure 1. Overall Flow of the work

The above diagram depicts the overall processing of the proposed research methodology is given. Description of this flow of research is discussed detailed in the following sub sections.

A. *Ontological Based User Profile Construction*

Ontology is a knowledge system which is used to represent the unstructured data in the structured way with the meaningful information. Ontological user profile construction make ease of users to generalize their sensitive information and provide a convenient way for web server to retrieve the content that most suits the user requirements. The proposed ontology construction methodology is based on the construction of the user profiles with the consideration of the sematic relation exists between the user required matches documents. It is done by finding similar attributes present in the documents and retrieves the association relationship present between the different terms present in the document.

98

Initially, all the documents that are present in the web server which matches the user query and the generalized web profile would be retrieved. After retrieving of documents, sentences present in the each document would be separated to identify the association rule present among them. After separation and extraction of individual sentences that are present in the system, parse tree would be constructed based on the phrases present in each sentences. Phrases of sentences are obtained with the knowledge of hand coded rules which are inbuilt available in the web sites.

Hand coded rules are defined as the some underlying concepts that are used to understand the basic structures and the underlying concepts that are evolved in the given input documents with the help of computer programming languages. It provides an convenient way for the users to understand the basic structure of the given input document within less period of time. Some of the hand code rules are listed as follows:

Table 1. Hand coded Rules

Rule No	Rule Components		Example attributes
	Attribute lex pattern	RHS of parent VP	
1	VBZ	NP, S	"is"
2	VBZ+DT	NP,S	"is a"
3	VBD	NP,S	"lived"
4	VBZ+IN	PP	"lives in"
5	VBD+IN	PP	"lived in"
6	VBG+IN	PP	"living in"
7	VBN+TO	PP	"used to"
8	VBN+IN	PP	"used for"
9	VB+RP	NP	"carry out"
10	VBP	ADJP	"are"
11	VBP	NP	"are"
12	VBP	ADVP	"drive west"

By comparing the sentences with these hand coded rules, triplets would be extracted for every sentence which consists of the attributes called the (predicate. Target, key word).

The relationship between every sentence triplets would be identified with the help of WordNet tool. It will compare the triplets with the WordNet tool based on the sentenced would classified. WordNet is one kind of lexical database by using which the syntactic meaning of the various kinds of sentences can be learned in terms of their grammatical representation. The actual meaning of the triplets can be identified by comparing it with the available WordNet which will parse the triplets and will compare it with the grammatical meaning based on which meaning of the sentences can be learnt easily.

The classification will result the types and semantic meaning of every sentences. In addition to that association rule between every sentence would also be identified. WordNet is an online knowledge source which would give the semantic and syntactic meaning of every sentence. This information would be used to construct the ontological profile where the user interest values can be identified in the efficient manner.

99

B. *Generalization Of User Profile*

After constructing the user profile, generalization would be done to preserve the sensitive information of the users which would prevent the third party server from accessing it. It will gather the input sensitive requirements from the users which would be forward to the user's local servers. GreedyIL algorithm would generalize the sensitive information present in the user profiles. The algorithm is given as follows:

ALGORTIHM 1: GreedyIL (H, q, δ)

Input: Seed Profile G_0; Query q; Privacy threshold δ

Output: Generalized profile G* satisfying δ-Risk

Let Q be the IL priority queue of prune-leaf decisions;

i be the iteration index, initialized to 0;

// online decisions whether personalize queue 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 ϵ 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 → 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*

The above algorithm will generalize the user profile by hiding the sensitive information of users in those profiles.

C. Retrieval Of Contents

After generalizing the user profile, both the user query and the generalized user profiles would be sent to personalized web server where similarity matching would be done. The web content which matches most with both query and the profile would be retrieved as output to the users. The similarity retrieval leads to an accurate retrieval of contents based on which one can improve the system efficiency.

After retrieval process, the current browsing behaviour would be monitored which is then send to the personalized web server to update the user profiles in which additional web browsing behaviour would be added.

NUMERICAL RESUTLS

In this section, performance evaluation is done to prove the effectiveness of the proposed methodology in terms of the average time and the average precision rate. The findings clearly demonstrate that the proposed methodology improved in terms of both time taken to retrieve the user search content and the accuracy. The comparison is shown as follows:

A. Average Time

The time taken to retrieve the contents is described in this section. The total time taken between the time of user query submission and the time of search retrieval is calculated for both existing approach and the proposed approach and compared in the following graph. The actual values that are obtained while processing the submitted queries are indicated in the table 1.

Table 1. Average Time Comparison values

Number of Queries	Average Time Comparison	
	Privacy Preserved Query Retrieval	Ontological Based Privacy Enhanced Query Retrieval
5	63	52
10	125	100
15	197	152
20	267	200
25	308	262

101

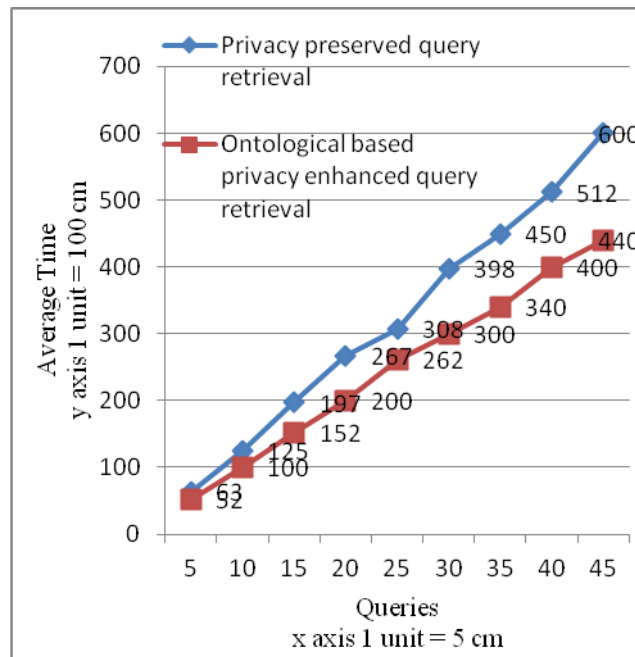| 30	398	300
35	450	340
40	512	400
45	600	440

Figure 1. Average Time Comparison

In the above graph, the time value is compared against the existing approach and the proposed approach. In the x axis numbers of queries gathered from the users are taken. And in y axis, average time is taken. From this graph, it can be proved that the time taken by the proposed methodology is less than the existing approach.

B. Average Precision

In information retrieval contexts, precision and recall are defined in terms of a set of retrieved documents (e.g. the list of documents produced by a web search engine for a query) and a set of relevant documents (e.g. the list of all documents on the internet that are relevant for a certain topic), of relevance.

In the field of information retrieval, precision is the fraction of retrieved documents that are relevant to the find:

$$\text{Precision} = \frac{|\{\text{number of Similar documents}\}| \cap \{\text{Number of retrieved documents}\}|}{|\{\text{Total number of retrieved documents}\}|}$$

Precision is calculated to know how well the proposed system can extract the similar contents in terms of the users personalized profile consideration. This is calculated by taking the ratio between the total number similar documents that are obtained from the total number of documents that are retrieved as output.

The actual precision values that are obtained while processing the submitted queries are indicated in the table 2.

Table 1. Precision Comparison values

102

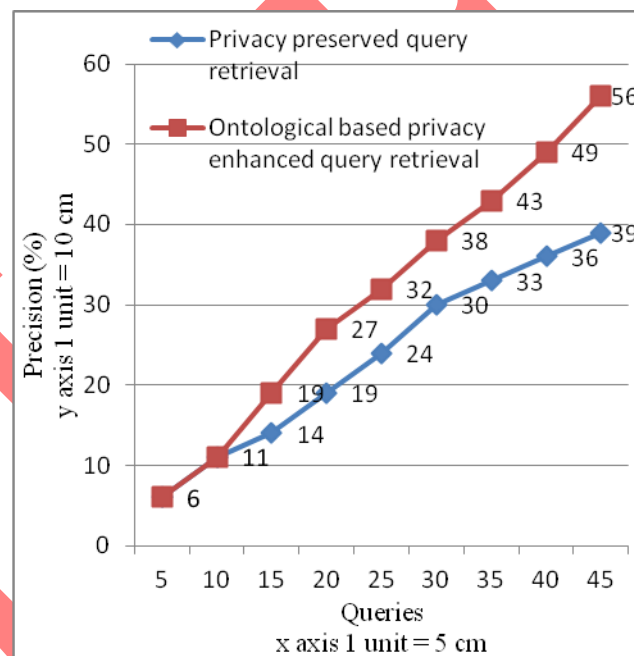| Number of Queries	Precision Comparison	
	Privacy Preserved Query Retrieval	Ontological Based Privacy Enhanced Query Retrieval
5	6	6
10	11	11
15	14	19
20	19	27
25	24	32
30	30	38
35	33	43
40	36	49
45	39	56

Figure 2. Average Precision Comparison

In the above graph, the precision value is compared against the existing approach and the proposed approach. In the x axis numbers of queries gathered from the users are taken. And in y axis, precision is taken. From this graph, it can be proved that the precision value of the proposed methodology is less than the existing approach

CONCLUSION

Personal web search retrieval provides an meaningful way for the web clients who intends to retrieve their contents in the efficient manner. In this work ontological based privacy enhanced user profile is

103

constructed which would attempt to retrieve the user needs contents in the efficient manner. The ontological profiles provide a semantic overview of users interest scores thus the retrieval contents becomes easier. The experimental tests were conducted in the java simulation environment from which it is proved that the proposed approach provides an secured and privacy enhanced accurate retrieval web contents to the users.

REFERENCES

[1]. Jaime Teevan, Susan T. Dumais, Eric 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, pp. 449-456, 2005

[2]. Kazunari Sugiyama, Kenji Hatano, Masatoshi Yoshikawa, "Adaptive Web Search Based on User Profile Constructed without Any Effort from Users", Proceedings of the 13th international conference on World Wide Web, pp. 675-684, 2004

[3]. Feng Qiu, Junghoo Cho, "Automatic Identification of User Interest For Personalized Search", Proceedings of the 15th international conference on World Wide Web, pp. 727-736, 2006

[4]. Yabo Xu, Benyu Zhang, Zheng Chen, Ke Wang, "Privacy-Enhancing Personalized Web Search", Proceedings of the 16th international conference on World Wide Web, pp. 591-600, 2007

[5]. Alexander Pretschner, Susan Gauch, "Ontology Based Personalized Search", Tools with Artificial Intelligence, 1999. Proceedings. 11th IEEE International Conference on 1999

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

[7] 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.

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

[9] J. Pitkow, H. Schu¨ tze, T. Cass, R. Cooley, D. Turnbull, A.Edmonds, E.Adar, and T. Breuel, "Personalized Search," Comm.ACM, vol. 45, no. 9, pp. 50-55, 2002.

[10] 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.

Author Biography

Dr. N. Umadevi working as Head and Associate Professor in the Department of Computer Science and Information Technology, Sri Jayendra Saraswathy Maha Vidyalaya College of Arts and Science. She has 3 years of industrial experience and 13 years of teaching experience. Her area of interest are image processing and Data mining. Her publications include 8 international journals. She has presented papers in 6 international conferences and 4 national conferences

V. Ramya pursuing her M.Phil in Computer Science, Sri Jayendra Saraswathy Maha Vidhyalaya College of Arts and Science, Singanallur, Coimbatore under the guidance of Dr. N. Umadevi working as Head and Associate Professor, Department of Computer Science and Information Technology, Sri Jayendra Saraswathy Maha Vidyalaya College of Arts and Science, Singanallur, Coimbatore. Her area of interest is Data mining. She has presented paper in 1 International Conference and 1 National Conference.

105