Search on Graphs: Theory Meets Engineering

Yuqing Wu[1] and George H.L. Fletcher[2]

[1] Indiana University, Bloomington, USA
yuqwu@cs.indiana.edu
[2] Eindhoven University of Technology, The Netherlands
g.h.l.fletcher@tue.nl

Abstract. The last decade has witnessed an explosion of the availability of and interest in graph structured data. The desire to search and reason over these increasingly massive data collections pushes the boundaries of search languages, from pure keyword search to structure-aware searches in the graph. These phenomena have inspired a rich body of research on query languages, data management and query evaluation techniques for graph data, both from the theoretical and engineering angles. In this tutorial, we present an overview of the progress on graph search queries, focusing specifically on how the theoretical and engineering perspectives meet and together advanced the field.

1 Tutorial Overview

Exploratory keyword-style search has been heavily studied in the past decade, both in the context of structured [18] and semi-structured [9] data. Given the ubiquity of massive (loosely structured) graph data in domains such as the web, social networks, biological networks, and linked open data (to name a few), there recently has been a surge of interest and advances on the problem of search in graphs (e.g., [1, 3, 12, 15, 17, 19, 20]). As graph exploration leads to deeper domain understanding, user queries begin to shift from unstructured searching to richer structure-based exploration of the graph. Consequently, there has been a flurry of language proposals specifically targeting this style of structure-aware querying in graphs (e.g., [2, 3, 5, 13, 16]).

In this tutorial, we survey this growing body of work, with an eye towards both bringing participants up to speed in this field of rapid progress and delimiting the boundaries of the state-of-the-art. A particular focus will be on recent results in the theory of graph languages on the design and structural characterization of simple yet powerful algebraic languages for graph search, which bridge structure-oblivious and structure-aware graph exploration [4–6]. At the heart of these results is the methodology of coupling the expressive power of a given query language with an appropriate structural notion on data instances. Here, the idea is to characterize language equivalence of data objects in instances (i.e., the inability of queries in the language to distinguish the objects) purely in terms of the structure of the instance (i.e., equivalence under notions such as homomorphism or bisimilarity). Recently, first steps towards graph

Y. Ishikawa et al. (Eds.): APWeb 2013, LNCS 7808, pp. 3–6, 2013.

indexing have shown promise in transferring this theoretical framework into practice [7, 14]. The basic intuition behind this approach is that data should be organized to optimally reflect the type and style of queries being asked, and that the optimality of this organization can be formally established, as above. Recent results have established the practical feasibility of computing and maintaining these structural organizations on massive graphs [8, 10, 11].

2 Tutorial Outline

The tutorial will be presented as follows:

Part 1: Searching the Graph
We start the tutorial with an overview of variants of graph data and the evolution of the search queries on graph, leading into the discussion of the challenges posed by the massive size and complex nature of graph data and the flexible nature of graph queries and their evaluation.

Part 2: Bridging Theory and Engineering
We next examine the lines of work in the theoretical study of query languages and the engineering efforts in developing novel techniques for managing graph data and evaluating various types of search queries on such data. We then present our methodology of coupling the expressive power of a given query language with an appropriate structural notion on data instances, as a tool for reasoning about and guiding engineering efforts.

Part 3: Indexing Graph Data – A Case Study
We follow this discussion with a presentation of the design of indices for semi-structured and graph data, as a case study, to illustrate our methodology.

Part 4: Looking Forward
We close the tutorial with indications of ongoing and future research directions.

3 Speakers

Yuqing Wu and George Fletcher, together with a group of collaborators in the USA, the Netherlands, and Belgium, have been conducting research in this area in recent years and have published several papers in both the theory and engineering branches of database research.

Yuqing Wu, Indiana University, Bloomington, USA
Prof. Wu is an Associate Professor at School of Informatics and Computing, Indiana University, Bloomington, USA. Prof. Wu received her Ph.D. degree from University of Michigan, Ann Arbor, in 2004. Her research area is in data management, especially semi-structure and non-structured data, with an emphasis on query language, query processing and query optimization.

George Fletcher, Eindhoven University of Technology, The Netherlands
Dr. Fletcher is an Assistant Professor in the Databases and Hypermedia group at Eindhoven University of Technology, The Netherlands. Dr. Fletcher was awarded a doctorate in computer science from Indiana University, Bloomington (2007), with a dissertation on the topic of query learning for data integration. His current research focuses on the study of database query languages for data integration and web data.

References

1. Delbru, R., Campinas, S., Tummarello, G.: Searching web data: An entity retrieval and high-performance indexing model. J. Web Sem. 10, 33–58 (2012)
2. Fazzinga, B., Gianforme, G., Gottlob, G., Lukasiewicz, T.: Semantic web search based on ontological conjunctive queries. J. Web Sem. 9(4), 453–473 (2011)
3. Fletcher, G.H.L., Van den Bussche, J., Van Gucht, D., Vansummeren, S.: Towards a theory of search queries. ACM Trans. Database Syst. 35(4), 28 (2010)
4. Fletcher, G.H.L., Gyssens, M., Leinders, D., Van den Bussche, J., Van Gucht, D., Vansummeren, S.: Similarity and bisimilarity notions appropriate for characterizing indistinguishability in fragments of the calculus of relations. CoRR, abs/1210.2688 (2012)
5. Fletcher, G.H.L., Gyssens, M., Leinders, D., Van den Bussche, J., Van Gucht, D., Vansummeren, S., Wu, Y.: Relative expressive power of navigational querying on graphs. In: Proc. ICDT, Uppsala, Sweden, pp. 197–207 (2011)
6. Fletcher, G.H.L., Gyssens, M., Leinders, D., Van den Bussche, J., Van Gucht, D., Vansummeren, S., Wu, Y.: The impact of transitive closure on the boolean expressiveness of navigational query languages on graphs. In: Lukasiewicz, T., Sali, A. (eds.) FoIKS 2012. LNCS, vol. 7153, pp. 124–143. Springer, Heidelberg (2012)
7. Fletcher, G.H.L., Hidders, J., Vansummeren, S., Picalausa, F., Luo, Y., De Bra, P.: On guarded simulations and acyclic first-order languages. In: Proc. DBPL, Seattle, WA, USA (2011)
8. Hellings, J., Fletcher, G.H.L., Haverkort, H.: Efficient external-memory bisimulation on DAGs. In: Proc. ACM SIGMOD, Scottsdale, AZ, USA, pp. 553–564 (2012)
9. Liu, Z., Chen, Y.: Processing keyword search on XML: a survey. World Wide Web 14(5-6), 671–707 (2011)
10. Luo, Y., de Lange, Y., Fletcher, G.H.L., De Bra, P., Hidders, J., Wu, Y.: Bisimulation reduction of big graphs on MapReduce (manuscript in preparation, 2013)
11. Luo, Y., Fletcher, G.H.L., Hidders, J., Wu, Y., De Bra, P.: I/O-efficient algorithms for localized bisimulation partition construction and maintenance on massive graphs. CoRR, abs/1210.0748 (2012)
12. Mass, Y., Sagiv, Y.: Language models for keyword search over data graphs. In: Proc. ACM WSDM, Seattle, Washington, USA (2012)
13. Pérez, J., Arenas, M., Gutierrez, C.: nSPARQL: A navigational language for RDF. J. Web Sem. 8(4), 255–270 (2010)
14. Picalausa, F., Luo, Y., Fletcher, G.H.L., Hidders, J., Vansummeren, S.: A structural approach to indexing triples. In: Simperl, E., Cimiano, P., Polleres, A., Corcho, O., Presutti, V. (eds.) ESWC 2012. LNCS, vol. 7295, pp. 406–421. Springer, Heidelberg (2012)
15. Tran, T., Herzig, D.M., Ladwig, G.: SemSearchPro - using semantics throughout the search process. J. Web Sem. 9(4), 349–364 (2011)

16. Tran, T., Wang, H., Rudolph, S., Cimiano, P.: Top-k exploration of query candidates for efficient keyword search on graph-shaped (RDF) data. In: Proc. IEEE ICDE, Shanghai, pp. 405–416 (2009)
17. Wu, Y., Van Gucht, D., Gyssens, M., Paredaens, J.: A study of a positive fragment of path queries: Expressiveness, normal form and minimization. Comput. J. 54(7), 1091–1118 (2011)
18. Yu, J.X., Qin, L., Chang, L.: Keyword search in relational databases: A survey. IEEE Data Eng. Bull. 33(1), 67–78 (2010)
19. Zhou, M., Pan, Y., Wu, Y.: Conkar: constraint keyword-based association discovery. In: Proc. ACM CIKM, Glasgow, UK, pp. 2553–2556 (2011)
20. Zhou, M., Pan, Y., Wu, Y.: Efficient association discovery with keyword-based constraints on large graph data. In: Proc. ACM CIKM, Glasgow, UK, pp. 2441–2444 (2011)