Syllabus

Graph and Web Mining - Motivation, Algorithms and Applications

Prof. Ehud Gudes, Ben-Gurion University, Israel

Whereas data-mining in structured data focuses on frequent data values, in semi-structured and graph data mining, the issue is frequent labels and common specific topologies. Here, the structure of the data is just as important as its content.
We study the problem of discovering typical patterns of graph data. The discovered patterns can be useful for many applications, including: compact representation of the information contained in a source, a road map for browsing and querying information sources, finding common structures of strongly connected groups in social networks and in several scientific domains like finding frequent molecular structures.

The discovery task is impacted by structural features of semistructured data in a non-trivial way, making traditional data mining approaches inapplicable. Difficulties result from the complexity of some of the required sub-tasks, such as graph and sub-graph isomorphism. In recent years the topic arises much interest, and there is an annual workshop: International Workshop on Mining Graphs, Trees and Sequences dedicated to it.

This course will discuss first the motivation and possible applications of Graph mining, and then will survey in detail the recent algorithms for this task, including: FSG, GSPAN and other recent algorithms by the Presentor. The differences between graph mining in the single graph setting and in the transaction setting will be described, and the problematic issue of Support in the single graph setting will also be discussed in detail. Next the use of graph mining for indexing and searching will be discussed. The last part of the course will deal with Web mining. Graph mining is central to web mining because the web links form a huge graph and mining its preoperties has large significance. Finally, applications of graph mining in several other areas will be described and future work and directions for further research on this subject will be outlined.

Course Outline

The outline of the course is as follows:

1. **Introduction** (1 hour)
   - Overview of Data mining and basic algorithms for Item-sets and Sequences
   - Motivation and Applications for Graph mining
2. **Algorithms for Graph mining** (4 hours)
   - FSG and GSPAN
   - Single vs. multiple transaction setting. The support issue in single graph setting
   - The edge-disjoint algorithm
   - New algorithms
3. **Searching graphs and Related algorithms** (3 hours)
   - Sub-graph isomorphism (Subsea)
   - Indexing and searching – graph indexing
   - A new sequence mining algorithm
4. **Web mining and other Applications** (4 hours)
   - Document classification
   - Web mining
   - Short students project presentations

5. **References**

[1] T. Washio and H. Motoda, "State of the art of graph-based data mining", SIGKDD Explorations, 5:59-68, 2003

[2] X. Yan and J. Han, "gSpan: Graph-Based Substructure Pattern Mining", ICDM'02

[3] X. Yan and J. Han, "CloseGraph: Mining Closed Frequent Graph Patterns", KDD'03

[4] X. Yan, P. S. Yu, and J. Han, “Graph Indexing: A Frequent Structure-based Approach”, SIGMOD'04

[5] M. Kuramochi, G. Karypis, "An Efficient Algorithm for Discovering Frequent Subgraphs" IEEE TKDE, September 2004 (vol. 16 no. 9)

[6] V. Lipets , N. Vanetik and E. Gudes **Subsea: an efficient heuristic algorithm for subgraph isomorphism**, Data Mining and Knowledge Discovery, Volume 19, Number 3, December, 2009
[7] Natalia Vanetik, Solomon Eyal Shimony, Ehud Gudes.**Support measures for graph data.** ; Data Min. Knowl. Discov. 13(2): 243-260 (2006)

[8] Ehud Gudes, Solomon Eyal Shimony, Natalia Vanetik **Discovering Frequent Graph Patterns Using Disjoint Paths** ; IEEE Trans. Knowl. Data Eng. 18(11): 1441-1456 (2006)

"[9] N. Vanetik and E. Gudes, Mining Frequent Labeled and Partially Labeled Graph Patterns, Proceedings of ICDE 2004: 91-102, Boston, MA, 2004

[10] J. Han and M. Kamber, Data minining – Concepts and Techniques, 2nd Edition, Morgan kaufman Publishers, 2006

[11] Bing Liu, Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data, Springer publishing, 2009

Course Requirements

The main requirement of this course (in addition to attending lectures) is a final project or a final paper to be submitted a month after the end of the course. In addition a small homework will be required to be submitted on Thursday May 27th. Homework is to be submitted individually.

In the **final project** the students (mostly 2) will implement one of studied graph mining algorithms and will test it on some public available data. In addition to the software , a report detailing the problem, algorithm, software structure and test results is expected.

In the **final paper** the student(mostly 1) will review at least two recent papers in graph mining not presented in class and explain them in detail.

Topics for projects and papers will be presented during the course.

The last hour of the course will be dedicated for students for presenting their selected project/paper. Presentation will be 7 -10mins.

Computation of grade:

1) Homework – 15%
2) Project/paper presentation – 5%
3) Project/paper – 80%