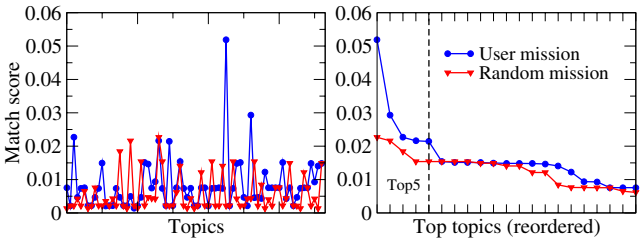

Figure 9: Example of match of missions sequences on a topical user profile. On the x axis are the topics in the profile; on the y axis are the match scores of the mission sequences for each topic. The sequence originated by the same user has more spikes compared to the sequence of the random user. When the values of the each curve are sorted, the best matching sequence is evident by looking at the top-N scores.

user we select two sequences of missions from the 30 days right after the 3 months considered. One sequence comprises of all the missions generated by the same user in the 30-day period. The other is a sequence of equal length generated by another user chosen at random among all other users.

Intuitively, a mission is likely to match at most a few of the several, possibly different topics in a user profile. Given this intuition, to decide which of the two sequences best matches the profile, we focus on the top-N ($N = 5$) elements of the match vectors between topic profile and mission sequences. We then apply a simple majority rule, i.e., which sequence has the most elements with higher match scores. This idea is exemplified by Figure 9.

Using this technique we are able to detect the user's own sequence in 65% of the cases. We stress that the mission sequences of randomly selected users are strongly biased toward high-frequency queries such as "facebook," "amazon," and so on. Since these are shared by a large number of users, any user profile is likely to match them, leading to a decay in detection performance. For this reason we divided the random sequences into three sets according to the average frequency of their queries. The accuracy rises to 72% when considering the sequences with lower frequency queries and drops to around 55% when considering the sequences with higher frequency queries. 65% success in prediction can be considered a good result. Even if the interest of a given user would presumably be quite stationary *within a particular domain*, in web search, where a much wider range of suitable topics is available, the user focus can be often inconstant in time, independent by past search sessions and made even more variable by bursty search activity triggered by external events (e.g., "Michael Jackson death"), hardly predictable by looking only at the user history. This issues make this prediction task much more difficult compared to domain-specific predicition or recomendation.

9. CONCLUSIONS

The *behavior* of the users in submitting queries to a search engine, including the implicit and explicit information that their actions reveal about their search intent, is a crucial element to determine what is the *topic* of a query or of a sequence of search actions. We introduce a novel definition of topic in the context of query log analysis and propose a topic extraction algorithm based on agglomerative clustering of sequences of queries that exhibit a coherent user intent. Our algorithm relies on a semi-supervised classifier that can tell if two query sets are topically coherent with excellent accuracy (AUC 0.95). We compare our method with a graph-based clustering baseline, showing its advantages on query coverage and on the trade-off between purity and resource coverage of the clusters. Finally, we define the *topical profile* of a user in terms of a topic vector that best defines the user search history. With our classifier we are able to discriminate a query sequence submitted by the profiled user from a random query sequence 72% of the time in the best case scenario.

One could consider more sophisticated baselines, for instance combining click co-occurrence with lexical similarity features, or even using query clustering algorithms based on alternative paradigms [7, 18, 11]. A more direct evaluation could be achieved with the golden standard of human judgments on the quality of the topics. A preliminary effort in this direction has pointed to the difficulty of human inspection of topic quality, as well as the challenge of identifying a suitable tradeoff between topic specificity and coverage.

This work opens several research directions. In particular, we are working on the formulation of a user-to-user similarity metric based on topics that can overcome the sparsity problem of similarity metrics based on exact query matches. Finally, we want to explore in greater depth the potential of the topical profiling technique to predict future search activity and provide novel search recommendation services.

10. REFERENCES

[1] P. Anick. Using terminological feedback for web search refinement: a log-based study. In *SIGIR '03*, 2003.

[2] R. Baeza-Yates. Graphs from search engine queries. In *SOFSEM'07: 33rd conference on Current Trends in Theory and Practice of Computer Science*, pages 1–8, Berlin, Heidelberg, 2007. Springer-Verlag.

[3] R. Baeza-Yates and A. Tiberi. Extracting semantic relations from query logs. In *KDD'07: 13th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 76–85, New York, NY, USA, 2007. ACM.

[4] N. Balasubramanian and S. Cucerzan. Automatic generation of topic pages using query-based aspect models. In *CIKM'09: 18th ACM conference on Information and knowledge management*, pages 2049–2052, New York, NY, USA, 2009. ACM.

[5] A. L. Barabasi. The origin of bursts and heavy tails in human dynamics. *Nature*, 435:207, 2005.

[6] D. Beeferman and A. Berger. Agglomerative clustering of a search engine query log. In *KDD'00: 6th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 407–416, New York, NY, USA, 2000. ACM.

[7] S. M. Beitzel, E. C. Jensen, D. D. Lewis, A. Chowdhury, and O. Frieder. Automatic classification of web queries using very large unlabeled query logs. *ACM Transactions on Information Systems*, 25, April 2007.

[8] P. Boldi, F. Bonchi, C. Castillo, D. Donato, A. Gionis, and S. Vigna. The query-flow graph: model and

applications. In *CIKM'08: Proceeding of the 17th ACM conference on Information and knowledge management*, pages 609–618, New York, NY, USA, 2008. ACM.

[9] I. Bordino, C. Castillo, D. Donato, and A. Gionis. Query similarity by projecting the query-flow graph. In *SIGIR'10*, pages 515–522. ACM, 2010.

[10] U. Brandes, D. Delling, M. Höfer, M. Gaertler, R. Görke, Z. Nikoloski, and D. Wagner. On Finding Graph Clusterings with Maximum Modularity. In *WG'07: Proceedings of the 33rd International Workshop on Graph-Theoretic Concepts in Computer Science*, Lecture Notes in Computer Science, 2007.

[11] H. Cao, D. Jiang, J. Pei, Q. He, Z. Liao, E. Chen, and H. Li. Context-aware query suggestion by mining click-through and session data. In *KDD'08: 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 875–883, New York, NY, USA, 2008. ACM.

[12] S.-L. Chuang and L.-F. Chien. A practical web-based approach to generating topic hierarchy for text segments. In *CIKM'04: 13th ACM international conference on Information and knowledge management*, pages 127–136, New York, NY, USA, 2004. ACM.

[13] D. Donato, F. Bonchi, T. Chi, and Y. Maarek. Do you want to take notes?: identifying research missions in Yahoo! search pad. In *WWW'10: 19th international conference on World wide web*, pages 321–330, New York, NY, USA, 2010. ACM.

[14] L. Fitzpatrick and M. Dent. Automatic feedback using past queries: social searching? In *SIGIR '97*, 1997.

[15] S. Fortunato. Community detection in graphs. *Physics Reports*, 486(3-5):75–174, 2010.

[16] S. Fortunato and M. Barthèlemy. Resolution limit in community detection. *Proceedings of the National Academy of Sciences*, 104(1):36–41, Jan 2007.

[17] S. Gollapudi and R. Panigrahy. Exploiting asymmetry in hierarchical topic extraction. In *CIKM'06: 15th ACM international conference on Information and knowledge management*, pages 475–482, New York, NY, USA, 2006. ACM.

[18] X. He and P. Jhala. Regularized query classification using search click information. *Pattern Recognition*, 41:2283–2288, July 2008.

[19] T. Hofmann. Probabilistic latent semantic indexing. In *SIGIR'99*, pages 50–57, 1999.

[20] R. Jones and K. L. Klinkner. Beyond the session timeout: automatic hierarchical segmentation of search topics in query logs. In *CIKM'08: Proceeding of the 17th ACM conference on Information and knowledge management*, pages 699–708, New York, NY, USA, 2008. ACM.

[21] R. Jones and K. L. Klinkner. Beyond the session timeout: automatic hierarchical segmentation of search topics in query logs. In *CIKM '08: 17th ACM conference on Information and knowledge mining*, pages 699–708, New York, NY, USA, 2008. ACM.

[22] A. Lancichinetti, S. Fortunato, and F. Radicchi. Benchmark graphs for testing community detection algorithms. *Physical Review E (Statistical, Nonlinear, and Soft Matter Physics)*, 78(4), 2008.

[23] A. Lancichinetti, F. Radicchi, J. J. Ramasco, and S. Fortunato. Finding statistically significant communities in networks. *PLoS ONE*, 6(4), 04 2011.

[24] J. Leskovec, K. J. Lang, A. Dasgupta, and M. W. Mahoney. Statistical properties of community structure in large social and information networks. In *WWW'08: Proceedings of the 17th international conference on World Wide Web*, pages 695–704, New York, NY, USA, 2008. ACM.

[25] M. Molloy and B. Reed. A critical point for random graphs with a given degree sequence. *Random Structures & Algorithms*, 6:161–179, March 1995.

[26] M. E. J. Newman and M. Girvan. Finding and evaluating community structure in networks. *Physical Review E*, 69, 2004.

[27] F. Radlinski and T. Joachims. Query chains: learning to rank from implicit feedback. In *KDD'05: 11th ACM SIGKDD international conference on Knowledge discovery in data mining*, pages 239–248, New York, NY, USA, 2005. ACM.

[28] V. V. Raghavan and H. Sever. On the reuse of past optimal queries. In *SIGIR '95*, 1995.

[29] G. Salton and M. J. McGill. *Introduction to Modern Information Retrieval*. McGraw-Hill, Inc., New York, NY, USA, 1986.

[30] L. Sarmento, A. Kehlenbeck, E. Oliveira, and L. Ungar. Efficient clustering of web-derived data sets. In *MLDM '09: 6th International Conference on Machine Learning and Data Mining in Pattern Recognition*, pages 398–412, Berlin, Heidelberg, 2009. Springer-Verlag.

[31] W. Song, Y. Zhang, T. Liu, and S. Li. Bridging topic modeling and personalized search. In *COLING'10: 23rd International Conference on Computational Linguistics: Posters*, pages 1167–1175, Stroudsburg, PA, USA, 2010. Association for Computational Linguistics.

[32] A. Veilumuthu and P. Ramachandran. Intent based clustering of search engine query log. In *CASE'09: 5th IEEE international conference on Automation science and engineering*, pages 647–652, Piscataway, NJ, USA, 2009. IEEE.

[33] J.-R. Wen, J.-Y. Nie, and Z. Hong-Jiang. Query clustering using user logs. *ACM Transactions on Information Systems*, 20:59–81, January 2002.

[34] I. H. Witten and E. Frank. *Data mining: practical machine learning tools and techniques with Java implementations*. Morgan Kaufmann, 1999.

[35] J. Ye, J.-H. Chow, J. Chen, and Z. Zheng. Stochastic gradient boosted distributed decision trees. In *CIKM'09: 18th ACM conference on Information and knowledge management*, pages 2061–2064, New York, NY, USA, 2009. ACM.

[36] J. Yi and F. Maghoul. Query clustering using click-through graph. In *WWW'09: 18th international conference on World wide web*, pages 1055–1056, New York, NY, USA, 2009. ACM.

[37] H.-J. Zeng, Q.-C. He, Z. Chen, W.-Y. Ma, and J. Ma. Learning to cluster web search results. In *SIGIR'04: 27th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 210–217, New York, NY, USA, 2004. ACM.