Short Texts Similarity: A Survey

ABSTRACT

1. INTRODUCTION

if two queries share a set of same clicked URLs, they will convey similar search intent [2, 3]. The wisdom of the crowds property thus makes clickthrough, especially the query co-clicks, a more precise resource for identifying similar search intent of queries. However, query co-click information is relative sparse (i.e. lower recall) in describing search intents compared with the search result snippets, since usually there are limited clicks for each query.

2. OVERVIEW

2.1 Similarity and relatedness

2.2 Major approaches

surface
corpus
user behavior
knowledget-base

2.3 Major procedures

Augment the data
Representation
Similarity function
Query Efficiency
Evaluation

3. WORD-TO-WORD SIMILARITY

4. AUGMENT THE DATA

5. REPRESENTATION

6. SIMILARITY FUNCTION

Proceedings of the VLDB Endowment, Vol. 5, No. 0

6.1 pair-wise

6.2 graph-based

6.3 learning a similarity function

7. QUERY EFFICIENCY

8. EVALUATION

9. CONCLUSION

[13, 25, 32, 17, 26, 23, 31, 22, 24, 14, 30, 12, 18, 29, 27, 16, 11, 9, 15, 21, 19, 4, 28, 33, 5, 1, 10, 20, 6, 7, 8]

10. REFERENCES

[1] I. Antonellis, H. G. Molina, and C. C. Chang. Simrank++: Query rewriting through link analysis of the click graph. *Proceedings of the VLDB Endowment*, 1(1):408421, 2008.

[2] R. Baeza-Yates, C. Hurtado, and M. Mendoza. Query recommendation using query logs in search engines. In *Current Trends in Database Technology-EDBT 2004 Workshops*, page 395397, 2005.

[3] D. Beeferman and A. Berger. Agglomerative clustering of a search engine query log. In *Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining*, page 407416, 2000.

[4] D. Beeferman and A. Berger. Agglomerative clustering of a search engine query log. In *Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining*, page 407416, 2000.

[5] P. Boldi, F. Bonchi, C. Castillo, D. Donato, A. Gionis, and S. Vigna. The query-flow graph: model and applications. In *Proceedings of the 17th ACM conference on Information and knowledge management*, page 609618, 2008.

[6] P. Boldi, F. Bonchi, C. Castillo, D. Donato, A. Gionis, and S. Vigna. The query-flow graph: model and applications. In *Proceedings of the 17th ACM conference on Information and knowledge management*, page 609618, 2008.

[7] P. Boldi, F. Bonchi, C. Castillo, D. Donato, and S. Vigna. Query suggestions using query-flow graphs. In *Proceedings of the 2009 workshop on Web Search Click Data*, page 5663, 2009.

[8] I. Bordino, C. Castillo, D. Donato, and A. Gionis. Query similarity by projecting the query-flow graph. In *Proceeding of the 33rd international ACM SIGIR conference on Research and development in information retrieval*, page 515522, 2010.

[9] C. Burges, T. Shaked, E. Renshaw, A. Lazier, M. Deeds, N. Hamilton, and G. Hullender. Learning to rank using gradient descent. In *Proceedings of the 22nd international conference on Machine learning*, page 8996, 2005.

[10] N. Craswell and M. Szummer. Random walks on the click graph. In *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, page 239246, 2007.

[11] L. Fitzpatrick and M. Dent. Automatic feedback using past queries: social searching? In *ACM SIGIR Forum*, volume 31, page 306313, 1997.

[12] J. Guo, X. Cheng, G. Xu, and X. Zhu. Intent-aware query similarity. In *Proceedings of the 20th ACM international conference on Information and knowledge management*, page 259268, 2011.

[13] V. Hatzivassiloglou, J. Klavans, and E. Eskin. Detecting text similarity over short passages: Exploring linguistic feature combinations via machine learning. In *Proceedings of the 1999 joint sigdat conference on empirical methods in natural language processing and very large corpora*, page 203212, 1999.

[14] A. Islam and D. Inkpen. Semantic text similarity using corpus-based word similarity and string similarity. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 2(2):10, 2008.

[15] R. Jones, B. Rey, O. Madani, and W. Greiner. Generating query substitutions. In *Proceedings of the 15th international conference on World Wide Web*, page 387396, 2006.

[16] T. K. Landauer, P. W. Foltz, and D. Laham. An introduction to latent semantic analysis. *Discourse processes*, 25(2-3):259284, 1998.

[17] M. D. Lee, B. M. Pincombe, and M. B. Welsh. An empirical evaluation of models of text document similarity. *Cognitive Science*, 2005.

[18] M. Lesk. Automatic sense disambiguation using machine readable dictionaries: how to tell a pine cone from an ice cream cone. In *Proceedings of the 5th annual international conference on Systems documentation*, page 2426, 1986.

[19] H. Ma, H. Yang, I. King, and M. R. Lyu. Learning latent semantic relations from clickthrough data for query suggestion. In *Proceeding of the 17th ACM conference on Information and knowledge management*, page 709718, 2008.

[20] H. Ma, H. Yang, I. King, and M. R. Lyu. Learning latent semantic relations from clickthrough data for query suggestion. In *Proceeding of the 17th ACM conference on Information and knowledge management*, page 709718, 2008.

[21] Q. Mei, D. Zhou, and K. Church. Query suggestion using hitting time. In *Proceedings of the 17th ACM conference on Information and knowledge management*, page 469478, 2008.

[22] D. Metzler, S. Dumais, and C. Meek. Similarity measures for short segments of text. *Advances in Information Retrieval*, page 1627, 2007.

[23] R. Mihalcea, C. Corley, and C. Strapparava. Corpus-based and knowledge-based measures of text semantic similarity. In *Proceedings of the national conference on artificial intelligence*, volume 21, page 775, 2006.

[24] J. O'Shea, Z. Bandar, K. Crockett, and D. McLean. A comparative study of two short text semantic similarity measures. In *Proceedings of the 2nd KES International conference on Agent and multi-agent systems: technologies and applications*, page 172181, 2008.

[25] P. Resnik. Semantic similarity in a taxonomy: An information-based measure and its application to problems of ambiguity in natural language. *Arxiv preprint arXiv:1105.5444*, 2011.

[26] M. Sahami and T. D. Heilman. A web-based kernel function for measuring the similarity of short text snippets. In *Proceedings of the 15th international conference on World Wide Web*, page 377386, 2006.

[27] P. Turney et al. Mining the web for synonyms: PMI-IR versus LSA on TOEFL. 2001.

[28] J. R. Wen, J. Y. Nie, and H. J. Zhang. Clustering user queries of a search engine. In *Proceedings of the 10th international conference on World Wide Web*, page 162168, 2001.

[29] Z. Wu and M. Palmer. Verbs semantics and lexical selection. In *Proceedings of the 32nd annual meeting on Association for Computational Linguistics*, page 133138, 1994.

[30] J. Xu and G. Xu. Learning similarity function for rare queries. In *Proceedings of the fourth ACM international conference on Web search and data mining*, page 615624, 2011.

[31] W. T. Yih and C. Meek. Improving similarity measures for short segments of text. In *Proceedings Of The National Conference On Artificial Intelligence*, volume 22, page 1489, 2007.

[32] S. Zelikovitz and H. Hirsh. Improving short text classification using unlabeled background knowledge to assess document similarity. In *Proceedings of the Seventeenth International Conference on Machine Learning*, page 11831190, 2000.

[33] Z. Zhang and O. Nasraoui. Mining search engine query logs for query recommendation. In *Proceedings of the 15th international conference on World Wide Web*, page 10391040, 2006.