A Privacy Enhancing Infomediary for Retrieving Personalized Health Information from the Web

Yinggang Li*
School of Informatics
Indiana University
Bloomington, Indiana
yinli@indiana.edu

Javed Mostafa†
School of Informatics
Indiana University
Bloomington, Indiana
jm@indiana.edu

Xiaofeng Wang‡
School of Informatics
Indiana University
Bloomington, Indiana
xw7@indiana.edu

ABSTRACT

With explosive growth in number of information sources, users now can access a wide variety of health information from the Web. However, information that may be potentially relevant to individual users remain highly scattered and users frequently have to "hunt" for and aggregate information from multiple sites. Additionally, in the process of finding information users often expose personal information in exchange of focussed and individualized information services. We introduce a trusted Infomediary model named MedSIFTER as a one-stop-shop access point to personalized health and medical information. The model centralizes personal information management at Infomediary level to facilitate specific information aggregation task of individual clients. It employs group query mixing and noise query mixing in order to hide user's profile from external eavesdropper. Experiments were conducted to demonstrate trade-off levels between retrieval performance and the degree of privacy preservation in our proposed query mixing strategies.

Categories and Subject Descriptors

H.4 [**Information Systems Applications**]: Personalization and personal information management; K.4.1 [**Computers and Society**]: Public Policy Issues—*privacy*

General Terms

Security, Experimentation

Keywords

Privacy Preservation, User-Profile, Web-security

*yinli@indiana.edu
†jm@indiana.edu
‡xw7@indiana.edu

1. INTRODUCTION

Personalization is increasingly becoming a standard function in many information portals and search engine sites, and searching for health information is a frequent activity of Web users. As health care organizations discover the benefits of digitization of health records, they are likely to invest more in this area and it is likely to encourage more direct and personalized access to health information. For these reasons, it is important focus on the complexities associated with supporting personalized access to health information.

The extent of privacy protection of personal health information often depends on where the information is located and the purpose for which the information is compiled. The federal Health Insurance Portability and Accountability Act (HIPAA) sets a national standard for privacy of health information(effective April 14, 2003). But HIPAA only applies to personal medical records maintained by health care providers, health plans, and health clearinghouses. A great deal of health-related information dispensed on the Web exists outside of health care facilities, thus beyond the reach of proper legal regulations.

To support personalization, interest-specific information is usually gathered either statically by directly querying the user, or dynamically by automatically monitoring user's interaction with information resources [1]. Typically, personalization is carried out on individual sites based on local strategies and content. Sites differ on representation format, content, and profiling algorithm. Compounding this issue is the fact that user's interaction varies from site to site. There exists no unifying patient-oriented personal information management system.

To ensure privacy, many web sites require "proof" of identity before allowing use. Individuals who do not provide essential identifications might be denied services as a consequence. This identification is required in many medical information services that serve dedicated user groups (e.g., registered patients), or in general health-related portals to ensure higher quality of information through paid membership. Even if such identifiers are not required, frequent query patterns when correlated with IP addresses may still lead to certain amount of undesirable leakage of personal information. In most sites, level of control offered to share personal information or "trade-off" such information for more refined personalization services is rather limited.

[1]Some systems use a combination of these approaches to minimize cognitive load on the user.

In this work, we present a multi-tier health information delivery environment, where a trusted Infomediary named MedSIFTER serves as a one-stop-shop access point to support individual's need of quality health and medical information. The system centralizes the personal information management at the Infomediary level. The personal information aggregation, management, and use are located at the Infomediary level. Additionally, the Infomediary supports personalized information retrieval without revealing such sensitive information to external eavesdroppers or to systems from which health information is retrieved.

2. THE MEDSIFTER MODEL

2.1 System Overview

Figure 1 illustrates the structure of a multi-tier information delivery system called MedSIFTER. It offers an intermediate layer between the information requestor and the information resources, called the Infomediary server, which acts on behalf of the requestor to retrieve information from the resource servers. The Infomediary plays two central roles in health information retrieval: *personal information management(PIM)*, and *privacy protection proxy*. This mechanism helps mitigate threats resulting from direct interactions while facilitating personalized information delivery.

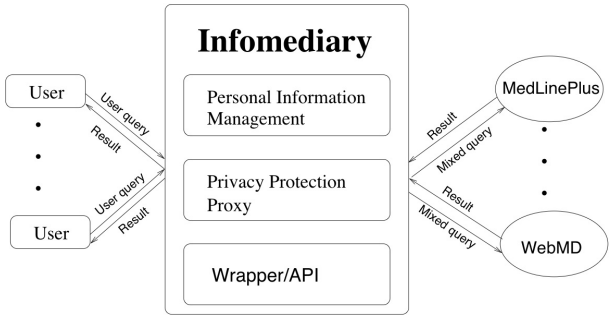

Figure 1: MedSIFTER System Architecture

2.2 Intelligent PIM

Information queries on health and medical topics are highly dependent on user's needs that remain stable over relatively long duration. Information personalization facilitates increased learning and greater engagement in the management of one's own health care. Realized in the MedSIFTER system, personal information management components handle user attributes and transaction states:

- *User's attributes* refer to user identifiers and profiles. Identifiers are in the form of legal identity or pseudonym that is necessary for getting information from external sources requiring authorization. A profile consists of user's interests on specific health topics that correspond to user's medical condition. It is either supplied by the user explicitly or it is derived from the user's interactions over time. The Infomediary uses the profile to predict user's behavior and determine the relevance of information retrieved from external sources. In this work, each profile is a health term vector, whereby each element is a numerical weight representing user's interest in specific health topics. The set of identifiers associated with an unifying profile describes an individual user model that the Infomediary manages and uses for providing personalized services.
- *Transaction states* refer to user's interaction behavior, statistics of activities, and feedback. It shows how the user model evolves over time. For instance, subsequent search queries may reinforce the previously built user profile, or shift it when the user develops interests in a new health topic.

By centralizing personal information management at the Infomediary, the system decouples PIM functionality from the information providers, thus reducing the possibility of undesirable or unintended loss of profile information. It is also more user-centric to provide an integrated access point to personal information.

2.3 Privacy Protection Proxy

The user profile is vital in getting personalized health information. In our model we use a weighted vector of significant health terms that a user may query from a health search engine. The weight distribution reflects the relevancy of topics for a specific user. Over time, the user is more likely to retrieve information related to those high weight terms. The eavesdropper at the information- provider end is capable of deriving accurate user profile by applying various acquisition methods on user's query terms.

When users need to identify themselves in order to receive services, a straight-forward approach of safeguarding profiles is to hide the actual profiles in a broader topical group or cluster. If the dimension of interest vector space is N, we send the complete set of all N terms for each query, i.e. mix the real query terms with the other $N-1$ terms. The probability of detecting each real term from the set is $1/N$. Unobservability is thus achieved within the domain of profile vector space. Ideally, the eavesdropper obtains no additional knowledge about user's real profile if each query term is mixed with the entire set of taxonomy terms. Due to the sheer number of possible terms, a more pragmatic way is to generate a fairly large evenly distributed set that can effectively blur the real query terms.

In MedSIFTER, the Infomediary sever acts as a mix server. It combines multiple requestors' queries into a larger batch, sanitizes these queries, and then sends the batch to resource servers. This foils an eavesdropper's attempt to build individual initiator's profile and also mitigates threats from timing attacks. To further control the inference over statistical data, the Infomediary server may generate dummy queries (i.e. noise) and mix them with the real ones to straighten the distribution trend of queries.

3. EXPERIMENTS

By implementing the MedSIFTER model a prototype system was developed. Several experiments were then conducted to examine the effectiveness of privacy protection options (i.e., query mixing strategies) provided by the MedSIFTER system.

3.1 Mixing Strategy

1. *Group mixing*
   If users form a non-homogenous group in terms of user profile, complete unlinkability can be achieved within

the set of all clients served by our Infomediary. External observer obtains no additional information about which specific user the query belongs to. In this respect, users form an anonymity set.

All query terms from the batch group query are mixed together by union to form a minimal group query-term set. A user is then randomly picked from the group as the representative. If the group query set is distinct or sufficiently blur from the chosen user's profile, we retrieve information using the chosen user's identifier and the group query terms set. Otherwise, we pick another user. After the search results are returned to the Infomediary, information is dispatched to each individual user according to their real profiles maintained at PIM component.

When the number of non-homogenous users served during a single session is large, the group query terms is broad enough to hide the terms in real query.

2. *Noise mixing*
Noise mixing strategy adds noise query terms to each individual user query until the number of elements in the query set is sufficiently large. When the search results are returned to the Infomediary, it only forwards to the user the information relevant to the original query terms.

   The quality of noise mixing depends on how the distribution of noise terms are related to the distribution of terms in a user's profile. Randomly generated noise queries may not hide a user's query well, because one could launch an "intersection attack" to find synonymous keywords in multiple consecutive batches of queries. Using this approach and clustering techniques, an information resource might be able to establish end user's profiles. Good noise candidates are terms that are "near" the real terms in the health-term taxonomy.

3. *Hybrid*
A key concern with group mixing is performance. To hide one's message using a batch, the Infomediary has to temporarily hold a requestor's query until sufficient number of messages are received. This increases the delay perceived by a user. Hybrid strategy alleviates the problem by introducing noise terms to generate a group batch query.

   When there is not sufficient number of users for the current retrieval session, the Infomediary first attempts to generate a small group query term set using group mixing. The remaining terms needed to form a large term set are complemented by noise terms. When the number of users are large, this strategy typically involves group mixing without noise. In case the group profile has a high similarity to members, noise terms are still needed to replace some of the terms in the group query generated by group mixing.

3.2 Results and Discussion

Experiments were conducted to evaluate the MedSIFTER's interaction with simulated clients under different mixing strategies. A user profile is represented as a health term vector weighted by user preference. A user query is a set of health terms to which the user has preferences higher than a preset threshold. Average turn-around time for each user query was plotted against the number of simulated users served simultaneously in a retrieval session.

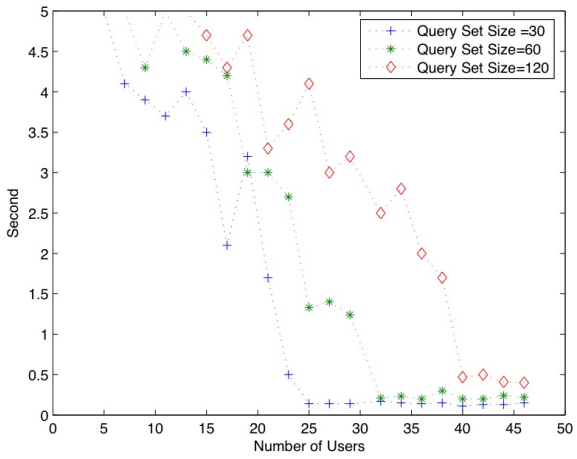

Figure 2: Average turn-around time with group mixing

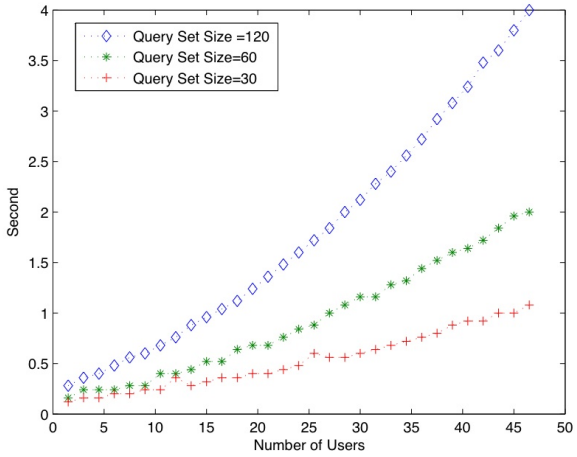

Figure 3: Average turn-around time with noise mixing

In the case of pure group mixing, a group query with fixed size of terms is generated by each MedSIFTER retrieval session. As described earlier, larger query size indicates lower probability of observing a real user query term. Figure 2 illustrates that the Infomediary has to wait a significant amount of time before starting a retrieval task when the number of users is small. The waiting time at the Infomediary contributes to the longer overall turn-around time for each user query. If enough queries arrive at the Infomediary, the preset group query set size can be easily satisfied, and the Infomediary is ready to start the retrieval without waiting. We therefore observe a drop in turn-around time, which is proportional to the average arrival time of simulated user in our experiment. The larger the mixed query set size, the more users are needed to reach this non-wait drop point. To

generate a mixed group query with 30, 60 and 120 terms, the minimum number of users needed for the Infomediary to start a retrieval without waiting are 25, 30, and 40 respectively. With pure noise mixing, the Infomediary adds noise terms to every individual query when it arrives, and then sends to external information sources. The total number of terms sent out at each retrieval session is the multiplication of number of user with the size of noise mixed querv.

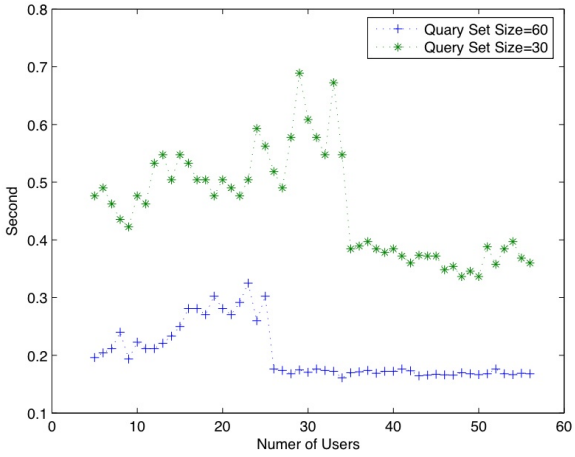

Figure 4: Avg. turn-around time with hybrid mixing

The increasing turn-around time we observe in Figure 3 is due to the effect of bandwidth overhead incurred by increasing query terms. For example, 6000 query terms are transferred by the Infomediary when 50 users are served simultaneously and a noise mixed query of 120 terms is formed for each user. This bandwidth overhead can be reduced if we cache some of the noise query terms search result for the subsequent retrieval sessions. For every inbound query, the server can check its cache and if there is a match, it simply sends back the results; otherwise, the Infomediary proceeds as usual.

Looking at Figure 4 it can be seen that hybrid mixing successfully inherits the merit of both group mixing and individual noise mixing. If the number of arriving users is small when a waiting timeout is reached, noise terms are added to form a group query of larger size. On the other hand, no noise terms are needed if sufficient number of distinct users are served in a batch retrieval session. The average turn-around time is relatively low for a wide range of concurrent user service requests.

We suggest that a practical mixing strategy would be a hybrid one. The size of group query should depend on the balance of privacy level and the acceptable waiting timeout set at the Infomediary. The solution would be more complete if the timeout and mixed query size are adaptive to changes of privacy protection level set directly by the user. The Infomediary may also monitor the performance measurement and calculate the observability of user profile from the group query, so that it can adjust the hybrid parameters accordingly.

4. CONCLUSIONS AND FUTURE WORK

We have presented a trusted Infomediary model that offers good privacy protection while providing personalized health information retrieval services to end users. By centralizing personal information management, it makes possible delivering personalized patient-centric health information based on user profiles. It also facilitates group query mixing among non-homogenous user-profiles maintained at the Infomediary level.

Further studies are required in order to explore if reusability of noise terms by applying caching strategies can reduce retrieval latencies. Noise terms with good reusability should effectively hide the user profiles in the current session and predict user-interests in future sessions. This calls for an intelligent learning of information that users are likely to retrieve in the future.

Although the Infomediary server can generally be treated as trusted mediators, a user may not be willing to commit full trust in a single PIM. We plan to develop a mechanism for users to establish anonymity groups in a peer-to-peer network. In such a P2P network, clients will be able to send encrypted profiles to an Infomediary server through other clients, thus hiding the origin of queries from the Infomediary as well.

5. ACKNOWLEDGEMENT

This research was partially funded by NSF grant 054931.