Preliminary Reading List for BMGT 828
Computational Challenges of Web 2.0 and Beyond

OK - Don't panic that the preliminary reading list is so long! Avi and I wanted to include a smorgasboard of papers in a subset of areas. We don't expect you to read or be familiar with all these papers. It will be okay if you take a look at the abstracts. We will identify a few papers for you to read for each session.

- Machine Learning
  - Video [20].
  - The structure and function of complex networks [25](Sections 1, 2, and 3)
  - Preserving the Privacy of Sensitive Relationships in Graph Data [38]
-
- Ranking Web pages
  - Classic readings: [6, 26, 21].
  - A mathematical treatment: [14, 23].
  - Various extensions: [1, 16, 2, 11, 12, 9].
  - Efficient and / or distributed computation: [37, 30, 18].
  - A practical approach: [19].
- Schema and Ontology Matching
  - Introductory reading: [15].
  - Efficient computation: [10].
  - Semantics: [13].
  - Schema matching and the Web: [31, 7].
- Online Monitoring
  - Publish-Subscribe: [24, 36, 35, 34, 32]
  - Coherency: [22, 8].
  - Web information sources: [27, 3, 17].
  - Continuous queries: [29].
  - Complex event processing: [4]
  - User Data needs: [33, 5, 28].

References

[1] `http://www.useit.com/alertbox/web-growth.html`.

[2] A. Balmin, V. Hristidis, and Y. Papakonstantinou. ObjectRank: Authority-based keyword search in databases. In *VLDB, 2004*, 2004.

[3] Nilesh Bansal and Nick Koudas. Searching the blogosphere. In *WebDB*, 2007.

[4] Lars Brenna, Alan J. Demers, Johannes Gehrke, Mingsheng Hong, Joel Ossher, Biswanath Panda, Mirek Riedewald, Mohit Thatte, and Walker M. White. Cayuga: a high-performance event processing engine. In *SIGMOD Conference*, pages 1100–1102, 2007.

[5] Laura Bright, Avigdor Gal, and Louiqa Raschid. Adaptive pull-based policies for wide area data delivery. *ACM Transactions on Database Systems (TODS)*, 31(2):631–671, 2006.

[6] Sergey Brin and Lawrence Page. The anatomy of a large-scale hypertextual web search engine. In *WWW7: Proceedings of the seventh international conference on World Wide Web 7*, pages 107–117, Amsterdam, The Netherlands, The Netherlands, 1998. Elsevier Science Publishers B. V.

[7] Adriana Budura, Philippe Cudré-Mauroux, and Karl Aberer. From bioinformatic web portals to semantically integrated data grid networks. *Future Gener. Comput. Syst.*, 23(3):485–496, 2007.

[8] Junghoo Cho and Uri Schonfeld. Rankmass crawler: A crawler with high pagerank coverage guarantee. In *VLDB*, pages 375–386, 2007.

[9] Jason V. Davis and Inderjit S. Dhillon. Estimating the global pagerank of web communities. In *KDD '06: Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 116–125, New York, NY, USA, 2006. ACM Press.

[10] C. Domshlak, A. Gal, and H. Roitman. Rank aggregation for automatic schema matching. *IEEE Transactions on Knowledge and Data Engineering (TKDE)*, 19(4):538–553, 2007.

[11] Cynthia Dwork, Ravi Kumar, Moni Naor, and D. Sivakumar. Rank aggregation methods for the web. In *WWW '01: Proceedings of the 10th international conference on World Wide Web*, pages 613–622, New York, NY, USA, 2001. ACM Press.

[12] Nadav Eiron, Kevin S. McCurley, and John A. Tomlin. Ranking the web frontier. In *WWW '04: Proceedings of the 13th international conference on World Wide Web*, pages 309–318, New York, NY, USA, 2004. ACM Press.

[13] Jérôme Euzenat and Pavel Shvaiko. *Ontology matching*. Springer-Verlag, Heidelberg (DE), 2007.

[14] Michalis Faloutsos, Petros Faloutsos, and Christos Faloutsos. On power-law relationships of the internet topology. In *SIGCOMM '99: Proceedings of the conference on Applications, technologies, architectures, and protocols for computer communication*, pages 251–262, New York, NY, USA, 1999. ACM Press.

[15] A. Gal. Why is schema matching tough and what can we do about it? *SIGMOD Record*, 35(4):2–5, 2007.

[16] Taher H. Haveliwala. Topic-sensitive pagerank. In *WWW '02: Proceedings of the 11th international conference on World Wide Web*, pages 517–526, New York, NY, USA, 2002. ACM Press.

[17] Seung Jun and Mustaque Ahamad. Feedex: collaborative exchange of news feeds. In *WWW*, pages 113–122, 2006.

[18] Sepandar D. Kamvar, Taher H. Haveliwala, Christopher D. Manning, and Gene H. Golub. Extrapolation methods for accelerating pagerank computations. In *WWW*, pages 261–270, 2003.

[19] Peter Kent. Search engine optimization for dummies. 2006.

[20] Jon Kleinberg. Keynote at acm sigkdd 2007.

[21] Jon M. Kleinberg. Authoritative sources in a hyperlinked environment. *Journal of the ACM*, 46(5):604–632, 1999.

[22] Ratul kr. Majumdar, Kannan M. Moudgalya, and Krithi Ramamritham. Adaptive coherency maintenance techniques for time-varying data. In *RTSS '03: Proceedings of the 24th IEEE International Real-Time Systems Symposium*, page 98, Washington, DC, USA, 2003. IEEE Computer Society.

[23] Amy N. Langville and Carl D. Meyer. Updating markov chains with an eye on google's pagerank. *SIAM J. Matrix Anal. Appl.*, 27(4):968–987, 2006.

[24] Hongzhou Liu, Venugopalan Ramasubramanian, and Emin Gün Sirer. Client behavior and feed characteristics of RSS, a publish-subscribe system for web micronews. In *IMC'05: Proceedings of the Internet Measurement Conference 2005 on Internet Measurement Conference*, pages 3–3, Berkeley, CA, USA, 2005. USENIX Association.

[25] M. E. J. Newman. The structure and function of complex networks, March 2003.

[26] Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. The pagerank citation ranking: Bringing order to the web. Technical report, Stanford Digital Library Technologies Project, 1998.

[27] Sandeep Pandey, Kedar Dhamdhere, and Christopher Olston. WIC: A general-purpose algorithm for monitoring web information sources. In *Proceedings of the 30th International Conference on Very Large Data Bases (VLDB)*, pages 360–371, 2004.

[28] Sandeep Pandey and Christopher Olston. User-centric web crawling. In *Proceedings of the 14th international conference on World Wide Web (WWW)*, pages 401–411, New York, NY, USA, 2005. ACM.

[29] Sandeep Pandey, Krithi Ramamritham, and Soumen Chakrabarti. Monitoring the dynamic web to respond to continuous queries. In *Proceedings of the 12th international conference on World Wide Web (WWW)*, pages 659–668, 2003.

[30] Josiane Xavier Parreira, Debora Donato, Sebastian Michel, and Gerhard Weikum. Efficient and decentralized pagerank approximation in a peer-to-peer web search network. In *VLDB '06: Proceedings of the 32nd international conference on Very large data bases*, pages 415–426. VLDB Endowment, 2006.

[31] A. Doan R. McCann, W. Shen. Matching schemas in online communities: A web 2.0 approachm.

[32] Venugopalan Ramasubramanian, Ryan Peterson, and Emin Gün Sirer. Corona: A high performance publish-subscribe system for the world wide web. In *NSDI*, 2006.

[33] Haggai Roitman, Avigdor Gal, and Louiqa Raschid. Satisfying complex data needs using pull-based online monitoring of volatile data sources. In *International Conference on Data Engineering (ICDE), Cancun, Mexico, to appear*, 2008. Available upon request from haggair@gmail.com.

[34] Ian Rose, Rohan Murty, Peter R. Pietzuch, Jonathan Ledlie, Mema Roussopoulos, and Matt Welsh. Cobra: Content-based filtering and aggregation of blogs and rss feeds. In *NSDI*, 2007.

[35] Ka Cheung Sia, Junghoo Cho, and Hyun-Kyu Cho. Efficient monitoring algorithm for fast news alerts. *IEEE Transactions on Knowledge and Data Engineering*, 19(7):950–961, 2007.

[36] Ka Cheung Sia, Junghoo Cho, Koji Hino, Yun Chi, Shenghuo Zhu, and Belle L. Tseng. Monitoring rss feeds based on user browsing pattern search. In *Proceedings of the International Conference on Weblogs and Social Media*, 2007.

[37] Yuan Wang and David J. DeWitt. Computing pagerank in a distributed internet search engine system. In *VLDB*, pages 420–431, 2004.

[38] Elena Zheleva and Lise Getoor. Preserving the privacy of sensitive relationships in graph data. In *First ACM SIGKDD Workshop on Privacy, Security, and Trust in KDD (PinKDD'07)*, 2007.