Ranking on Large-Scale Graphs with Rich Metadata

Bin Gao
Microsoft Research Asia
4F, Sigma Center, No. 49,
Zhichun Road
Beijing, 100190, P. R. China
bingao@microsoft.com

Taifeng Wang
Microsoft Research Asia
4F, Sigma Center, No. 49,
Zhichun Road
Beijing, 100190, P. R. China
taifengw@microsoft.com

Tie-Yan Liu
Microsoft Research Asia
4F, Sigma Center, No. 49,
Zhichun Road
Beijing, 100190, P. R. China
tyliu@microsoft.com

ABSTRACT

For many Web applications, one needs to deal with the ranking problem on large-scale graphs with rich metadata. However, it is non-trivial to perform efficient and effective ranking on them. On one aspect, we need to design scalable algorithms. On another aspect, we also need to develop powerful computational infrastructure to support these algorithms. This tutorial aims at giving a timely introduction to the promising advances in the aforementioned aspects in recent years, and providing the audiences with a comprehensive view on the related literature.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval; H.5.4 [**Information Interface and Presentation**]: Hypertext/Hypermedia.

General Terms

Algorithm, Experimentation, Theory

Keywords

Large-scale graph, graph ranking, Markov process, Map-Reduce.

1. INTRODUCTION

In many Web applications, we need to tackle the problem of ranking on large-scale graphs with rich metadata. For example, to compute page importance ranking for search, one may need to analyze the link structure of the Web graph; to understand Web user behaviors and preferences, one may need to rank the nodes on user-page bipartite graphs extracted from search engine logs; and to make recommendations in online community, one may need to conduct ranking on the social network graph. All these graphs are of very large scale and contain rich information (represented by rich metadata). As a result, it is non-trivial to perform efficient and effective ranking on them. On one aspect, we need to design scalable algorithms. On another aspect, we also need to develop powerful computational infrastructure to support these algorithms. We observe that in recent years, there are some promising advances in the aforementioned aspects, which can potentially enhance many important Web search and data mining applications, and greatly advance the state of the art of the related research. This tutorial aims at giving a timely introduction to these works, and providing the audiences with a comprehensive view on the related literature. We believe many researchers in the Web search and data mining community would have interest in listening to this tutorial, and we hope that they can be motivated to participate in the research of large-scale graph ranking with rich metadata.

WWW 2011, March 28–April 1, 2011, Hyderabad, India.
ACM 978-1-4503-0637-9/11/03.

2. TUTORIAL DESCRIPTION

We will organize the tutorial into six parts.

In the first part, a brief introduction of large-scale graphs with rich metadata and their properties will be given. The widely-used mathematical tools and models to represent these graphs and rich metadata will be introduced.

In the second part, we will focus on unsupervised ranking on large-scale graphs. We propose a unified view on the unsupervised graph ranking algorithms, i.e., they are generative models based on some stochastic processes. Specifically, a general ranking framework based on Markov skeleton process [8, 9] will be introduced. The algorithms in the literature are organized in two categories.

- **Graph ranking algorithms on single graph.** To compute page importance ranking in Web graphs, HITS [10] and PageRank [13] were proposed, which are based on the analysis of link structures of Web graphs. These methods can be explained using random walks on a discrete-time Markov process. After that, quite a few link analysis algorithms like TrustRank [7] and PopRank [12] were developed to improve HITS and PageRank to robustly deal with web spam, and to handle heterogeneous graphs. BrowseRank [11] was proposed recently to consider rich metadata (e.g., visiting frequency and staying time) in user behavior data for page importance ranking, which is based on a new mathematical tool - continuous time Markov process. MobileRank [6] and BrowseRank Plus [6] further improved BrowseRank in considering more dependency between different nodes in the graph and more metadata. It has been shown that most of these algorithms can be summarized into a general framework based on Markov skeleton process.
- **Graph ranking algorithms on graph series.** Mining on a time series of graph snapshots also attract much interests in recent years. Basically with the graph series, one can find some trends or to obtain

a more stable ranking results. We will introduce such kind of work like TemporalRank [16].

In the third part, we will introduce supervised large-scale graph ranking algorithms. We propose a unified view on the supervised graph ranking algorithms, i.e., they are discriminative models on some graph-based smoothing function. Specifically, a supervised/semi-supervised graph ranking framework will be introduced.

- **Supervised/Semi-supervised graph ranking algorithms.** Recently people have realized that unsupervised graph ranking sometimes might not be consistent with human intuition. To solve the problem, supervised or semi-supervised graph ranking schemes have been considered. Take importance ranking on graphs as an example, there are several pieces of work along this line. LiftHITS [5] learns from user feedback to realign the eigenvectors of the link matrix in HITS. Adaptive PageRank [15] alters the PageRank scores according to human feedback, using a quadratic programming technique. NetRank [2, 4, 1] provides a uniform framework for learning the parameters of Markov random walks on graphs according to supervision in terms of pairwise preference between nodes. In Laplacian Rank [3, 17, 14], the supervised graph ranking problem was formulated as minimizing the combination of the empirical error and a regularization term represented by graph. Semi-supervised PageRank considers both graph structure and the rich metadata contained in node features and edge features in the ranking process.

In the fourth part, we make complexity analysis on the above mentioned algorithms, to discuss their scalability and efficiency. Based on the discussion, we give some guidelines on designing highly scalable and efficient algorithms for large-scale graph ranking algorithms.

In the fifth part, we discuss how to design a distributed system to support large-scale graph ranking algorithms, and how to parallel existing algorithms to better fit into the system. The design principles for such a distributed system will first be discussed, and then a real system will be explained in detail, which has been used in a commercial search engine. This system contains a map-reduce engines specifically optimized for graph operations, as well as a rich pool of graph operators such as graph traverse, graph partitioning, and graph sampling. It is easy to implement a graph ranking algorithm on this platform. We will take Semi-supervised PageRank as an example to make detailed discussion on this process.

In the last part of the tutorial, the future research directions regarding large-scale graph ranking with rich metadata, and open questions and challenges will be discussed.

3. REFERENCES

[1] A. Agarwal and S. Chakrabarti. Learning random walks to rank nodes in graphs. In the proceedings of the 24th International Conference on Machine Learning (*ICML*), pages 9-16, 2007.

[2] A. Agarwal, S. Chakrabarti, and S. Aggarwal. Learning to rank networked entites. In the proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pages 14-23, 2006.

[3] S. Agarwal. Ranking on graph data. In the proceedings of the 23th International Conference on Machine Learning (*ICML*), pages 25-32, 2006.

[4] S. Chakrabarti and A. Agarwal. Learning parameters in entity relationship graphs from ranking preferences. In the proceedings of the 10th European Conference on Principles and Practice of Knowledge Discovery in Databases (*PKDD*), volume 4213, pages 91-102, 2006.

[5] H. Chang, D. Cohn, and A. K. McCallum. Learning to create customized authority lists. In the proceedings of the 17th International Conference on Machine Learning (*ICML*), pages 127-134, 2000.

[6] B. Gao, T. Liu, Z. Ma, T. Wang, and H. Li. A general markov framework for page importance computation. In CIKM'09, pp, 1835-1838, 2009.

[7] Z. Gyongyi, H. Garcia-Molina, and J. Pedersen. Combating web spam with trustrank. In *VLDB '04*, pages 576–587. VLDB Endowment, 2004.

[8] Z. Hou and G. Liu. Markov Skeleton Processes and their Applications. Science Press and International Press, 2005.

[9] Z. Hou, Z. Liu, and J. Zou. Markov Skeleton Processes. In *Chinese Science Bulletin*, vol 43, no 11, pages 881-889, June, 1998.

[10] J. M. Kleinberg. Authoritative sources in a hyperlinked environment. In the proceedings of the 9th Annual ACM-SIAM Symposium on Discrete Algorithms (*SODA*), pages 668-677, 1998.

[11] Y. Liu, B. Gao, T. Liu, Y. Zhang, Z. Ma, S. He, and H. Li. BrowseRank: letting users vote for page importance. In *SIGIR '08*, pages 451–458, 2008.

[12] Z. Nie, Y. Zhang, J. Wen, and W. Ma. Object-Level Ranking: Bringing Order to Web Objects. In *WWW'05*, 2005.

[13] L. Page, S. Brin, R. Motwani, and T. Winograd. The PageRank citation ranking: bringing order to the web. Technical report, Stanford Digital Library Technologies Project, 1998.

[14] D. Rao, and D. Yarowsky. Ranking and semi-supervised classification on large scale graphs using map-reduce. In the proceedings of the 4th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing (*ACL-IJCNLP*), 2009.

[15] A. C. Tsoi, G. Morini, F. Scarselli, M. Hagenbuchner, and M. Maggini. Adaptive ranking of Web pages. In the proceedings of the 12th International World Wide Web Conference (*WWW*), pages 356-365, 2003.

[16] L. Yang, L. Qi, Y. Zhao, B. Gao, and T. Liu Link Analysis using Time Series of Web Graphs. In CIKM'07, pp. 1011-1014, 2007.

[17] D. Zhou, J. Huang, and B. Scholkopf. Learning from labeled and unlabeled data on a directed graph. In the proceedings of the 22th International Conference on Machine Learning (*ICML*), pages 1041-1048, 2005.