Analysis and study of Mining of Social Network Data

[1]Nirav S. Shukla ,[2]Nirav K. Dave,[3]Parthik R.Patel, [4]Tejal S.Pandya [5]Parag B. Makwana

[1]Assistant Professor, [2]Assistant Professor, [3]Assistant Professor, [4]Assistant Professor, [5]Assistant Professor
[1]Shri Chimanbhai Patel Institutes of Computer Applications,
Ahmedabad, India
[2]Nirav K. Dave
[2]Shree Swaminarayan College of Compter Science , Bhavnagar,India
[3]Parthik R. Patel
[2]Shree Swaminarayan College of Compter Science , Bhavnagar,India
[4]Tejal S. Pandya
[2]Shree Swaminarayan College of Compter Science , Bhavnagar,India
[5]Parag B. Makwana
[2]Shree Swaminarayan College of Compter Science , Bhavnagar,India

Abstract : Data mining is a great innovative technology which helps corporate to focus on the important information in the data of there storage. Data mining is used in various machine learning, statistical and in graphical methods.

***IndexTerms* – Mining, Social Network, OLAP, Node**

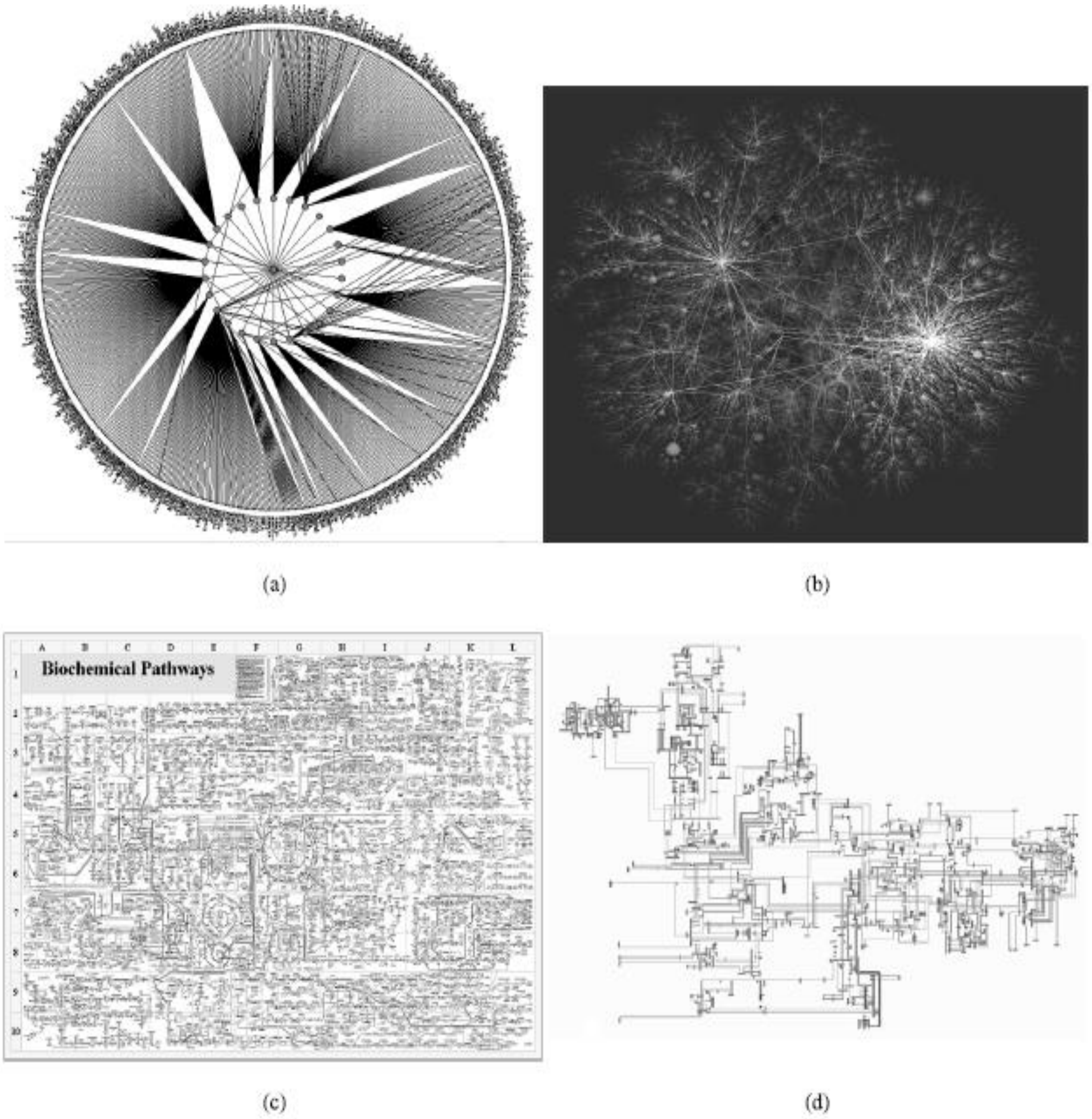
I. Introduction OF Mining

Data Mining is the extraction or “Mining” of knowledge from a large amount of data or data warehouse. To do this extraction data mining combines artificial intelligence, statistical analysis and database management systems to attempt to pull knowledge form stored data. Data mining is the process of applying intelligent methods to extract data patterns. This is done using the front-end tools. Data mining is the process of applying intelligent methods to extract data patterns. This is done using the front-end tools. The spreadsheet is still the most compiling front-end application for Online Analytical Processing (OLAP). The challenges in supporting a query environment for OLAP can be crudely summarized as that of supporting spreadsheet operation effectively over large multi-gigabytes databases. To distinguish information extraction through data mining from that of a traditional database querying, the following main observation can be made. In a database application the queries issued are well defined to the level of what we want and the output is precise and is a subset of operational data. In data mining there is no standard query language and the queries are poorly defined. Thus the output is not precise (fuzzy) and do not represent a subset of the database. Beside the data used not the operational data that represents the today transactions.

II What Is a Social Network & how its work?

From the point of view of data mining, a social network is a heterogeneous and multi relational data set represented by graph. The graph is typically very large, with nodes corresponding to objects and edges corresponding to links representing relationships or interactions between objects. Both nodes and links have attributes. Objects may have class labels. Links can be one-directional and are not required to be binary. Social networks need not be social in context. There are many real-world instances of technological, business, economic, and biologic social networks. Examples include electrical power grids, telephone call graphs, the spread of computer viruses, the World Wide Web, and co authorship and citation networks of scientists. Customer networks and collaborative filtering problems (where product recommendations are made based on the preferences of other customers) are other examples. In biology, examples range from epidemiological networks, cellular and metabolic networks, and food webs, to the neural network of the nematode worm Caenorhabditis elegans (the only creature whose neural network has been completely mapped). The exchange of e-mail messages within corporations, newsgroups, chat rooms, friendships, sex webs (linking sexual partners), and the quintessential “old-boy” network (i.e., the overlapping boards of directors of the largest companies in the United States) are examples from sociology.

Small world (social*) networks* have received considerable attention as of late. They reflect the concept of “small worlds,” which originally focused on networks among individuals. The phrase captures the initial surprise between two strangers (“What a small world!”) when they realize that they are indirectly linked to one another through mutual acquaintances. In 1967, Harvard sociologist, Stanley Milgram, and his colleagues conducted experiments in which people in Kansas and Nebraska were asked to direct letters to strangers in Boston by forwarding them to friends who they thought might know the strangers in Boston. Half of the letters were successfully delivered through no more than five intermediaries. Additional studies by Milgram and others, conducted between other cities, have shown that there appears to be a universal “six degrees of separation” between any two individuals in the world. Examples of small world networks are shown in Figure. Small world networks have been characterized as having a high degree of local clustering for a small fraction of the nodes (i.e., these nodes are interconnected with one another), which at the same time are no more than a few degrees of separation from the remaining nodes. It is believed that many social, physical, human-designed, and biological networks exhibit such small world characteristics.



(a) (b)

(c) (d)

Figure of Real-world examples of social networks: (a) science coauthor network, (b) connected pages ona part of the Internet, (c) biochemical pathway network, and (d) New York state electric power grid.

III Characteristics of Social Networks

These may be used to predict how a network may look in the future, answering "what-if" questions. Taking the Internet as an example, we may ask *"What will the Internet look like when the number of nodes doubles?"* and *"What will the number of edges be?"*. If a hypothesis contradicts the generally accepted characteristics, this raises a flag as to the questionable plausibility of the hypothesis. This can help detect abnormalities in existing graphs, which may indicate fraud, spam, or Distributed Denial of Service (DDoS) attacks. Models of graph generation can also be used for simulations when real graphs are excessively large and thus, impossible to collect (such as a very large network of friendships). In this section, we study the basic characteristics of social networks as well as a model for graph generation. *"What qualities can we look at when characterizing social networks?"* Most studies examine the nodes' degrees, that is, the number of edges incident to each node, and the *distances* between a pair of nodes, as measured by the *shortest path length*. (This measure embodies the small world notion that individuals are linked via short chains.) In particular, the network diameter is the maximum distance between pairs of nodes. Other nodeto- node distances include the average distance between pairs and the effective diameter (i.e., the minimum distance, *d*, such that for at least 90% of the reachable node pairs, the path length is at most *d*). Social networks are rarely static. Their graph representations evolve as nodes and edges are added or deleted over time. In general, social networks tend to exhibit the following phenomena:

1. Densification power law: Previously, it was believed that as a network evolves, the number of degrees grows linearly in the number of nodes. This was known as the *constant average degree assumption*.However, extensive experiments have shown that, on the contrary, networks become increasingly *dense* over time with the average degree increasing (and hence, the number of edges growing super linearly in the number of nodes). The densification follows the densification power law (or growth power law),which states $e(t)$ µ $n(t)a$; (9.1) where $e(t)$ and $n(t)$, respectively, represent the number of edges and nodes of the graph at time t, and the exponent a generally lies strictly between 1 and 2. Note that if $a = 1$, this corresponds to constant average degree over time, whereas $a = 2$ corresponds to an extremely dense graph where each node has edges to a constant fraction of all nodes.

2. Shrinking diameter: It has been experimentally shown that the effective diameter tends to *decrease* as the network grows. This contradicts an earlier belief that the diameter slowly increases as a function of network size. As an intuitive example, consider a citation network, where nodes are papers and a citation from one paper to another is indicated by a directed edge. The out-links of a node, *v* (representing the papers cited by *v*), are "frozen" at the moment it joins the graph. The decreasing distances between pairs of nodes consequently appears to be the result of subsequent papers acting as "bridges" by citing earlier papers from other areas.

3. Heavy-tailed out-degree and in-degree distributions: The number of out-degrees for a node tends to follow a heavy-tailed distribution by observing the power law, $1=na$, where n is the rank of the node in the order of decreasing out-degrees and typically, $0 < a < 2$ (Figure 9.17). The smaller the value of a, the heavier the tail. This phenomena is represented in the preferential attachment model, where each new node attaches to an existing network by a constant number of out-links, following a

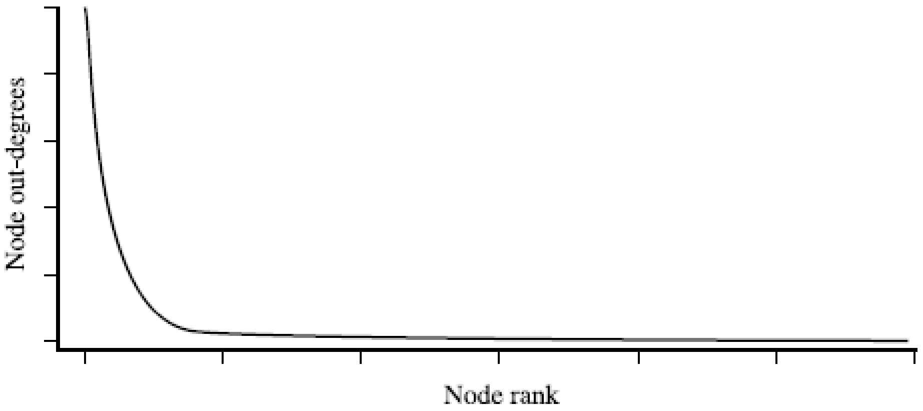


Figure :The number of out-degrees (*y*-axis) for a node tends to follow a heavy-tailed distribution.The node rank (*x*-axis) is defined as the order of deceasing out-degrees of the node.

IV MINING OF SOCIAL NETWORKS

In this section, we explore exemplar areas of mining on social networks, namely, link prediction, mining customer networks for viral marketing, mining newsgroups using networks, and community mining from multirelational networks. Other exemplars include characteristic subgraph detection. and mining link structures on the Web.

A. Link Prediction: What Edges Will Be Added to theNetwork?

Approaches to link prediction have been proposed based on several measures for analyzing the "proximity" of nodes in a network. Many measures originate from techniques in graph theory and social network analysis. Thegeneral methodology is asfollows: All methods assign a connection weight, *score*(***X***, ***Y***), to pairs of nodes, ***X*** and ***Y***, based on the given proximity measure and input graph, *G*. A ranked list in decreasing order of *score*(***X***, ***Y***) is produced. This gives the predicted new links in decreasing order of confidence. The predictions can be evaluated based on real observations on experimental data sets.

B. MINING CUSTOMER NETWORK FOR VIRAL MARKETING

Viral marketing is an application of social network mining that explores how individuals can influence the buying behavior of others. Traditionally, companies have employed direct marketing (where the decision to market to a particular individual is based solely on her characteristics) or mass marketing (where individuals are targeted based on the population segment to which they belong). These approaches, however, neglect the influence

that customers can have on the purchasing decisions of others. For example, consider a person who decides to see a particular movie and persuades a group of friends to see the same film. Viral marketing aims to optimize the positive word-of-mouth effect among customers. It can choose to spend more money marketing to an individual if that person has many social connections. Thus, by considering the interactions between customers, viral marketing may obtain higher profits than traditional marketing, which ignores such interactions.

REFERENCES

[1] Data mining techniques : Jiawei Han
[2] Data mining : Doglous rechards