Topic Model based Privacy Protection in Personalized Web Search

Wasi Uddin Ahmad, Md Masudur Rahman, Hongning Wang
{wua4nw, mr5ba, hw5x}@virginia.edu

Department of Computer Science, University of Virginia, VA USA

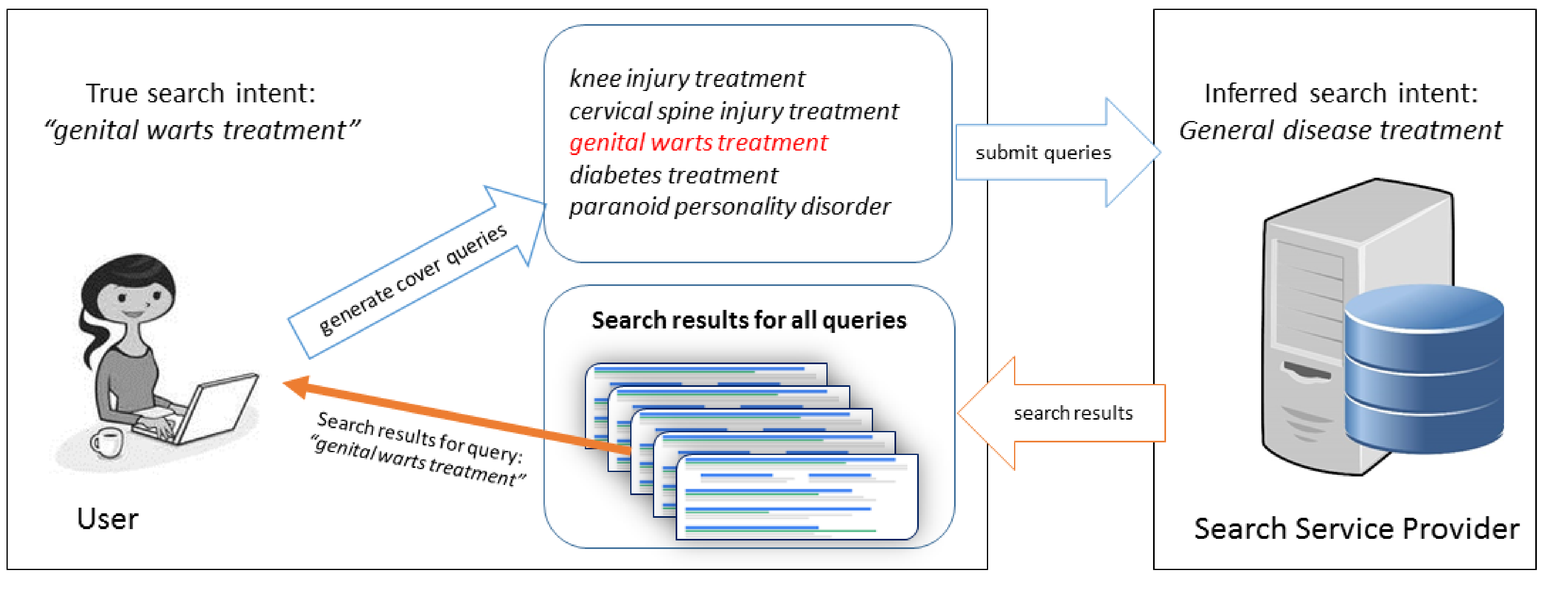
Objective

Personalized search has a potential risk of revealing users' privacy by identifying their underlying intention from their logged search behaviors.

- Google and Facebook privacy issues in Europe
- AOL search logs release in 2006

Our aim is to protect user privacy by injecting controlled noise to users' search log while still providing necessary utility for the search engine to perform personalization.

Topic-based Privacy Protection (TPP) Framework



Methodology

- Used probabilistic topic models to infer search intent from their issued queries.
- Injected k cover queries with the original query
- Re-ranked the **original query's results** based on the user profile constructed and maintained on the client side.

Cover Query Generation

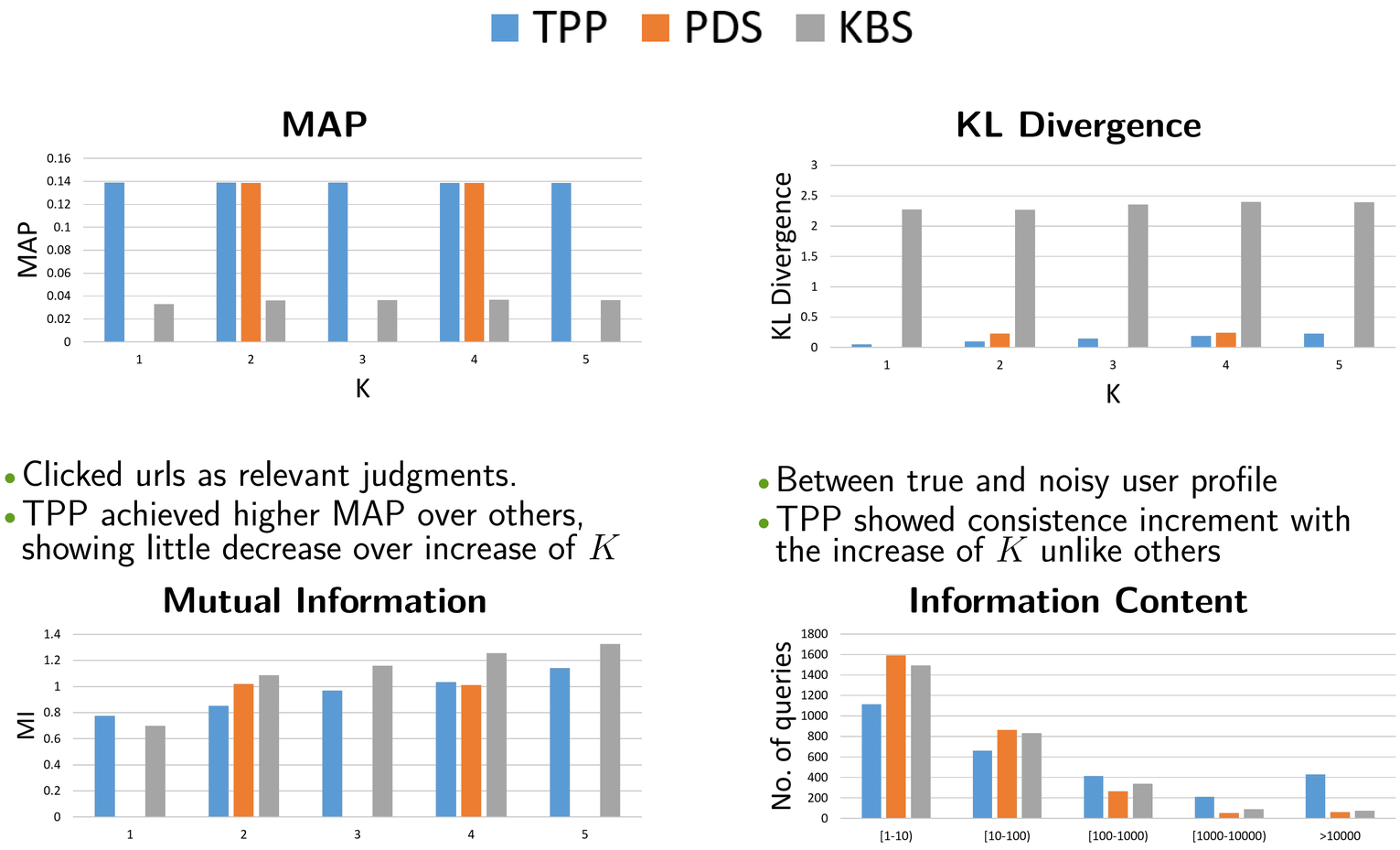
- Generated k cover queries with similar entropy to the original query randomly based on the inferred topics by LDA topic model trained on BBC news data set.
- Example: If the query is highly concentrated in sports, fewer cover queries will be generated from the topic of sports but more from business, entertainment etc.
- Improved the plausibility by randomizing the length of cover queries using Poisson distribution with target user's average query length as rate parameter

Improving Search Effectiveness

- Built user profiles, using language model, on client-side with users' true queries and click documents for search result re-ranking.
- Calculated client side score for each returned documents, only for the **true user query,** using true user profile.
- Re-ranked the returned documents based on a linear combination of client side score and server side ranking score

Experimental Results

- Dataset: AOL search logs - 45,200 queries over 250 users.
- Baselines: Plausible Deniable Search (PDS) [1] and Knowledge-based Scheme (KBS) [2]
- *PDS cannot generate result for k = 1,3 and 5, indicated by zero or none



- Clicked urls as relevant judgments.
- TPP achieved higher MAP over others, showing little decrease over increase of K

- Between true and noisy user profile
- TPP showed consistence increment with the increase of K unlike others

- Between original and cover query set
- TPP and KBS showed consistence increment with the increase of K

- Between the original and corresponding cover query
- TPP generated greater number of queries of higher IC compared to PDS and KBS

Conclusion

- Novel solution to protect user privacy based on their inferred search intent and provide personalized search

Acknowledgements

We would like to thank the anonymous reviewers for their insightful comments. This paper is based upon work supported in part by a Yahoo Academic Career Enhancement Award and the National Science Foundation under grants IIS-1553568. We would also like to thank SIGIR for providing *SIGIR Student Travel Grant* to facilitate the participation in this conference.

References

[1] M. Murugesan and C. Clifton.
Providing privacy through plausibly deniable search.
In *SDM*, pages 768–779. SIAM, 2009.

[2] D. Sánchez, J. Castellà-Roca, and A. Viejo.
Knowledge-based scheme to create

UNIVERSITY of VIRGINIA