©2015 IEEE. Personal use of this material is permitted. However, permission to reprint/republish this material for advertising or promotional purposes or for creating new collective works for resale or redistribution to servers or lists, or to reuse any copyrighted component of this work in other works must be obtained from the IEEE.

Querying of Geo-Textual Web Content: Concepts and Techniques

Christian S. Jensen Department of Computer Science Aalborg University Denmark csj@cs.aau.dk

TALK OUTLINE

Much has happened with the web over the past decade. In brief, a spatial, or geographical, web has emerge where content and users are associated with locations that are used in a wide range of services. Billions of queries are being processed by web search engines. According to one report, Google processed a daily average of between 5 and 6 billion queries in 2013 and 2014 [29]. A substantial fraction of these have local intent and target geo-textual web objects, i.e., points of interest with a web presence that have locations as well as textual descriptions, among possibly other attributes. One article reports that 53% of mobile searches on Bing have local intent [30], and another source reports that 20% of Google searches are related to location [19].

A prototypical geo-textual web query takes a user location and user-supplied keywords as arguments and returns geotextual objects that are spatially and textually relevant to these arguments. Due perhaps to the rich semantics of geographical space and its importance to our daily lives, many different kinds of useful geo-textual web query functionality may be envisioned.

For example, some queries aim to find a few near-by points of interest that each satisfy a users needs as indicated by query keywords, other queries aim to find a set of points of interest that collectively satisfy a users needs, and yet other queries aim to find densely populated regions that enable a user to conveniently explore different relevant points of interest.

Based on recent and ongoing work by the speaker and his colleagues, the talk presents key functionality, concepts, and techniques relating to the querying of geo-textual web content; it covers functionality that addresses different kinds of user intent; and it offers directions for the future development of keyword-based geo-textual web querying.

READINGS

The querying of geo-textual web content is a very active research area. Somebody who is new to the area and wants to understand key concepts may find the relatively short survey paper by Cao et al. [3] to be an appropriate starting point. This paper gives structure to the area and should give a good basis for selecting the papers to read next. Someone interested in indexing may find the experimental comparison paper by Chen et al. [11] to be a good starting point. In addition to its experimental comparisons, the paper offers brief explanations of all the techniques compared.

For convenience, the references section lists 40 papers that are interesting, that offer different perspectives on the area, and that address a range of interesting aspects. These papers cover the aspects covered in the talk and beyond, and they are good candidates for papers to read next. The two abovementioned papers were published in 2012 and 2013, and much has happened since. Thus, about half of the 40 papers listed were published after 2012.

ACKNOWLEDGMENTS

The author wishes to thank those of his colleagues with whom he has collaborated on the topics covered here: Xin Cao, Lisi Chen, Gao Cong, Ilkcan Keles, Siyuan Liu, Beng Chin Ooi, Qiang Qu, Simonas Šaltenis, Darius Šidlauskas, Anders Skovsgaard, Dingming Wu, and Man Lung Yiu.

The author's research is supported in part by a grant from the Obel Family Foundation.

REFERENCES

- Bøgh, K., Skovsgaard, A., and Jensen, C. S. 2013. GroupFinder: a new approach to top-k point-of-interest group retrieval. *PVLDB* 6, 12, 1226–1229.
- [2] Cao, X., Chen, L, Cong, G., Guan, J., Phan, N.-T., and Xiao, X. 2013. KORS: keyword-aware optimal route search system. *ICDE*. 1340–1343.
- [3] Cao, X., Chen, L, Cong, G., Jensen, C. S., Qu, Q., Skovsgaard, A., Wu, D., and Y, M. L. 2012. Spatial keyword querying *ER*. 16–29.
- [4] Cao, X., Cong, G., Guo, T., Jensen, C. S., and Ooi, B. C.. 2015. Efficient Processing of Spatial Group Keyword Queries *TODS 40*, 2, 45 pages, to appear.
- [5] Cao, X., Cong, G., and Jensen, C. S. 2010. Mining significant semantic locations from GPS data. *PVLDB 3*, 1, 1009–1020.
- [6] Cao, X., Cong, G., and Jensen, C. S. 2010. Retrieving topk prestige-based relevant spatial web objects. *PVLDB 3*, 1, 373–384.

- [7] Cao, X., Cong, G., Jensen, C. S., Ng, J. J., Ooi, B. C., Phan, N.-T., and Wu, D. 2012. SWORS: A system for the efficient retrieval of relevant spatial web objects. *PVLDB* 5, 12, 1914– 1917.
- [8] Cao, X., Cong, G., Jensen, C. S., and Ooi, B. C. 2011. Collective spatial keyword querying. *SIGMOD*. 373–384.
- [9] Cao, X., Cong, G., Jensen, C. S., and Yiu, M. L. 2014. Retrieving regions of interest for user exploration. *PVLDB* 7, 9, 733–744.
- [10] Chen, L., Cong, G., and Cao, X. 2013. An efficient query indexing mechanism for filtering geo-textual data. *SIGMOD*. 749–760.
- [11] Chen, L., Cong, G., Jensen, C. S., and Wu, D. 2013. Spatial keyword query processing: an experimental evaluation. *PVLDB* 6, 3, 217–228.
- [12] Chen, L., Cui, Y., Cong, and Cao, X. 2014. SOPS: a system for efficient processing of spatial-keyword publish/subscribe. *PVLDB* 7, 13, 1601–1604.
- [13] Chen, L., Lin, X., Hu, H., Jensen, C. S., and Xu, J. 2015. Answering why-not questions on spatial keyword top-k queries. *ICDE*.
- [14] Chen, Y.-Y., Suel, T., and Markowetz, A. 2006. Efficient query processing in geographic web search engines. SIG-MOD. 277–288.
- [15] Christoforaki, M., He, J., Dimopoulos, C., Markowetz, A., and Suel, T. 2011. Text vs. space: efficient geo-search query processing. *CIKM*. 423–432.
- [16] Cong, G., Jensen, C. S., and Wu, D. 2009. Efficient retrieval of the top-k most relevant spatial web objects. *PVLDB 2*, 1, 337–348.
- [17] Derczynski, L., Yang, B., Jensen, C. S. 2013. Towards context-aware search and analysis on social media data. *EDBT*. 137–142.
- [18] Felipe, I. D., Hristidis, V., and Rishe, N. 2008. Keyword search on spatial databases. *ICDE*. 656–665.
- [19] GOOGLE 2015. Google Places. Stats & Facts. sites.google. com/a/pressatgoogle.com/googleplaces/metrics.
- [20] Hage, C., Jensen, C. S., Pedersen, T. B., Speicys, L., and Timko, I. 2003. Integrated data management for mobile services in the real world. *VLDB*. 1019–1030.
- [21] Li, Z., Lee, K. C. K., Zheng, B., Lee, W.-C., Lee, D. L., and Wang, X. 2011. IR-tree: an efficient index for geographic document search. *TKDE* 23, 4, 585–599.
- [22] Lu, J., Lu, Y., and Cong, G. 2011. Reverse spatial and textual k nearest neighbor search. *SIGMOD*. 349–360.
- [23] Lu, S., Luo, Y., Zhou, S., Cong, G., and Guan, J. 2014. Distributed spatial keyword querying on road networks. *EDBT*. 235–246.

- [24] Qu, Q., Liu, S., Yang, B., and Jensen, C. S. 2014. Efficient top-k spatial locality search for co-located spatial web objects. *MDM*. 269-.278.
- [25] Qu, Q., Liu, S., Yang, B, and Jensen, C. S. 2014. Integrating non-spatial preferences into spatial location queries. *SSDBM*. Article 8.
- [26] Rocha-Junior, J. B., Gkorgkas, O., Jonassen, S., and, Nørvåg, K. 2011. Efficient processing of top-k spatial keyword queries. *SSTD*. 205–222.
- [27] Skovsgaard, A. and Jensen, C. S. 2014. Top-k point of interest retrieval using standard indexes. SIGSPATIAL/GIS. 173–182.
- [28] Skovsgaard, A., Šidlauskas, D., and Jensen, C. S. 2014. A clustering approach to the discovery of points of interest from geo-tagged microblog posts. *MDM*. 178–188.
- [29] Statistic Brain. 2015. Google Annual Search Statistics. www. statisticbrain.com/google-searches.
- [30] Search Engine Land. 2010. Microsoft: 53 percent of mobile searches have local intent. searchengineland.com/ microsoft-53-percent-of-mobile-searches-have-local-intent-55556.
- [31] Venetis, P., Gonzalez, H., Jensen, C. S., and Halevy, A. 2011. Hyper-local, directions-based ranking of places. *PVLDB* 4, 5, 290–301.
- [32] Wang, X., Zhang, Y., Zhang, W., Lin, X., and Wang, W. 2015. AP-tree: efficiently support continuous spatial-keyword queries over stream. *ICDE*.
- [33] Wu, D., Choi, B., Xu, J., and Jensen, C. S. 2015. Authentication of moving top-k spatial keyword queries. *TKDE 27*, 4, 922–935.
- [34] Wu, D., Cong, G., and Jensen, C. S. 2012. A framework for efficient spatial web object retrieval. VLDBJ 21, 6, 797–822.
- [35] Wu, D., Yiu, M. L., Cong, G., and Jensen, C. S. 2012. Joint top-k spatial keyword query processing. *TKDE 24*, 10, 1889– 1903.
- [36] Wu, D., Yiu, M. L., and Jensen, C. S. 2013. Moving spatial keyword queries: Formulation, methods, and analysis. *TODS 38*, 1, Article 7.
- [37] Wu, D., Yiu, M. L., Jensen, C. S., and Cong, G. 2011. Efficient continuously moving top-k spatial keyword query processing. *ICDE*. 541–552.
- [38] Zhang, D., Chee, Y. M., Mondal, A., Tung, A. K. H., and Kitsuregawa, M. 2009. Keyword search in spatial databases: Towards searching by document. *ICDE*. 688–699.
- [39] Zhen, K., Su, H., Zheng, B., Shang, S., Xu, J., Liu, J., and Zhou, X. 2015. Interactive top-k spatial keyword queries. *ICDE*.
- [40] Zhou, Y., Xie, X., Wang, C., Gong, Y., and Ma, W.-Y. 2005. Hybrid index structures for location-based web search. *CIKM*. 155–162.