TUTORIAL FOR APWEB/WAIM 2009

Tutorial title: **Mining Evolution of Complex Structured Data**

Author: Sourav S Bhowmick
Affiliation: Associate Professor
School of Computer Engineering
Nanyang Technological University
Singapore
Email: assourav@ntu.edu.sg
Web: http://www.ntu.edu.sg/home/assourav/

Audience: Researchers and practitioners interested in tree and graph structured data mining.

Tutorial history: To the best of my knowledge, this tutorial has not been presented in any major data mining conference.

Abstract:

Many real-life data can be represented as tree or graph. Different data mining techniques have been recently proposed to mine such complex structured data. These techniques can be broadly classified into two categories: (i) A large number of algorithms have been designed to find different types of pattern from such data by considering them as snapshot data. These techniques are *evolution-unaware*. (ii) Recently, there has been increasing research efforts in wide variety of domains such as XML, Web, and life sciences that mine evolutionary features of tree and graph structured data to discover novel knowledge. Typically, such knowledge cannot be discovered by mining snapshot data. This tutorial focuses on this second issue. That is, we highlight recent efforts in *discovering novel knowledge from the historical evolution patterns of tree and graph structured data*.

The tutorial is structured as follows. We motivate the necessity for mining evolution and give an overview of the evolutionary features of various types of tree and graph-structured data. Next, we identify various research issues involved in evolution mining. Specifically, our discussion can be categorized into the following three main components:

- Study of evolution mining for tree-structured data. We use XML and web usage data as representatives of tree structured data.
- Study of evolution mining for graph-structured data. We use web communities, click-through data, and biological networks as representatives of graph-structured data.

- This tutorial session also reveal various application domains of evolution mining (such as social networks, blogs, web event detection, web personalization, XML query caching, protein function prediction, protein-protein interaction prediction, etc).

We conclude by identifying potential research directions in this area.

Duration: Half day.

Outline:

Section 1: Introduction and Motivation: This section includes a brief overview on the dynamic nature of the complex structured data (XML, Web, and biological networks) and how it is affecting our information need. Then we motivate the need for mining evolution of such data. Lastly, we identify the technical challenges associated with this problem.

Section 2: Mining Evolution of Tree-Structured Data: In this section we report and analyze various studies related to mining evolutionary features of tree-structured data. In particular, we discuss the followings:

- We present efforts related to mining structure of historical XML documents and XML queries [3, 4, 27, 32, 33]
- As web usage data is evolutionary in nature and can be represented as web session trees, we discuss efforts related to mining evolutionary features of web user sessions [2, 5, 19, 20, 26, 29, 30, 31].

Section 3: Mining Evolution of Graph-Structured Data: Next, we move from tree to graph-structured data. We explore works that propose novel data mining techniques to mine the evolution to graph data in order to discover interesting evolutionary patterns. Specifically, we present state-of-the-art techniques related to the followings:

- Mining evolution of web communities (social networks, blogs) [1, 6, 9, 10, 11, 14, 15, 16, 17, 18, 24, 25, 34, 35]
- Mining evolution of search engine log data [20, 28]
- Mining evolution of biological networks [7, 8, 12, 13, 21, 22, 23]

Section 4: Applications of Evolution Mining: Here we highlight how evolution techniques are used in several application domains such as XML query caching, web personalization, web event detection, Web 2.0, and life sciences.

Section 5: The Road Ahead: We expose potential research issues in mining evolution of complex-structured data. We also throw some open questions for the audience to ponder about.

Biographies:

Sourav S Bhowmick is an Associate Professor in the School of Engineering, Nanyang Technological University and the Director of Centre for Advanced Information Systems (CAIS) . He is currently Visiting Associate Professor at the Biological Engineering Division, Massachusetts Institute of Technology (MIT), USA. He also holds the position of Singapore-MIT Alliance (SMA) Fellow in Computation and Systems Biology program (2005-2008). Sourav received his Ph.D. in computer engineering in 2001. His current research interests include XML data management, systems biology data management, web data management, and data mining. He has published more than 100 papers in major international database and data mining conferences and journals such as VLDB, IEEE ICDE, ACM WWW, ACM SIGMOD, ACM SIGKDD, ACM CIKM, ER, PAKDD, IEEE TKDE, ACM CS, Information Systems, and DKE. Sourav is serving as a PC member of various database conferences and workshops and reviewer for various database journals. He is also serving as a program chair/co-chair of several international workshops in biological and XML data management. He is a member of the editorial boards of several international journals. He has given tutorial in ER 2006, APWeb 2008, and WAIM 2008. He has co-authored a book entitled "Web Data Management: A Warehouse Approach" (Springers Verlag, October 2003). Sourav is a member of ACM and an affiliate member of IEEE.

Sourav has received **Best Interdisciplinary Paper Award** (along with Q Zhao, M Mohania, Y Kambayashi) at ACM CIKM 2004 for the paper titled "Discovering Frequently Changing Structures from Historical Structural Deltas of Unordered XML" . He was also nominated for Excellence in Teaching Award for three consecutive years (2003 – 2005).

Representive References Used in the Tutorial:

1. N. Agarwal, H. Liu, L. Tang, P. S. Yu. **Identifying the Influential Bloggers in a Community.** In *WSDM,* 2008.

2. S Baron, M Spiliopoulou. **Monitoring the Evolution of Web Usage Patterns.** *EWMF 2003.*

3. Ling Chen, Sourav S Bhowmick, and L-T Chia. **FRACTURE Mining: Mining Frequently and Concurrently Mutating Structures from**

Historical XML Documents. In *Data and Knowledge Engineering Journal* (DKE), 59(2), Elsevier Science, 2006.

4. Ling Chen, Sourav S Bhowmick, and L-T Chia. **Mining Association Rules from Structural Deltas of Historical XML Documents.** In *PAKDD*, Sydney, 2004.

5. Ling Chen, Sourav S Bhowmick, W Nejdl. **COWES: Web User Clustering Based on Evolutionary Web Sessions.** To appear in DKE, 2009.

6. Bi Chen, Qiankun Zhao, Bingjun Sun, Prasenjit Mitra. **Predicting Blogging Behavior Using Temporal and Social Networks.** In ICDM 2007.

7. J. Dutkowski J. Tiuryn. **Identification of functional modules from conserved ancestral protein-protein interactions.** *Bioinformatics*, 23, *i*149–*i*158, 2007.

8. J. Flannick, A. Novak, B.S. Srinivasan et al. **Græmlin: General and robust alignment of multiple large interaction networks.** *Genome Research*, 2006.

9. T. Falkowski, J. Bartelheimer, and M. Spiliopoulou. **Mining and visualizing the evolution of subgroups in social networks**. In *IEEE/WIC/ACM WI*, 2006.

10. D. Gruhl, R. Guha, R. Kumar, J. Novak, A Tomkins. **On the Predictive Power of Online Chatters.** In *ACM KDD*, 2005.

11. A. Java, P. Kolari, T. Finin, and T. Oates. **Modeling the Spread of Influence on the Blogsphere.** In *ACM WWW*, 2006

12. M. Kalaev, V. Bafna, R. Sharan. **Fast and Accurate Alignment of Multiple Protein Networks.** *In Proc. of ACM RECOMB*, 2008.

13. B. P. Kelley et al. **Pathblast: a tool for alignment of protein interaction networks.** *Nucleic Acids Res.*, 32, W83-W88.

14. J Leskovec, K J Lang et al. **Statistical Properties of Community Structure in Large Social and Information Networks.** *In WWW, 2008.*

15. Jure Leskovec, Mary McGlohon, Christos Faloutsos, Natalie S. Glance, Matthew Hurst. **Patterns of Cascading Behavior in Large Blog Graphs.** SDM 2007

16. Jure Leskovec, Jon M. Kleinberg, Christos Faloutsos. **Graph evolution: Densification and shrinking diameters.** In TKDD, 2007.

17. Y-R Lin, H Sundaram, Y Chi et al. **Blog Community Discovery and Evolution Based on Mutual Awareness Expansion.** In *IEEE/ACM WI,* 2007.

18. Y-R Lin, Y Chi et al. **FacetNet: A Framework for Analyzing Communities and Their Evolutions in Dynamic Networks.** In WWW, 2008.

19. Sophia G. Petridou, Vassiliki A. Koutsonikola, Athena Vakali, Georgios I. Papadimitriou: T**ime-Aware Web Users' Clustering.** IEEE Trans. Knowl. Data Eng. 20(5): 653-667, 2008.

20. B. Piwowarski, Hugo Zaragoza. **Predictive user click models based on click-through history.** In CIKM, 2007.

21. B,-S Seah, Sourav S Bhowmick, C F Dewey, Jr, H Yu. **DiscriAlign: A Discriminative Two-Stage Approach for Alignment of Protein Interaction Networks.** Under review.

22. Roded Sharan, Trey Ideker. **Modeling Cellular Machinery Through Biological Network Comparison**. Nature Biotechnology, 24(4), April 2006.

23. R. Sharan, S. Suthram, R. M. Kelley et al. **Conserved patterns of protein interaction in multiple species.** *PNAS*, 102, pp. 1974-1979, 2005.

24. P Singla, M Richardson. **Yes, There is a Correlation - From Social Networks to Personal Behavior on the Web.** In WWW 2008.

25. C-Y Teng, H-H Chen. **Detection of Bloggers' Interests: Using Textual, Temporal, and Interactive Features.** In *ACM WI*, 2006.

26. Vincent S. Tseng, Kawuu Weicheng Lin, Jeng-Chuan Chang. **Prediction of user navigation patterns by mining the temporal web usage evolution.** Soft Comput. (SOCO) 12(2): 157-163, 2008)

27. Qiankun Zhao, Ling Chen, Sourav S Bhowmick, Sanjay Madria. **XML Structural Delta Mining: Issues and Challenges**. In *Data and Knowledge Engineering Journal* (DKE), 59(3), Elsevier Science, 2006.

28. Qiankun Zhao, T-Y Liu, Sourav S Bhowmick, and W-Y Ma. **Event Detection from the Evolution of Click-through Data**. In *ACM SIGKDD*, Philadelphia, 2006.

29. Qiankun Zhao, Sourav S Bhowmick, and L Gruenwald. **CLEOPATRA: Evolutionary Pattern-based Clustering of Web Usage Data**. In *PAKDD*, 2006.

30. Qiankun Zhao, Sourav S Bhowmick, and A. Sun. **iWED: An Integrated Multigraph Cut-based Approach for Detecting Events from a Website**. In *PAKDD*, Singapore, 2006.

31. Qiankun Zhao, Sourav S. Bhowmick, and L Gruenwald. **WAM-Miner: In the Search of Web Access Motifs from Historical Web Log Data**. In ACM CIKM, Bremen, Germany, 2005.

32. Q Zhao, Sourav S Bhowmick, L Gruenwald. **Mining Conserved XML Query Paths for Dynamic-Conscious Caching.** In ACM CIKM, 2005.

33. Qiankun Zhao, Sourav S. Bhowmick, M Mohania, and Y Kambayashi. **Discovering Frequently Changing Structures from Historical Structural Deltas of Unordered XML Documents** . In ACM CIKM, Washington D.C, 2004.

34. Qiankun Zhao, Prasenjit Mitra, Bi Chen. **Temporal and Information Flow Based Event Detection from Social Text Streams.** In AAAI 2007.

35. Q Zhao, Sourav S Bhowmick, X Zheng, K Yi. **Characterizing and Predicting Community Members from Evolutionary and Heterogeneous Networks.** In ACM CIKM, 2008.