A Novel Query Obfuscation Scheme with User Controlled Privacy and Personalization

Saraswathi Punagin
Asst. Professor, Dept. of CSE
PESIT – Bangalore
South Campus

Arti Arya
Head, Dept. of MCA
PESIT – Bangalore
South Campus

ABSTRACT
Web Search Engines are tools that help users find information. These search engines use the information provided by users, in terms of their search history to build their "user profiles". Rich user profiles enable the search engines to provide better personalized search results. However, this puts the user's privacy at risk. Apart from the risk of exposing one's identity, there is the added disadvantage of being subjected to unsolicited advertising and potential disclosure of sensitive information. Rich user profiles contain a lot of personally identifiable information, which can attract unwarranted malicious interests. It is important that sensitive data collection be curbed or at least obfuscated at the source. To that effect this work is a novel approach towards providing a balance between privacy preservation and personalization by keeping the user in control of his privacy Vs personalization decisions. This work supports complex queries and obfuscates them by adding a set of fake queries that are semantically related to the original query where both the semantic distance and the number of fake queries are user controlled parameters.

General Terms
Data Privacy, Obfuscation

Keywords
Private Search, Query Anonymization, User Control, Privacy, Personalization, Web Search

1. INTRODUCTION
Web Search Engines (WSE), are a vital part of our everyday life because they are easy to use and generate useful results quickly. In fact, Web Search Engines, have played an important role in boosting the growth of the World Wide Web. WSEs present search results using several pages. 68% of the users click a result on the first page while 92% of them click a result on the first three pages [1]. This has led to page ranking [2] where results are ordered via two factors: Sponsored links (direct revenue) or Enhanced User Experience (indirect revenue). A search engine is only as successful as the number of useful search results it provides. The WSEs provide Enhanced User Experience by displaying search results that make sense to the user which is in turn made possible by understanding the true intent of the user or by knowing what interests the user.

Knowing a user's true interest is complicated. One way of doing this is by creating a user profile based on the user's search history. The user profiles not only enable the WSEs to provide a better user experience but are also a source of revenue to them when these profiles are sold at a price to law enforcement agencies or other marketing partners.

User profiles contain a lot of sensitive information and the WSEs are supposed to store them securely which is not always guaranteed. Search engines and recommendation systems benefit from users who make personal information available, thereby providing tailor-made search results and/or recommendations. However, a user risks his/her privacy to gain personalized recommendations.

As part of this research work, a study was conducted with 660 participants to gauge the privacy and personalization perceptions of the Indian demographic with respect to web searches. The results indicate that while there is a very low percentage of Indian consumers who are Fully Privacy Aware (11%), there exists a moderate number of consumers who are Fully Customization/Personalization Aware (55%). Percentage of consumers who dislike being tracked online is 37% and consumers who took some action to protect their online privacy during web searches were 25%. Details of this study can be found in [3].

Personalized search results and/or recommendations are a result of rich user profiling. Search engines employ usage mining techniques to build user profiles. Usage mining puts user's privacy at risk. While there are solutions that attempt privacy preservation or user profiling exclusively, there is a need for a solution that provides both and puts users in control of the level of privacy preserved Vs usefulness of user profiles.

Objective of this research is to come up with a new, robust and effective, ontology based approach to obfuscating complex search queries by way of adding a set of fake queries that are at a given semantic distance from the original query. The solution is a layer that resides on top of Web Search Engine (WSE), which accepts search queries and obfuscation parameters from the user, anonymizes the queries and finally submits both the original as well as the fake queries to the WSE. The novelty lies in the fact that the user is put in control of the amount of Privacy Vs Personalization. The semantic of the original query is preserved and required privacy levels are achieved via obfuscation without sacrificing response time.

2. RELATED WORK
In the age of pervasive internet where people are communicating, networking, buying, paying bills, managing their health and finances over the internet, where sensors and machines are tracking real-time information and communicating with each other, it is but natural that big data will be generated and analyzed for the purpose of "smart business" and "personalization". Today storage is no longer a bottleneck and the benefit of analysis outweighs the cost of making user profiling omnipresent. However, this brings with it several privacy challenges – risk of privacy disclosure without consent, unsolicited advertising, unwanted exposure of sensitive information and unwarranted attention by malicious interests. Punagin and Arya in [4] survey privacy risks associated with personalization in Web Search, Social

Networking, Healthcare, Mobility, Wearable Technology and Internet of Things. The authors review current privacy challenges, existing privacy preserving solutions and their limitations.

Viejo et al. in [5] propose profiling users locally based on their social network account usage instead of their web search history. They then use the local profile as a base to obfuscate the user's search queries. By not using a user's web search history to create a user profile, this approach forces search engines to focus on a user's "macro" interests instead of their fine-grained "micro" interests. However local profiles built thus are static and may not represent a user's true interests over time. Also, query obfuscation in this approach, is based on a high-level characterization of user's interests and ignores the semantic preservation of the original query.

Sanchez et al. in [6], propose a mechanism where users are in control of the amount of private information they reveal vs. the degree of richness their user profiles retain. Their solution obfuscates the original query by adding "k" number of fake queries to the original query set. The contributions of Sanchez et al. include the proposal of a new scheme to generate distorted queries from a semantic point of view, thereby preserving utility of user profiles. Their work also supported complex queries and provided a tradeoff between Utility and Privacy addressed via configurable parameters.

While the contributions of their work towards providing users with control over their privacy preservation and personalization are exciting, the performance impact is not studied in detail. The solution makes use of knowledge bases like WordNet and Open Data Project (ODP) to extract query topics and concepts that are at a given semantic distance from the query topic. This is an important step in their obfuscation technique. While the average time taken for a query found in WordNet is 30 ms, it is 1500 ms for one not found in WordNet and found in ODP in the second iteration. Also, the approach obfuscates the search queries, but the submission of these queries to search engine is not linked to a particular protocol.

Hassan et al. in [7] discuss the ongoing European Union funded EEXCESS (www.eexcess.eu) project as an example of providing improved user recommendations by making use of intensive user profiling techniques. One of the major challenges is that the EEXCESS architecture is based on a federated recommender system in which future partners may join. The trustworthiness and the intent of these partners are not necessarily known. The information collected and disclosed to recommenders may not, in itself, be sensitive; however, cross-referencing it with external big data sources and analyzing it through big data techniques may create breaches in user privacy.

Since, untrustworthy partners may have access to such big data sources and analysis techniques, privacy becomes a clear challenge. The EEXCESS project addresses the challenges of guaranteeing privacy, based on flexible privacy policies and evaluating the trust and reputation of a recommender.

Although EEXCESS project proposes a novel user-controlled approach to privacy preserving searches that ensure rich recommendations as results, it requires a complete architectural change to how web searches are done.

Batet et al. in [8] present a query anonymization method based on semantic micro aggregation which focuses on reducing the risk of query log disclosure while retaining the utility of the query logs at the same time. In this approach, the semantic concepts of the query logs are retrieved from Open Data Project. The query logs are then semantically aggregated using a variation of the Maximum Distance Average Vector (MDAV) algorithm. An approach to generate fake queries which replace original ones is also proposed.

Like most methods based on query aggregation, this approach also suffers from a moderate to high chance of record linkage. Concept recall using ODP is not very high, therefore bringing to light the need to explore this scheme using other ontologies.

Carpineto et al. in [9] propose a scheme to anonymize query logs by leveraging the affinity between frequent canonical concepts and their infrequent refinements. This approach is able to mask identification of queries while retaining the semantic of the query logs. However, computation of the affinity measure between queries can be improved by using better similarity measures and auxiliary information. K-anonymization based on taxonomic generalizations are known to yield better results than this scheme. Also, the sensitivity of this scheme to an attack is yet to be analyzed.

Yeye et al. in [10] propose a top-down, partition based approach to anonymizing set-value data that preserves better utility most of the time. While this approach works sufficiently well for query log anonymization it does not work well with market-basket type of data. Also, this work does not acknowledge the distinction between sensitive vs. non-sensitive attributes or the notion of quasi identifiers.

Hong et al. in [11] applied k-anonymity at user level where they cluster users based on the similarity of their query data. However, clustering takes away the order of the queries. This scheme also deletes the original data and adds new fake data. There is little proof about the amount of utility retained using this method.

Zhu et al. in [12], proposed a grouping approach for anonymizing user profiles with a p-linkability notion where they bound the probability of linking a potentially sensitive term to a user by p. This scheme utilizes a greedy clustering technique with semantic similarity metric that takes into account the semantic relationships between user profiles. The authors were successful in experimenting the tradeoff between search quality and privacy protection. However, this work was limited to the AOL dataset which has several limitations with respect to extraction of users' specific interests and therefore warrants further research.

3. PROPOSED METHOD

This work is motivated by the need to give control back to the user as to how much of his privacy he is willing to forgo to reap the benefits of personalization. Query data is obfuscated "at the source" with fake queries that are semantically related to the original query. This solution supports complex queries, accepts obfuscation and semantic relation parameters from the user and also includes a protocol to submit the user's original query along with the fake ones to a WSE, thereby simulating a real user. The scheme also creates and manages a local user profile (for performance and future re-ranking of search results).

Another significant contribution of this work is recognizing compound noun phrases like "White House" along with independent nouns like "White" and "House" to identify the query topic. Query topics in this approach are calculated based on their Information Content (IC). Depending on the linguistic usage, a compound noun phrase might have a higher

IC than the individual nouns considered by themselves. This approach is explained in section 4.2.

The following sub sections describe the salient features of the proposed approach.

3.1 Support for Complex Queries

This solution supports queries with more than one term or with multiple noun phrases. Dealing with complex queries such as these requires analyzing of the query by employing Natural Language Programming (NLP) techniques like auto-correction, stemming, stop-word removal, tokenization, part of speech (POS) tagging and Noun Phrase (NP) chunking. The solution also considers compound Noun Phrase term combinations for assessment of the query topic instead of using individual Noun Phrase terms. This provides an opportunity to accurately assess query topics stemming out of compound words.

3.2 User Controlled Parameters

Users decide the balance between the amount of privacy they are ready to forgo to reap the benefits of personalization and vice versa. This control is facilitated by way of two parameters: semantic distance and number of fake queries. Semantic distance decides how far away from the meaning of the original query, the fake queries will be. Number of fake queries is instrumental in deciding the level of privacy exposure.

3.3 Local User Profile

Table 1 provides a sample of the Local User Profile (LUP) that is created from the user's browsing history where all query topics along with their related fake concepts (that are at a particular semantic distance from the query topics), are stored. When analyzing each user query, the LUP is consulted for fake concepts at first. If a particular query topic is not found in the LUP, then WordNet is explored to extract fake concepts that are at a given semantic distance from the query topic.

LUP improves performance for query topics that find a hit in it. LUP can also be harnessed for future enhancements to the project by using it to re-rank the search results returned by the WSE according to the true interests of the user.

Table 1. Local User Profile

Semantic Unit	IC	Semantic Distance	Fake Concepts
Depression	4.84	1	Psychological state, psychological condition, mental state, mental condition
Depression	4.84	2	Condition, state, situation
Water Sport	4.39	1	Sport, athletic, surfing, surf boarding
Water Sport	4.39	2	Diversion, recreation, bathe, dip

3.4 Use of ontologies like WordNet

The solution has been designed to use any knowledgebase that has a taxonomic structure. At present, it works with a domain-independent knowledgebase WordNet [13], which has over 100,000 general English concepts which are semantically structured in an ontological fashion. It contains words expressed as synsets, corresponding to a word sense. Synsets are linked to each other by way of hypernyms, hyponyms, meronyms and antonyms creating a graph like structure that can be exploited to interpret the meaning of a given concept. WordNet can be downloaded and consulted offline and has APIs for many programming languages including python. A partial taxonomy for the word "Depression" is presented in Figure 1 below.

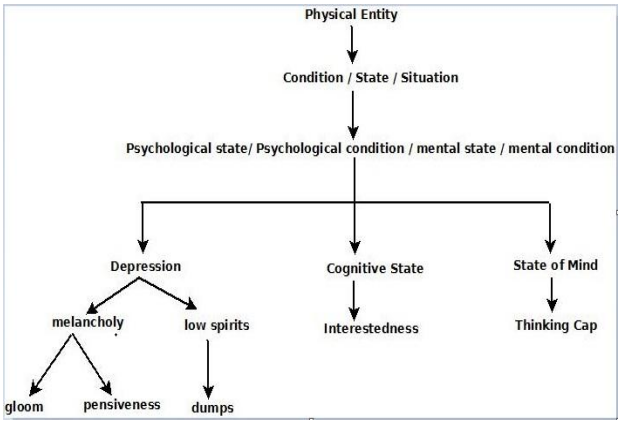

Fig 1: Partial Wordnet Taxonomy for the word "Depression"

3.5 Architecture

The architectural diagram in Figure 2 illustrates all the components of the system and the interactions between them. The Privacy Component is at the core of the architecture and sits on the client machine. It accepts user queries, consults the Local User Profile or the Ontology as needed and extracts the query topics as well as fake concepts to anonymize the query.

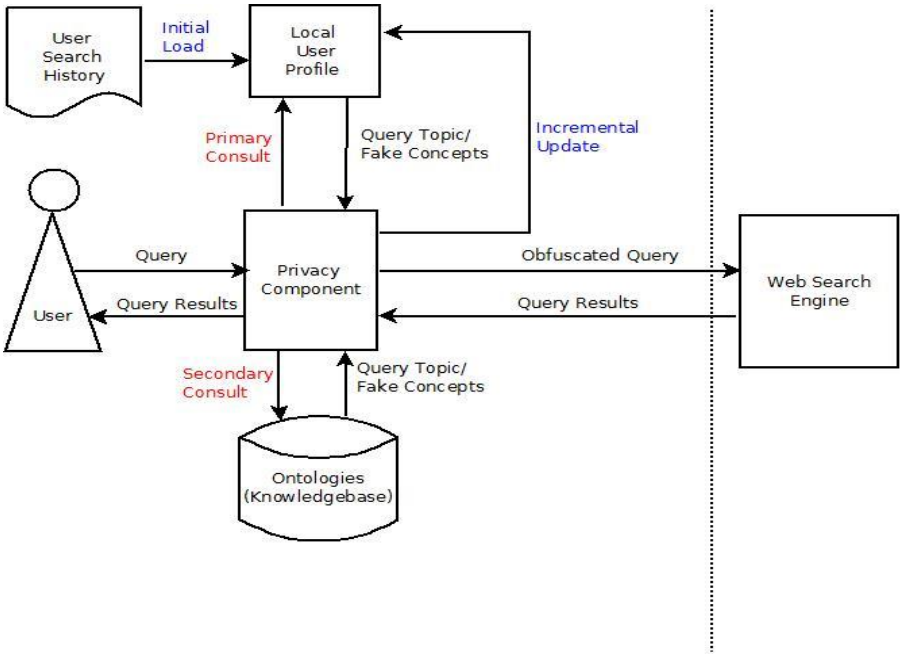

Fig 2: Architectural Overview

4. ALGORITHM AND WORKFLOW

4.1 Algorithm

Algorithm: Anonymize_Query

Input: original query q_i, semantic distance sem_dis, number of fake queries k and local profile LP, where each $QT_i \in \text{LP} = \{mi\_sui_i, sem\_dis, hypernhyms, hyponyms, siblings\}$

Output: set of k fake queries $NQ_i = \{nq_{i1}, nq_{i2}, .., nq_{ik}\}$

1. Generate semantic units of $q_i, SU_i = \{su_{i1}, su_{i2}, \ldots., su_{im}\}$
2. If $\text{su}_{ij} \in LP.QT_i.mi\_su_i \wedge sem\_dis = LP.QT_i.sem\_dis$
   a. Retrieve new concepts $NC_i = \{LP.QT_i.hypernyms \cup LP.QT_i.hyponyms \cup LP.QT_i.siblings\}$
3. Else
   a. Calculate taxonomy based Information Content (IC)
   b. Identify the semantic unit with the highest information content mi_su_i
   c. Retrieve from the taxonomy, a set of new concepts $NC_i = \{nc_{i1}, nc_{i2}, \ldots, nc_{ik}\}$ such that $distance(nc_i, mi\_su_i) = \text{min\_path}(\text{nc}_i, \text{mi\_su}_i) = \text{sem\_dis}$
   d. $LPUpdate(LP, mi\_su_i, NC_i, sem\_dis)$
4. Select k of these $NC_i = \{nc_{i1}, nc_{i2}, \ldots., nc_{ik}\}$ to construct k new fake queries $NQ_i = \{nq_{i1}, nq_{i2}, \ldots., nq_{ik}\}$
5. Submit $q_i \cup NQ_i$ to WSE

4.2 Analysis of the Query

- Input Query: Suppose the input query is of the form "Statistics of deaths from postpartum depression".
- Tokenization: As a first step this query is tokenized into the following tokens: "statistics" / "of" / "deaths" / "from" / "postpartum" / "depression"
- Part of Speech (POS) tagging: The tokens are then tagged based on the part they play in a given sentence. "statistics": Adjective; "of" / "from": Prepositions; "death" / "postpartum" / "depression": Nouns
- Stop word Removal: Words with general meanings like determinants and prepositions can be removed without altering the meaning of a sentence. For example, "some lady" becomes "lady". In the above example, the two prepositions "of" and "from" are removed.
- Stemming: Remove derivational affixes of words. For example, get rid of plural forms. This ensures we identify the morphological variations of the same term during processing natural language. In the above example "statistics" becomes "statistic" and "deaths" become "death" without losing their meaning.

  Noun Phrase chunking (NP Chunking): The part of speech tagged tokens are then fed to a regular expression parser which chunks these tokens into Noun Phrases based on a universal grammar. The grammar used in this approach is as follows:

  "NP: {(<JJR>|<JJS>|<JJ>|<NN>|<NNS>|<NNP>|<NNPS>)*}"
- As a result of NP Chunking we get the following Noun Phrases: "statistic" / "death" / "postpartum" / "depression"
- Compound Noun Phrase terms: If a noun phrase comes back as "white" / "house", it is important that we consider the concatenated compound NP terms as a semantic unit before we look at the individual terms. This is important because the term "white house" provides more information content than the individual terms "white" and "house". At the end of this process, we will have a set of semantic units $\text{SU}_i = \{\text{su}_{i1}, \text{su}_{i2}, \ldots., \text{su}_{im}\}$ that represent the input query q_i.

4.3 Assessment of the Query Topic

The semantic unit with the highest Information Content (IC) is chosen as the query topic as:

$$\text{mi\_su}_i = \{\text{su}_{ij} \mid \text{IC}(\text{su}_{ij}) = \max\left(\text{IC}(\text{su}_j)\right) \forall\, \text{su}_{ij}\} \quad (1)$$

The project uses the classic approach to calculating the IC by using the taxonomical structure of WordNet. IC is the inverse of the appearance probability. Higher the term in the taxonomy, lower is its IC. More specialized the term is it is going to be lower in the taxonomy and have a higher IC. Therefore, IC is calculated as:

$$\text{IC}(\text{su}_{ij}) = -\log\left(\frac{\frac{|\text{leaves}\ (\text{su}_{ij})|}{|\text{subsumers}\ (\text{su}_{ij})|}+1}{|\max\ \text{\_leaves}\ |+1}\right) \quad (2)$$

4.4 Retrieval of Fake concepts from the Ontology

User specified parameters semantic distance and number of fake queries are used to retrieve fake concepts from the ontology which are at a distance of sem_dis from the query topic concept.

The first step is to retrieve the concept c_i associated with the query topic mi_su_i from the ontology. Next, we retrieve new concepts $\text{NC}_i = \{\text{nc}_{i1}, \text{nc}_{i2}, \ldots, \text{nc}_{\_ij}\}$ such that the distance between the query topic concept and the fake ones is the minimum economic path between the two concepts and is given as:

$$\text{distance}(\text{c}_i, \text{nc}_{ij}) = |\text{min\_path}(\text{c}_i, \text{nc}_{ij})| = \text{sem\_dis} \quad (3)$$

The new concepts that are retrieved include the hypernyms and hyponyms of the synset associated with the query topic concept in the WordNet ontology.

4.5 Construction of fake queries

From the new concepts, NC_i that are retrieved from the ontology, a random set of k concepts are chosen to create a set of fake queries $\text{NQ}_i = \{\text{nq}_{i1}, \text{nq}_{i2}, \ldots\ldots, \text{nq}_{ik}\}$

4.6 Query Submission Protocol

All of the fabricated *k* queries are first submitted to the WSE in quick succession using a background process. The original query is submitted last whose results are fetched and displayed to the user.

5. RESULTS & SYSTEM VALIDATION

The prime motivation behind this work is to obfuscate search queries to make it difficult for the WSEs to distinguish between the real queries made by the user and the fake ones that are submit as a result of the obfuscation process. This ensures that the user's personal information collected by the WSEs is distorted to some extent thereby diluting the user's WSE profile. The secondary aim of this work is to ensure that the user's profile is distorted only to a degree determined by the user himself and that the semantics of the profile is just diluted and not completely lost. A byproduct of such distortion process is the performance overhead and needs to be contained.

As a result of the above considerations, the work is therefore evaluated on the Semantic Preservation, Profile Exposure Level and Response time.

5.1 Test Data Set

To evaluate the scheme described in the previous sections, a set of 1000 randomly selected queries from real user query logs released by AOL in 2006 [14] were picked and anonymized. These are real web queries submit by real users and therefore most of them were complex queries that contain more than one noun phrase. The results were evaluated against three parameters namely semantic preservation, profile exposure level and response time.

5.2 Semantic Preservation

The difference between the Information Content of the original queries to the associated fake ones is a measure of the Semantic Preservation. Ideally, we would want the IC Difference to be as less as possible.

Information Content (IC) of a query is calculated using the classic approach. That is IC is the inverse of the appearance probability of the query in the largest corpus available, i.e. the World Wide Web. Therefore, IC of a given query a, is calculated using the web hit count as:

$$IC(a) = -\log\big(p(a)\big) = -\log\left(\frac{hit\_count\ (a)}{total\ \_webs}\right) \quad (4)$$

Here $hit\_count(a)$ is the number of search results received by querying the WSE with the search query a and $total\_webs$ is the amount of web resources indexed by the WSE. Similarly, the IC of a fake concept b is calculated as:

$$IC(b) = -\log\big(p(b)\big) = -\log\left(\frac{hit\_count\ (b)}{total\ \_webs}\right) \quad (5)$$

Considering $total\_webs$ is a common factor and $-\ log$ is a monotonic function and hence does not alter the relative order between queries a and b, IC calculation can be deduced to:

$$IC(a) = hit\_count(a) \quad (6)$$

The IC Difference there is calculated as:

$$IC\_Differrence(a,b) = \frac{Max\left(hit\_coun\,t(a), hit\_count\ (b)\right)}{Min\,(hit\_count\ (a), hit\_count\ (b))} \quad (7)$$

A total of 1000 random AOL queries with different *semantic distance* and *number of fake queries or k value* were tested. The results are presented below:

5.2.1 *User specified Semantic Distance and k value:*

From the graphs in Figures 3, 4, 5 and 6, it is clear that with a *semantic distance* of 1, 2, 3 and 4, over 93% of the fake queries fall within an *IC Difference* of 1 to 1000 units from the original query. As the *k value or the number of fake queries* increases from 1 to 4, number of fake queries within an IC Difference of 1-10 decreases slightly indicating that the *k value* has little effect on the *IC Difference*.

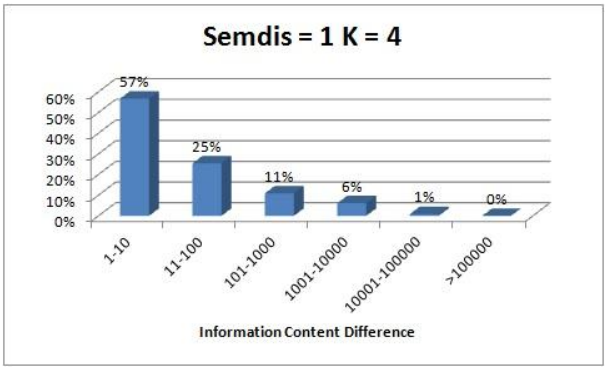

Fig. 3. Information Content Difference Semdis = 1 K = 4

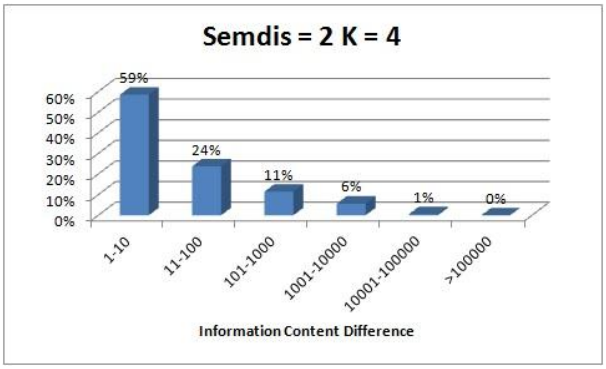

Fig. 4. Information Content Difference Semdis = 2 K = 4

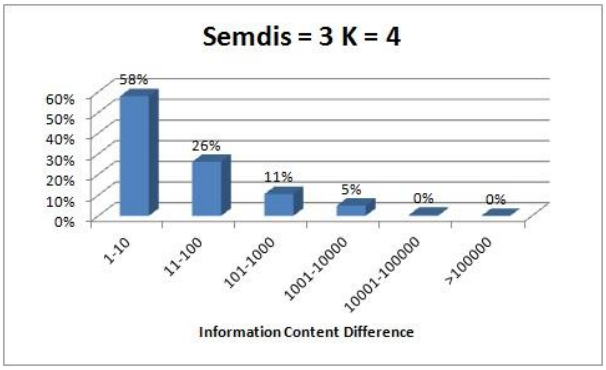

Fig. 5. Information Content Difference Semdis = 3 K = 4

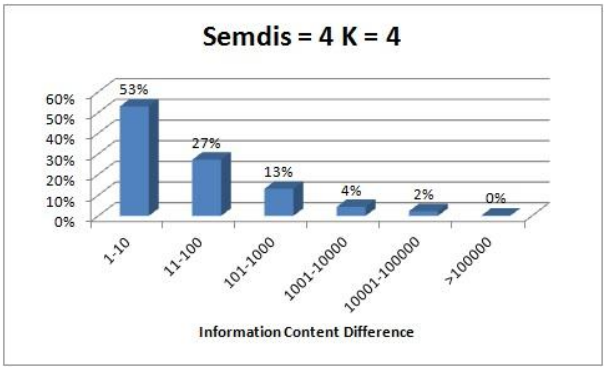

Fig. 6. Information Content Difference Semdis = 4 K = 4

5.2.2 *Random Fake Queries*

The graph in Fig. 7 shows that when k random concepts with no semantic association with the concept of the query topic were chosen and processed for $k\ value = 4$, only 24% of the fake queries stayed within an $IC\ Difference = 1\ to\ 1000\ units$ of the query topic. Rest of the 76% of the fake queries were at an $IC\ Difference > 1{,}000\ units$ from the query topic with a 64% of them being at an $IC\ Difference > 10{,}000\ units$ from the query topic.

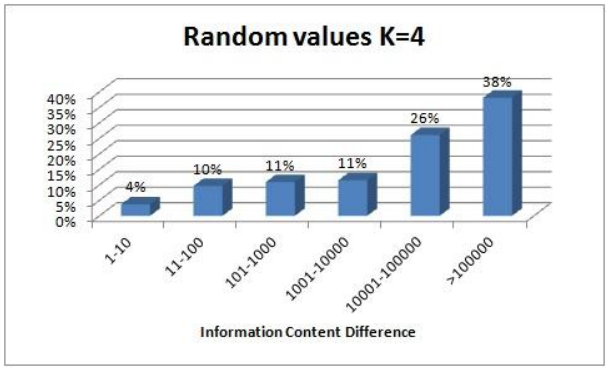

Fig. 7. IC Difference (Random fake queries)

This indicates that running fake queries using randomly selected concepts will severely distort the user's profile and speaks volumes about the need for providing a user controlled semantic distance parameter which controls how far from the meaning of the original query does a user want to go thereby controlling the amount of distortion applied to his user profile.

5.3 Profile Exposure Level (PEL)

PEL is a measure of privacy exposure and was used by Navarro-Arribas, Guillermo, et al in [15]. Suppose X and Y are random variables that can take on any value in the sets of original queries and fake ones respectively. Then Profile Exposure Level [15] can be expressed as:

$$PEL = \frac{I(X,Y)}{H(X)} \tag{8}$$

Here $I(X,Y)$ is the Mutual Information between X and Y and is expressed as:

$$I(X,Y) = H(X) - H(X|Y) \tag{9}$$

Here $H(X)$ is the Entropy of X and $H(X|Y)$ is the Conditional Entropy of X given Y. These equations are given by:

$$H(X) = -\sum_x \mathrm{p(x)}.\log_2 p(x) \tag{10}$$

$$H(X|Y) = -\sum_{xy} \mathrm{p(y)}.\log_2\left(\frac{p(x|y)}{p(y)}\right) \tag{11}$$

The test data set revealed the following results as far as PEL metric is concerned.

5.3.1 *Semantic Distance = 1*: As shown in Fig. 8, a semantic distance of 1 provides a profile exposure level of 94% with one fake query added and about 91% with just four fake queries added.

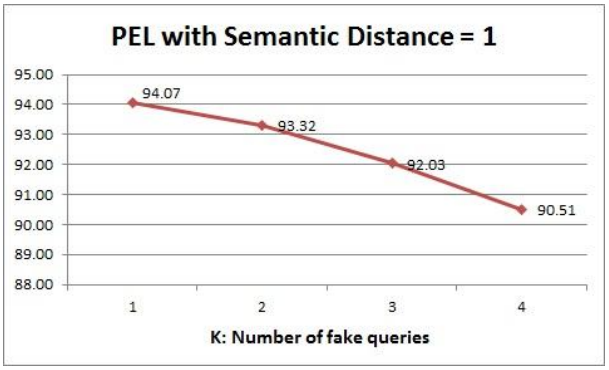

Fig. 8. Profile Exposure Level (Semantic Distance = 1)

5.3.2 *Semantic Distance = 2:* As shown in Fig. 9, a semantic distance of 2 provides a profile exposure level of 81% with one fake query added and only about 73% with just four fake queries added.

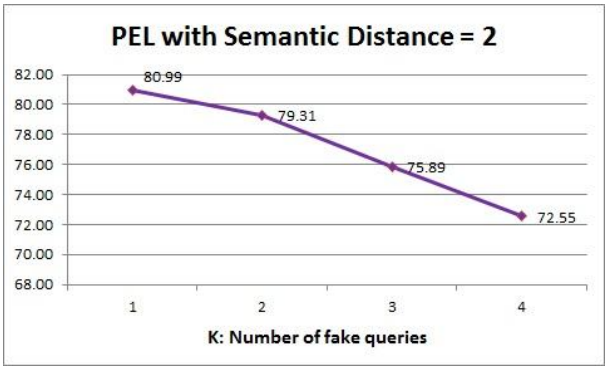

Fig. 9. Profile Exposure Level (Semantic Distance = 2)

5.3.3 *Semantic Distance = 3:* As shown in Fig. 10, a semantic distance of 3 provides a profile exposure level of 65% with one fake query added and only about 47% with just four fake queries added.

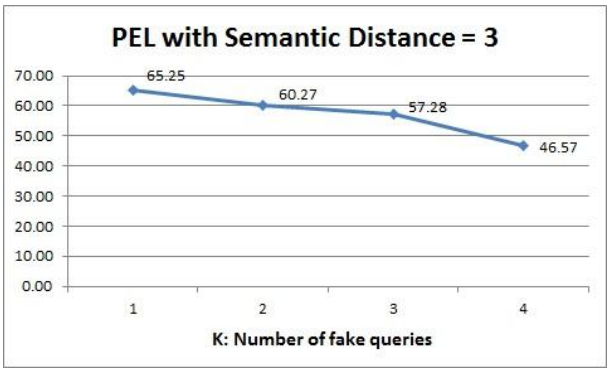

Fig. 10. Profile Exposure Level (Semantic Distance = 3)

These results suggest that both semantic distance and number of fake queries or k value have an effect on the Profile Exposure Level. Lower the semantic distance lower the PEL band is. Within a given band, higher the k value, lower the PEL value. Through these validations, we conclude that with a semantic distance of 3 and a k value of 4, a user using our scheme can achieve 53% privacy or 47% PEL.

5.3.4 *Semantic Distance = 4:* As shown in Fig. 11, a semantic distance of 4 provides a profile exposure level of 29% with one fake query added and a negative exposure level of -5.59% with just four fake queries added.

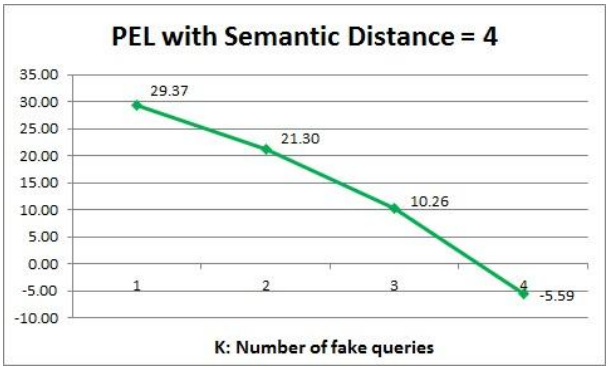

Fig. 11. Profile Exposure Level (Semantic Distance = 4)

A PEL of -5.59% (when semantic distance is 4 and number pf fake queries is 4) indicates that the conditional entropy of X given Y is greater than the entropy of X. This means knowing Y has increased the uncertainty of X. In other words, having more generalized fake queries is increasing the uncertainty of identifying the original queries.

Owing to the larger semantic distance, the original query topics and the fake query topics are more disassociated than associated and this could be because of linguistic reasons.

For example, when a user submits a query like "effects of hodgkins radiation treatment 25 years later", our approach processes this complex query to identify the query topic as "radiation" (NP with the highest IC). It then extracts the following fake query set: {"phenomenon", "event", "information", "process"} from WordNet based on the sem_dis = 4 (semantic distance) and k = 4 (number of fake queries). The query submission protocol then submits all five queries (original query plus four fake queries) to the WSE. This protocol submits a bunch of fake queries around the same time when the original query is submitted. The process is transparent to the user and hence does not interfere with the user's experience. However, obfuscating the original query with a bunch of fake queries ensures, the user's profile with the WSE is muddled.

Through these validations, we conclude that with a semantic distance of 4 and a k value of 4, a user using our scheme can achieve 100% privacy or 0% Privacy Exposure Level (PEL).

5.4 Response Time Evaluation

A direct search on a WSE (Bing in our case), takes about 300ms. Searching for concepts in WordNet takes about 30ms. However, 75% of the time a user runs repeat queries and therefore finds a query topic hit in the Local User Profile saving 30ms per concept search.

6. CONCLUSION & FUTURE WORK

6.1 Conclusion

This scheme puts the user back in control of the privacy Vs personalization decision when it comes to web searches. The user via obfuscation parameters decides the amount of privacy he wants to compromise in order to reap the benefits of personalization and user profiling. The project supports complex queries, preserves semantics of a user's profile and obfuscates his queries with great balance.

The results show that with a *semantic distance* value of 4 and a *number of fake queries or k value* of 4, a user can achieve 100% privacy or a 0% Profile Exposure Level. With these user-defined parameters, the semantic preservation is at high level, with over 93% of the fake queries lying within an *IC*

Difference of < 1000 units from the query topic. The scheme performs well with respect to response time as well with an added overhead of just 30ms per concept search in WordNet. Also, with over 75% of the queries being frequent queries, most of the times a WordNet search is not needed and a query topic hit is found within the Local User Profile.

6.2 Future Work

This work will benefit further from context based disambiguation of the query topic and also considering siblings at a given semantic distance from the query topic concept, while retrieving fake concepts from the ontology. Apart from this adding randomness when picking k fake concepts from a set of $\mathrm{NC_i} = \{\mathrm{nc_{i1}}, \mathrm{nc_{i2}}, \ldots, \mathrm{nc_{ij}}\}$ fake concepts will help increase entropy. Emulating real sentences while constructing fake queries instead of just using the lemma names of the selected synsets will enhance the obfuscation and make it more human like. One could also explore *domain dependent* ontologies to increase hits for *domain specific query terms.* The Local User Profile constructed and maintained can be harnessed to re-rank the query results that are returned by the WSEs as per the user's true interests.

7. REFERENCES

[1] Sullivan, Danny. "How search engines work." SEARCH ENGINE WATCH, at http://www. searchenginewatch. com/webmasters/work. html (last updated June 26, 2001)(on file with the New York University Journal of Legislation and Public Policy) (2002).

[2] Fellbaum, Christiane. WordNet. Blackwell Publishing Ltd, 1998.

[3] Punagin, Saraswathi, and Arti Arya. "Privacy and Personalization Perceptions of the Indian Demographic with respect to Online Searches." Proceedings of the Third International Symposium on Women in Computing and Informatics. ACM, 2015.

[4] Punagin, Saraswathi, and Arti Arya. "Privacy in the age of pervasive internet and big data analytics - challenges and opportunities." International Journal of Modern Education & Computer Science 7.7 (2015).

[5] Viejo, A; Sanchez, D. Providing useful and private Web search by means of social network profiling. Privacy, Security and Trust (PST), 2013 Eleventh Annual International Conference on , vol., no., pp.358,361, 10-12 July 2013

[6] David S´aNchez, Jordi Castell´ı-Roca, and Alexandre Viejo. Knowledge-based scheme to create privacy-preserving but semantically-related queries for web search engines. Inf. Sci., 218:17–30, January 2013.

[7] Omar Hasan, Benjamin Habegger, Lionel Brunie, Nadia Bennani, and Ernesto Damiani. A Discussion of Privacy Challenges in User Profiling with Big Data Techniques: The EEXCESS Use Case. Proceedings of the 2013 IEEE International Congress on Big Data (BIGDATACONGRESS '13). IEEE Computer Society, Washington, DC, USA

[8] Batet, Montserrat, et al. "Utility preserving query log anonymization via semantic microaggregation." Information Sciences 242 (2013): 49-63.

[9] Carpineto, Claudio, and Giovanni Romano. "Semantic search log k-anonymization with generalized k-cores of query concept graph." Advances in Information Retrieval. Springer Berlin Heidelberg, 2013. 110-121.

[10] He, Yeye, and Jeffrey F. Naughton. "Anonymization of set-valued data via top-down, local generalization." Proceedings of the VLDB Endowment 2.1 (2009): 934-945.

[11] Hong, Yuan, et al. "Effective anonymization of query logs." Proceedings of the 18th ACM conference on Information and knowledge management. ACM, 2009.

[12] Zhu, Yun, Li Xiong, and Christopher Verdery. "Anonymizing user profiles for personalized web search." Proceedings of the 19th international conference on World wide web. ACM, 2010.

[13] Miller, George A. "WordNet: a lexical database for English." Communications of the ACM 38.11 (1995): 39-41.

[14] Barbaro, Michael, Tom Zeller, and Saul Hansell. "A face is exposed for AOL searcher no. 4417749." New York Times 9.2008 (2006): 8For.

[15] Navarro-Arribas, Guillermo, et al. "User k-anonymity for privacy preserving data mining of query logs." Information Processing & Management 48.3 (2012): 476-487.