Towards Query Obfuscation Strategies for Information Retrieval

Francesco Luigi De Faveri[0009−0005−8968−9485]

University of Padova, Padova, Italy
francescoluigi.defaveri@phd.unipd.it

Abstract. Preserving privacy in Information Retrieval (IR) remains a significant issue for users when interacting with Information Retrieval Systems (IRSs). Conducting a private search when an IRS does not cooperate towards the protection of the user privacy can lead to unwanted information disclosure through the analysis of the queries sent to the system. Recent investigations in Natural Language Processing (NLP) and IR have adopted the use of ε-Differential Privacy (DP) to obfuscate the real information need contained in the user queries. Although privacy is protected from a formal point of view, such methods do not consider the fact that the obfuscations can be irrelevant if the lexical or semantic meaning of the obfuscated terms remains unchanged with respect to the real user text. This paper outlines the author's PhD research in designing new techniques based on ε-DP for preserving the real user information need when interacting with IRSs that aim to disclose private information.

Keywords: Privacy Preserving Information Retrieval · Differential Privacy · Information Retrieval · Information Security.

1 Motivation of the Research

As large amounts of data, including sensitive ones, are generated daily, the challenge of protecting user privacy becomes increasingly significant for the Information Retrieval (IR) community. A typical scenario of privacy exposure occurs when a user interacts with Information Retrieval Systems (IRSs) to issue sensitive queries when retrieving documents [13]. For example, when looking for medical information, the risk is that personally identifiable information is leaked, posing a significant threat to patient privacy. If sensitive information is revealed, malicious employers could exploit this knowledge to terminate employment before covering medical expenses related to the disease, thus harming the employee's welfare [17]. Another example of privacy violation is represented by *ego-surfing*, i.e., when a user searches for their name or social security number, and personalized advertisement: by linking the actual information need with the interests of a user, it is possible to produce tailored advertisements, exposing some of the personal sphere information like the sexual orientation of users [5].

The current State of the Art definition of privacy is the ε-Differential Privacy (DP) [9]. Different obfuscation mechanisms, i.e., algorithms that change

the original user texts [11, 19–21, 4, 2], have been released in NLP scenarios and adapted to IR tasks [10, 8]. With such algorithms, the embedding vector of a query term is changed with a specific amount of statistical noise, defined by the privacy budget ε, theoretically proving the privacy definition introduced in [9]. However, if the noise is too small, the mechanisms fail to mask the original word, thus missing concrete privacy for the user.

This paper outlines the author's PhD studies [8, 7] to provide robust privacy guarantees when users interact with IRSs. The main strategies analysed concern the obfuscation of the real information need in the queries submitted to the system, balancing the trade-off between effectiveness and privacy obtained.

2 Background and Related Work

Text obfuscation mechanisms are divided into two main approaches: the *Heuristics* methodologies or *ε-Differential Privacy (DP)*.

Heuristics IR Privacy Approaches. Different heuristics methods have been proposed [1, 12], specifically for IR tasks. Arampatzis et al. [1] employ WordNet [15] to substitute original query terms within the original text using synonyms, hypernyms, and holonyms. The obfuscation is performed based on a hierarchical degree aligned with the user's desired obfuscation. The concept of privacy, in this case, is represented by the idea that if the original term is leaked in the obfuscated query, privacy is broken; therefore, the solution to preserve the utility of the task is to substitute the original term with a generalization. Such an approach is extended by Fröbe et al. [12]. The obfuscation method retrieves locally the top-k documents from a local corpus. Then, using a sliding window parameter, the sequences of n terms within the documents are considered candidate obfuscation queries. Unlike Arampatzis et al., [1], those queries that include synonyms and holonyms are discarded. The top-k documents retrieved on the local corpus are considered pseudo-relevant, and finally, the queries to submit are selected based on the nDCG@10 achieved.

ε-Differential Privacy Approaches. Dwork et al. [9] introduced the ε-DP framework to formalize the privacy guarantees when releasing data. Given a privacy budget $\varepsilon \in \mathbb{R}^+$, and any pair of neighbouring datasets D, D', i.e., datasets that differ for only one entry, an obfuscation mechanism $\mathcal{M}$, i.e., an algorithm that takes a dataset D and returns a randomized version D', is DP if it holds the inequality $\Pr[\mathcal{M}(D) \in S] \leq e^{\varepsilon} \cdot \Pr[\mathcal{M}(D') \in S] \;\; \forall S \subset \mathrm{Im}(\mathcal{M})$. DP introduces calibrated noise levels during output computation using the privacy budget ε, which controls the balance between data privacy and utility. The DP framework for metric spaces, and therefore for NLP tasks, was proposed in [3]. Metric-DP relaxes the traditional DP definition by ensuring that the probability of obfuscating two distinct points x, x' in the metric space is proportional to the distance $d(x, x')$ between them. Therefore, different strategies based on noisy sampling [4, 21, 2, 8] and perturbed word embeddings [11, 19, 20] was introduced to achieve

formal privacy guarantees during textual obfuscation. The idea behind the former strategies modifies the singular word embedding independently. The vectors are obtained, for example, from pre-computed vectors like the GloVe ones [16], and using a metric function, the mechanism finds new terms to use as obfuscation candidates. On the other hand, the sampling embedding strategies use the DP property to model a probability distribution to sample the new obfuscated query terms for the query.

3 Proposed Research

The proposed research involves the development of new mechanisms to be used in a search obfuscation protocol. In a search obfuscation protocol, the user generates N obfuscations of the original queries and submits these obfuscated queries to an IRS to obtain relevant documents. Upon receiving the documents, the user conducts a secure reranking process using the original query on their end.

The main research questions (RQ1-3) are:

RQ1. Can the approach achieve effective query obfuscation that maintains strong privacy guarantees, preventing possible obfuscation failures?
RQ2. Does the obfuscation induced by such a method effectively allow the retrieval of relevant documents, even with strong privacy settings? To what extent do the obfuscations preserve the user utility of search results?
RQ3. How can the concrete privacy obtained be evaluated? Can we rely only on analyzing the mechanism's parameters and theoretical guarantees?

4 Research Methodology and Proposed Experiments

Current state of the art obfuscation mechanisms either preserve the privacy of the obfuscated queries by providing formal privacy via the DP framework or account for the presence of synonyms and holonyms. The Words Blending Boxes (WBB) mechanism [8] bridges the gap between these obfuscation strategies. The mechanism controls that the top-k most similar words, i.e., synonyms and holonyms that hang closer to the original term embedding, are excluded from the obfuscation process. Moreover, the mechanism uses as obfuscation candidates the n similar words outside the top similar k, sampling the final obfuscation term according to the DP exponential mechanism [14] that models the probability of sampling a certain word depending on the privacy parameter ε.

To test the proposed WBB mechanism, we employed two TREC Collections, the MSMARCO Deep Learning'19 (DL'19) [6] and the Robust'04 [18]. The embedding function used is represented by the precomputed vectors of GloVe [16] with 300 features per array. We developed the open-source framework pyPANTERA [7] to implement the ε-DP state-of-the-art mechanisms[1] used as com-

[1] Here we report the heuristics – Arampatzis et al. (AEA) [1] and Fröbe et al. (FEA) [12] – and DP mechanisms – Cumulative Multivariate Perturbations (CMP) [11] and Mahalanobis (Mhl) [19] for noisy embedding based, Customized Text (CusText) [4] and Sanitization Text (SanText) [21] for sampling-based.

parison baselines and also the implementation of the WBB mechanism. We test the obfuscated query variants with sparse bag-of-word (BM25 and TF-IDF) and neural retrieval models (Contriever and TAS-B) to study the effectiveness of the search process. We report the results[2] of the nDCG@10 achieved in Table 1.

Table 1. Mean nDCG@10 achieved pooling the documents retrieved and reranked using the Contriever model. The WBB(k, n) mechanism is parametrized using the cosine similarity function.

IR System	Strategy	Mechanism	Robust'04					DL'19				
			ε - Privacy Budget					ε - Privacy Budget				
			1	5	15	50	No-DP	1	5	15	50	No-DP
BM25	*Original*	No-Privacy	-	-	-	-	0.477	-	-	-	-	0.675
	Heuristics	AEA	-	-	-	-	0.423	-	-	-	-	0.557
		FEA	-	-	-	-	0.147	-	-	-	-	0.069
	DP	CMP	0.000	0.002	0.338	0.421	-	0.000	0.000	0.293	0.403	-
	Embeddings	Mhl	0.000	0.037	0.204	0.421	-	0.000	0.000	0.154	0.403	-
	DP	CusText	0.143	0.157	0.345	0.371	-	0.131	0.154	0.357	0.401	-
	Sampling	SanText	0.028	0.374	0.375	0.375	-	0.002	0.399	0.401	0.401	-
	Our Method	WBB (2,20)	0.092	0.103	0.117	0.116	-	0.230	0.230	0.215	0.236	-
		WBB (4,15)	0.092	0.106	0.104	0.111	-	0.225	0.201	0.239	0.213	-
Contriever	*Original*	No-Privacy	-	-	-	-	0.466	-	-	-	-	0.676
	Heuristics	AEA	-	-	-	-	0.430	-	-	-	-	0.567
		FEA	-	-	-	-	0.200	-	-	-	-	0.056
	DP	CMP	0.000	0.002	0.373	0.466	-	0.000	0.000	0.397	0.598	-
	Embeddings	Mhl	0.000	0.001	0.211	0.466	-	0.000	0.000	0.178	0.597	-
	DP	CusText	0.113	0.175	0.400	0.408	-	0.230	0.309	0.542	0.560	-
	Sampling	SanText	0.000	0.402	0.404	0.407	-	0.000	0.559	0.563	0.569	-
	Our Method	WBB (2,20)	0.460	0.451	0.449	0.444	-	0.597	0.623	0.604	0.603	-
		WBB (4,15)	0.450	0.447	0.432	0.442	-	0.599	0.616	0.606	0.611	-

Table 1 illustrates the impact of the analyzed obfuscation strategies on retrieval utility (nDCG@10) across different settings. The WBB mechanism consistently performs better in balancing privacy and utility, especially on the DL'19 dataset with the Contriever IRS. While noisy embedding methods show notable performance degradation under strict privacy settings, noisy sampling techniques offer comparatively good results. However, with the BM25 retrieval model, CMP and Mhl emerge as more robust strategies, mainly at higher privacy budgets ε.

One of the limitations of the WBB mechanism is the fact that across different ε parametrization, the performance in the downstream task remains bounded below the performance of the *no privacy* scenario probably due to the size of the candidate obfuscation term pool.

[2] We refer for the full experimental pipeline to the key references [8, 7].

5 Research Issues for the Doctoral Consortium

During the Doctoral Consortium, the main questions (DC1-3) that require discussion with experienced researchers are:

DC1 The proposed method shows important limitations when the queries are submitted to IRS like BM25 and TF-IDF. Which possible solution strategies could limit these adverse effects?

DC2 The obfuscation process (of any obfuscation method) does not consider the final performance of the search. Still, it only takes into account the text of the query. Which techniques can be employed towards this direction?

DC3 The evaluation of the actual privacy provided to the user queries remains an open issue, and it is still evaluated by tuning the formal privacy budget of the DP mechanisms and observing the loss in terms of performance. What key factors should be considered when measuring the actual privacy in DP mechanisms beyond the formal ones?

References

1. Arampatzis, A., Drosatos, G., Efraimidis, P.: A versatile tool for privacy-enhanced web search. In: Serdyukov, P., Braslavski, P., Kuznetsov, S.O., Kamps, J., Rüger, S.M., Agichtein, E., Segalovich, I., Yilmaz, E. (eds.) Advances in Information Retrieval - 35th European Conference on IR Research, ECIR 2013, Moscow, Russia, March 24-27, 2013. Proceedings. Lecture Notes in Computer Science, vol. 7814, pp. 368–379. Springer (2013). https://doi.org/10.1007/978-3-642-36973-5_31, https://doi.org/10.1007/978-3-642-36973-5_31
2. Carvalho, R.S., Vasiloudis, T., Feyisetan, O., Wang, K.: TEM: high utility metric differential privacy on text. In: Shekhar, S., Zhou, Z., Chiang, Y., Stiglic, G. (eds.) Proceedings of the 2023 SIAM International Conference on Data Mining, SDM 2023, Minneapolis-St. Paul Twin Cities, MN, USA, April 27-29, 2023. pp. 883–890. SIAM (2023). https://doi.org/10.1137/1.9781611977653.CH99, https://doi.org/10.1137/1.9781611977653.ch99
3. Chatzikokolakis, K., Andrés, M.E., Bordenabe, N.E., Palamidessi, C.: Broadening the scope of differential privacy using metrics. In: Cristofaro, E.D., Wright, M.K. (eds.) Privacy Enhancing Technologies - 13th International Symposium, PETS 2013, Bloomington, IN, USA, July 10-12, 2013. Proceedings. Lecture Notes in Computer Science, vol. 7981, pp. 82–102. Springer (2013). https://doi.org/10.1007/978-3-642-39077-7_5, https://doi.org/10.1007/978-3-642-39077-7_5
4. Chen, S., Mo, F., Wang, Y., Chen, C., Nie, J.Y., Wang, C., Cui, J.: A customized text sanitization mechanism with differential privacy. In: Rogers, A., Boyd-Graber, J., Okazaki, N. (eds.) Findings of the Association for Computational Linguistics: ACL 2023. pp. 5747–5758. Association for Computational Linguistics, Toronto, Canada (Jul 2023). https://doi.org/10.18653/v1/2023.findings-acl.355, https://aclanthology.org/2023.findings-acl.355
5. Cohn, J.: My tivo thinks i'm gay: Algorithmic culture and its discontents. Television & New Media **17**(8), 675–690 (2016). https://doi.org/10.1177/1527476416644978, https://doi.org/10.1177/1527476416644978

6. Craswell, N., Mitra, B., Yilmaz, E., Campos, D., Voorhees, E.M.: Overview of the TREC 2019 deep learning track. CoRR **abs/2003.07820** (2020), https://arxiv.org/abs/2003.07820
7. De Faveri, F.L., Faggioli, G., Ferro, N.: py-PANTERA: A Python PAckage for Natural language obfuscaTion Enforcing pRivacy & Anonymization. In: Proceedings of the 33rd ACM International Conference on Information and Knowledge Management (CIKM '24), October 21–25, 2024, Boise, ID, USA. p. 6. Springer (2024). https://doi.org/10.1145/3627673.3679173, https://doi.org/10.1145/3627673.3679173
8. De Faveri, F.L., Faggioli, G., Ferro, N.: Words Blending Boxes. Obfuscating Queries in Information Retrieval using Differential Privacy. CoRR **abs/2405.09306** (2024). https://doi.org/10.48550/ARXIV.2405.09306, https://doi.org/10.48550/arXiv.2405.09306
9. Dwork, C., McSherry, F., Nissim, K., Smith, A.: Calibrating noise to sensitivity in private data analysis. In: Halevi, S., Rabin, T. (eds.) Theory of Cryptography. pp. 265–284. Springer Berlin Heidelberg, Berlin, Heidelberg (2006)
10. Faggioli, G., Ferro, N.: Query obfuscation for information retrieval through differential privacy. In: Goharian, N., Tonellotto, N., He, Y., Lipani, A., McDonald, G., Macdonald, C., Ounis, I. (eds.) Advances in Information Retrieval - 46th European Conference on Information Retrieval, ECIR 2024, Glasgow, UK, March 24-28, 2024, Proceedings, Part I. Lecture Notes in Computer Science, vol. 14608, pp. 278–294. Springer (2024). https://doi.org/10.1007/978-3-031-56027-9_17, https://doi.org/10.1007/978-3-031-56027-9_17
11. Feyisetan, O., Balle, B., Drake, T., Diethe, T.: Privacy- and utility-preserving textual analysis via calibrated multivariate perturbations. In: Caverlee, J., Hu, X.B., Lalmas, M., Wang, W. (eds.) Proceedings of the 13th International Conference on Web Search and Data Mining. pp. 178–186. ACM (Jan 2020). https://doi.org/10.1145/3336191.3371856
12. Fröbe, M., Schmidt, E.O., Hagen, M.: Efficient query obfuscation with keyqueries. In: He, J., Unland, R., Jr., E.S., Tao, X., Purohit, H., van den Heuvel, W., Yearwood, J., Cao, J. (eds.) WI-IAT '21: IEEE/WIC/ACM International Conference on Web Intelligence, Melbourne VIC Australia, December 14 - 17, 2021. pp. 154–161. ACM (2021). https://doi.org/10.1145/3486622.3493950, https://doi.org/10.1145/3486622.3493950
13. Houssiau, F., Liénart, T., Hendrickx, J.M., de Montjoye, Y.: Web privacy: A formal adversarial model for query obfuscation. IEEE Trans. Inf. Forensics Secur. **18**, 2132–2143 (2023). https://doi.org/10.1109/TIFS.2023.3262123, https://doi.org/10.1109/TIFS.2023.3262123
14. McSherry, F., Talwar, K.: Mechanism design via differential privacy. In: 48th Annual IEEE Symposium on Foundations of Computer Science (FOCS 2007), October 20-23, 2007, Providence, RI, USA, Proceedings. pp. 94–103. IEEE Computer Society (2007). https://doi.org/10.1109/FOCS.2007.41, https://doi.org/10.1109/FOCS.2007.41
15. Miller, G.A.: Wordnet: A lexical database for english. Commun. ACM **38**(11), 39–41 (1995). https://doi.org/10.1145/219717.219748, https://doi.org/10.1145/219717.219748
16. Pennington, J., Socher, R., Manning, C.D.: Glove: Global Vectors for Word Representation. In: Moschitti, A., Pang, B., Daelemans, W. (eds.) Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest

Group of the ACL. pp. 1532–1543. ACL (2014). https://doi.org/10.3115/v1/d14-1162, https://doi.org/10.3115/v1/d14-1162

17. Ragan, E.D., Kum, H., Ilangovan, G., Wang, H.: Balancing privacy and information disclosure in interactive record linkage with visual masking. In: Mandryk, R.L., Hancock, M., Perry, M., Cox, A.L. (eds.) Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems, CHI 2018, Montreal, QC, Canada, April 21-26, 2018. p. 326. ACM (2018). https://doi.org/10.1145/3173574.3173900, https://doi.org/10.1145/3173574.3173900
18. Voorhees, E.M.: Overview of the TREC 2004 robust track. In: Voorhees, E.M., Buckland, L.P. (eds.) Proceedings of the Thirteenth Text REtrieval Conference, TREC 2004, Gaithersburg, Maryland, USA, November 16-19, 2004. NIST Special Publication, vol. 500-261. National Institute of Standards and Technology (NIST) (2004), http://trec.nist.gov/pubs/trec13/papers/ROBUST.OVERVIEW.pdf
19. Xu, Z., Aggarwal, A., Feyisetan, O., Teissier, N.: A differentially private text perturbation method using regularized mahalanobis metric. In: Proceedings of the Second Workshop on Privacy in NLP. Association for Computational Linguistics (2020). https://doi.org/10.18653/v1/2020.privatenlp-1.2
20. Xu, Z., Aggarwal, A., Feyisetan, O., Teissier, N.: On a utilitarian approach to privacy preserving text generation. CoRR **abs/2104.11838** (Apr 2021). https://doi.org/10.48550/ARXIV.2104.11838
21. Yue, X., Du, M., Wang, T., Li, Y., Sun, H., Chow, S.S.M.: Differential privacy for text analytics via natural text sanitization. In: Zong, C., Xia, F., Li, W., Navigli, R. (eds.) Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021. pp. 3853–3866. Association for Computational Linguistics, Online (Aug 2021). https://doi.org/10.18653/v1/2021.findings-acl.337, https://aclanthology.org/2021.findings-acl.337