Comparative Analysis of Data Mining Tools and Techniques

N. Abinaya[#1], B.Bensujitha[#2]
Kalasalingam Academy of Research and Education, udhayasri95@gmail.com
Kalasalingam Academy of Research and Education, bensujitha@gmail.com

Abstract: - Data mining is the process of discovering patterns in large data sets. It is an emerging research field of computer science, is defined as the significant process of identifying suitable, prospectively useful, and crucially understandable patterns in data. The overall goal of the data mining process is to transform it into an understandable structure for further use and extract information from a data set. Graph mining is the process of analyzing and gathering the data represented as graphs. The incredible rising of on-line social networks gives a new and very strong interest to the set of techniques developed since several decades to mining graphs and social networks. In this paper we overviewed graph mining tasks and the tools which are used for the mining of data represented as graphs.

Keywords: Data Mining, Graphs, Graph Mining, Gephi, Network X.

1. Introduction:

Data mining is a term from computer science. Sometimes it is also called knowledge discovery in databases (KDD). Although Tim Berners-Lee envisioned a read/write Web (the very first browser also worked as an HTML editor), the Web was a read-only medium for a majority of users. The web of the 1990s was much like the combination of a phone book and the yellow pages (a mix of individual postings and corporate catalogs) and

despite the connecting power of hyperlinks it instilled little sense of community among its users. This passive attitude toward the Web was broken by a series of changes in usage patterns and technology that are now referred to as Web 2.0, a buzzword coined by Tim O'Reilly.

In the following, we summarize the history and the defining aspects of Web 2.0. The set of innovations in the architecture and usage patterns of the Web led to an entirely different role of the online world as a platform for intense communication and social interaction. A recent major survey based on interviews with 2200 adults shows that the internet significantly improves Americans' capacity to maintain their social networks despite early fears about the effects of diminishing real life contact. The survey confirms that not only networks are maintained and extended online, but they are also successfully activated for dealing with major life situations such as getting support in case of a major illness, looking for jobs, informing about major investments etc.

The first wave of socialization on the Web was due to the appearance of blogs, wikis and other forms of web-based communication and collaboration. Blogs and wikis attracted mass popularity from around 2003[1]. What they have in common is that they both significantly lower the requirements for adding content to the Web: editing blogs and wikis did not require any knowledge of HTML any more. Blogs and wikis allowed individuals and groups to claim their personal space on the Web and fill it with content at relative ease.

1.1 Social Network:

Network analysis, a branch of sociology and mathematics that is increasingly applied also to questions outside the social domain. By no means do we expect to provide a complete coverage of any topic involved. For a more encyclopedic treatment of network analysis we refer the reader to the social network analysis reference of Wasserman and Faust[2]. Network analysis should appeal to all as one of the most formalized branches of Social Science. Most of these formalisms are based on the simple nodes and edges representations of social networks to which a large array of measures and statistics can be applied.

What is network analysis?

Social Network Analysis (SNA) is the study of social relations among a set of actors. The key difference between network analysis and other approaches to social science is the focus on relationships between actors rather than the attributes of individual actors.

Network analysis takes a global view on social structures based on the belief that types and patterns of relationships emerge from individual connectivity and that the presence(or absence) of such types and patterns have substantial effects on the network and its constituents. In particular, the network structure provides opportunities and imposes constraints on the individual actors by determining the transfer or flow of resources (material or immaterial) across the network.

The focus on relationships as opposed to actors can be easily understood by an example. When trying to predict the performance of individuals in a scientific community by some measure (say, number of publications), a traditional social science approach would dictate to look at the attributes of the researchers such as the amount of grants they attract, their age, the size of the team they belong to etc. A statistical analysis would then proceed by trying to relate these attributes to the outcome variable, i.e. the number of publications.

In the same context, a network analysis study would focus on the interdependencies within the research community. For example, one would look at the patterns of relationships that scientists have and the potential benefits or constraints such relationships may impose on their work. For example, one may hypothesize that certain kinds of relationships arranged in a certain pattern may be beneficial to performance compared to the case when that pattern is not present. The patterns of relationships may not only be used to explain individual performance but also to hypothesize their impact on the network itself (network evolution). Attributes typically play a secondary role in network studies as control variables.

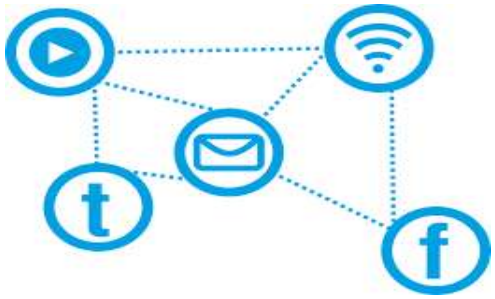

Fig.1.Social Network

1.1 Community Detection:

Relations of real-world entities are often represented as networks, such as social networks connected with friendships or co-authorships. In many cases, real social networks contain denser parts and sparser parts. Denser sub networks correspond to groups of people that are closely connected with each other. Such denser sub-networks are called “communities”. Detecting communities from given social networks are practically important for the following reasons:

1. Communities can be used for information recommendation because members of the communities often have similar tastes and preferences. Membership of detected communities will be the basis of collaborative filtering[3].

2. Communities will help us understand the structures of given social networks. Communities are regarded as components of given social networks, and they will clarify the functions and properties of the networks.

3. Communities will play important roles when we visualize large-scale social networks. Relations of the communities clarify the processes of information sharing and information diffusions, and they may give us some insights for the growth the networks in the future.

In general, detecting communities is not an easy task. There are many definitions of communities and most of them are computationally expensive. For example, detecting communities by optimizing modularity is NP-hard.

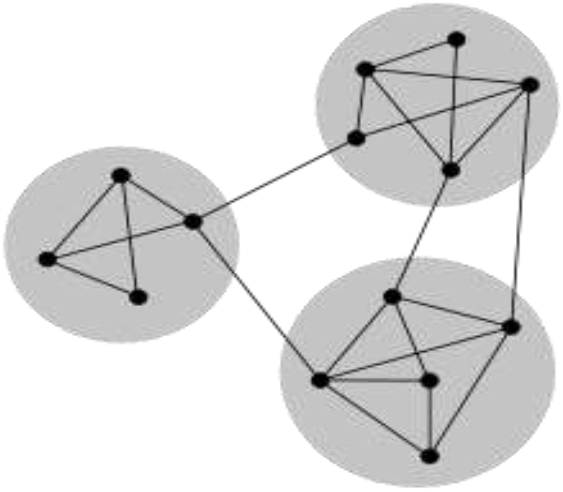

Fig.2 Community Detection

2. *Graph Mining Techniques and Tools:*

2.1 Techniques:

The main approaches which are followed for mining of graph data are

- Mining frequent sub graphs
- Classification
- Clustering

2.1.1 Mining Frequent Sub Graphs:

Frequent sub graphs, are sub graphs that occur frequently in data represented as graphs. Frequent Sub graph Mining (FSM) is the graph mining essence. To extract all the frequent sub graphs, in a given data set, whose occurrence counts are above a specified threshold is the objective of FSM[4]. A substructure may be different structural forms such as trees, graphs or lattices, which may be combined with item sets or subsequences. Although graph mining may include mining frequent sub graph patterns, clustering, graph classification and other analysis tasks.

2.1.2 Classification:

Classification is a general process related to categorization, the process in which ideas and objects are differentiated, recognized and understood. A classification system is an approach to manage classification. There are various other methods for constructing classification models, such as Naive Bayesian classification, k-nearest neighbor classification and support vector machines.

2.1.3 Clustering:

Clustering or Cluster analysis is the task of grouping a set of objects in such a way that objects in the same group are more similar (in some sense or another) to each other than to those in other groups. In other words, similar objects are grouped in one cluster and dissimilar objects are grouped in another cluster. This problem is challenging because of the need to match the structures of the underlying graphs, and use these structures for clustering purposes.

2.2 Tools:

There are several tools available for graph mining. Some of them are given here.

2.2.1 Cytoscape:

Cytoscape is open source software platform for visualizing with integrating and interaction networks with gene expression profiles and other state data. Plug in are available for molecular profiling analyses, new layouts, additional file format support, networks and connection with databases and searching in large networks[6]. Plug in may be developed using the Cytoscape open Java software architecture by anyone and plug in community development is encouraged. Cytoscape can be used to visualize and analyze network graphs of any kind involving edges (e.g., social networks). A key aspect of the software architecture of Cytoscape is the use of plug in for specialized features.

Gephi:

Gephi is an open-source software for network analysis and visualization. It helps data analysts to intuitively reveal patterns and trends, highlight outliers and tells stories with their data. Classic metrics of social network analysis, such as node degree or betweens centrality measures, can be computed and used in the visualization as well. The network can also be altered based on attributes.

2.2.2 Graph Insight:

Graph Insight is a visualization software that lets you explore graph data through high quality interactive representations[7]. Data exploration and knowledge extraction from graphs is of great interest nowadays. Humans are extremely good in identifying outliers and patterns. Graph Insight is useful for interacting visually with the data can give us a better intuition and higher confidence on the field.

2.2.3 Network X:

Network X is a Python package for the creation, study of the structure, dynamics manipulation and functions of complex networks. Elasticity ideal for representing networks found in many different fields.

2.2.5 *Social Networks Visualizer:*

Social Network Visualizer (SocNetV) is a cross-platform, user-friendly free software application for social network visualization and analysis. Analyze graph and social network properties, produce beautiful HTML reports and embed visualization layouts to the network.

2.2.6 Knime:

KNIME, the Konstanz Information Miner, is an open source data analytics, reporting and integration platform. KNIME integrates various components for data mining and machine learning through its modular data pipelining concept.

3. Issues:

- Scalability
- Data Ownership and Distribution
- Dimensionality
- Privacy Preservation
- Streaming Data
- Elasticity

4. Conclusion:

In this paper we briefly discuss about the graph mining techniques, tools and issues from its initiation to the upcoming research. This paper provides a new perspective of a researcher to overcome the challenges in methods, data and other issues of graph mining in social network. We also provide tools of social network analysis .Tools are used to identify the graph mining techniques.

5. References:

[1] Aggarwal, C.: “An introduction to social network data analytics”, Springer US, 2011.

[2] Aiello, L. M., Petkos, G., Martin, C. et al.: “0Sensing trending topics” in Twitter. 2013.

[3] Asur, S., and Huberman, B.: "Predicting the future with social network." Web Intelligence and Intelligent Agent Technology (WIIAT)”, 2010 IEEE/WIC/ACM International Conference on. Vol. 1. IEEE, 2010.

[4] Au Yeung, C. M., and Iwata, T.: “Strength of social influence in trust networks in product review sites”. In Proceedings of the fourth ACM international conference on Web search and data mining (pp. 495-504). ACM, 2011.

[5] Bakshy, E., Hofman, J. M., Mason, W. A., Watts, D. J.: “Identifying influencers on twitter”. In Fourth ACM International Conference on Web Seach and Data Mining (WSDM), 2011.

[6] Becker, H., Naaman, M., Gravano, L.: Beyond Trending Topics: “Real-World Event Identification on Twitter.” ICWSM, 11, 438-441, 2011.

[7] Becker, H., Chen, F., Iter, D., Naaman, M., Gravano, L.: “Automatic Identification and Presentation of Twitter Content for Planned Events”. In ICWSM, 2011.