An Analysis Of Link Strength In Social Networks

R. Hema Latha
Mphil Scholar
Department Of Computer Science
PSGR Krishnammal College For Women
Coimbatore

K .Sathiya Kumari
Associate Professor
Department Of Computer Science
PSGR Krishnammal College For Women
Coimbatore

Abstract

A social structure made of nodes that are generally individuals or organizations. A social network represents relationships and flows between people, groups, organizations, animals, computers or other information or knowledge processing entities. Social networking websites allow users to be part of a virtual community. Data mining, the extraction of hidden predictive information from large databases, is a powerful new technology with great potential to help companies focus on the most important information in their data warehouses. Link prediction in Facebook and Twitter can be done at a familiar class of graph generation model, where the nodes are united with locations in a latent metric space and connections are most likely between closer nodes. In this paper, GEPHI and NODEXL tools are used for the comparison measures to predict betweenness centrality of particular user's account in Facebook and Twitter.

Keywords

Facebook,; Twitter; Gephi; Average degree; Metrics; Page Rank; Centrality; NodeXL.

1. Introduction

Data mining is the process that attempts to discover patterns in large data sets. It utilizes methods at the intersection of artificial intelligence, machine learning, statistics, and database systems. The overall goal of the data mining process is to extract information from a data set and transform it into an understandable structure for further use. Aside from the raw analysis step, it involves database and data preprocessing, model and inference considerations, interestingness metrics, complexity considerations, post-processing of discovered structures, visualization, and online updating. The actual data mining task is the automatic or semi-automatic analysis of large quantities of data to extract previously unknown interesting patterns such as groups of data records (cluster analysis) , unusual records (anomaly detection) and dependencies (association rule mining). This usually involves using database techniques such as spatial indexes. These patterns can then be seen as a kind of summary of the input data, and may be used in further analysis or, for example, in machine learning and predictive analytics. For example, the data mining step might identify multiple groups in the data, which can then be used to obtain more accurate prediction results by a decision support system.

Data mining techniques are the result of a long process of research and product development. This evolution began when business data was first stored on computers, continues with improvements in data access, and more recently, generated technologies that allow users to navigate through their data in real time. Data mining takes this evolutionary process beyond retrospective and proactive information delivery. Data mining is ready for application in the business community because it is supported by three technologies that are now sufficiently mature:

- Massive data collection
- Powerful multiprocessor computers
- Data mining algorithms

Data Mining applications are computer software programs or packages that enable the extraction and identification of patterns from stored data. Data mining is widely used by companies and public bodies for such uses as marketing, detection of fraudulent activity, and scientific research. There are wide varieties of data mining applications available, particularly for business uses, such as Customer Relationship Management (CRM). These applications enable marketing managers to understand the behaviors of their customers and also to predict the potential behavior of prospective clients.

Data mining applications are often structured around the specific needs of an industry sector or even tailored and built for a single organization. This is because the patterns within data may be very specific. Banking data mining applications may, for example, need to track client spending habits in order to detect unusual transactions that might be fraudulent.

Social Network Analysis and Mining (SNAM) is intended to be a multidisciplinary journal to serve both academia and industry as a main venue for a wide range of researchers and readers from computer science, social sciences, mathematical sciences, medical and biological sciences. Social network analysis and mining using different techniques from sociology, social sciences, mathematics, static's and computer science.

The main areas covered by SNAM include: (1) data mining advances on the revelation and anatomizing of communities, personalization for solitary (like discovery of potential friends), the anatomizing of user behavior in open forums (like conventional sites, blogs and forums) and in commercial platforms (like e-auctions), and the associated security and privacy-preservation challenges; (2) social network modeling, construction of scalable, growth, and evolution patterns using machine learning approaches or multi-agent based simulation.

Some analysis of social network analysis is: (1) content analysis (2) web metrics analysis (3) sentiment and affect analysis (4) video analysis. SNA techniques are used to examine website and forum posting relationships. Various topological metrics (betweenness, degree, etc.) and properties (preferential attachment, growth etc.). There are several clustering (e.g., block modeling) and projection (e.g., multi-dimensional scaling, spring embedded) techniques to visualize their relationships.

Facebook is a social networking website that was originally designed for college students, but is now open to anyone 13 years of age or older. Facebook users can create and customize their own profiles with photos, videos, and information about themselves. Friends can browse the profiles of other friends and write messages on their pages.

Each Facebook profile has a "wall", where friends can post comments. Since the wall is viewable by all the user's friends, wall postings are basically a public conversation. Therefore, it is usually best not to write personal messages on the friend's walls. Instead, can send a person a private message, which will show up in their private inbox, similar to an e-mail message.

Facebook allows each user to set privacy settings, which by default are pretty strict. Another feature of Facebook, which makes it different from MySpace, is the ability to add applications to the profile. Facebook applications are small programs developed specifically for Facebook profiles. Facebook provides an easy way for friends to keep in touch and for individuals to have a presence on the Web without needing to build a website. Since Facebook makes it easy to upload pictures and videos, nearly anyone Can publish a multimedia profile.

Twitter is an online service that allows to share updates with others users by answering one simple question: "what are you doing?". Twitter limits each tweet to 140 characters, which means there is no room for rambling. The character limit is also within the 160 character limit of SMS text messages. Many people also use Twitter to blog about the news, politics, TV shows, or any other hot topic. Some people even use it to share their thoughts on lecturers or sermons.

Link prediction is one of the challenging research topics in social network. There are two main data sources for predicting links between nodes: 1) attributes of nodes, and 2) structural properties of networks that connect nodes. In the case of online social networks, nodes represent users and their attributes (personal information) are not always available. The latter data source (structural properties) is preferable for the purpose of predicting links of online social networks.

1.1.Gephi

Gephi is an interactive visualization and exploration platform for all kinds of networks and complex systems, dynamic and hierarchical graphs. Gephi is a tool for people that have to explore and understand graphs. Like Photoshop but for data, the user interacts with the representation, manipulate the structures, shapes and colors to reveal hidden properties. The goal is to help data analysts to make hypothesis, intuitively discover patterns, isolate structure singularities or faults during data sourcing. It is a complementary tool to traditional statistics. This is a software for Exploratory Data Analysis. Gephi tool provides a fastest graph visualization engine to speed-up understanding and pattern discovery in large graphs.

Gephi is powered by a ad-hoc OpenGL engine, it is pushing the envelope on how interactive and efficient network exploration can be.

- Networks up to 50,000 nodes and 1,000,000 edges
- Iterate through visualization using dynamic filtering
- Rich tools for meaningful graph manipulation

Gephi is a modular software and can be extended with plug-ins. Plug-ins can add new features like layout, filters, metrics, data sources, etc. or modify existing features. Gephi is written in Java so anything that can be used in Java can be packaged as a Gephi plug-in.

1.2.NodeXL

NodeXL: Network Overview, Discovery and Exploration for Excel. It is a free, open-source template for Microsoft Excel 2007 and 2010 that makes it easy to explore network graphs. With NodeXL, it can enter a network edge list in a worksheet, it displays the entire graph with all in the familiar environments of the Excel window.

1.2.1.NodeXL Features

- *Flexible Import and Export:* An import and export graph is GraphML, Pajek, UCINet, and matrix formats.
- *Direct Connections to Social Networks:* Import social networks directly from Twitter, YouTube, Flicker and email, or use one of several available plug-ins to get networks from Facebook, Exchange and WWW hyperlinks.
- *Zoom and Scale:* Zoom into areas of interest, and scale the graph's vertices to reduce clutter.
- *Flexible Layout:* Use one of several "force-directed" algorithms to layout the graph, or drag vertices around with the mouse. Have NodeXL move all of the graph's smaller connected components to the bottom of the graph to focus on the importance.
- *Easily Adjusted Appearance:* Set the color, shape, size, label, and opacity of individual vertices by filling in worksheet cells, or NodeXL will perform based on vertex attributes such as degree, betweenness centrality or PageRank.
- *Dynamic Filtering:* Instantly hide vertices and edges using a set of sliders-hide all vertices with degree less than five.
- *Powerful Vertex Grouping:* Group the graph's vertices by common attributes, or have NodeXL analyze their connectedness and automatically group them into clusters. Make groups distinguishable using shapes and color, collapse them with a few clicks, or put each group in its own box within the graph. "Bundle" intergroup edges to make them more manageable.
- *Graph Metric Calculation:* Easily calculate degree, betweenness centrality, closeness centrality, eigenvector centrality, PageRank, Clustering Coefficient, graph density and more.

The NodeXL template displays graphs using a custom Windows Presentation Foundation (WPF) control that can be reused in custom applications. In fact, the template is just an application wrapper around a set of reusable, prebuilt class libraries. It is used to create a .NET assembly that will import graph data from a custom source into the NodeXL template. These "plug-ins" appear in NodeXL's Data, Import menu.

2. Methodology

2.1.Gephi

Gephi offers a set of powerful calculation tools that allow exploring the qualities of particular user network. The in-degree and out-degree is the basic and useful calculation in network nodes. The Gephi tool is used to analyze a strong link of particular user account in Facebook. The particular user invokes 243 nodes and 2055 edges in Facebook account. The k-core algorithm is implemented in Gephi tool to predict strong link in particular user of Facebook account.

2.1.1.Network Diameter

The average graph-distance between all pairs of nodes. Connected nodes have graph distance 1. The diameter is the longest graph distance between any two nodes in the network (i.e., how far apart are the two most distant nodes).

2.1.2.Betweenness Centrality

It measures how often a node appears on the shortest paths between nodes in the network.

2.1.3.Closeness Centrality

The average distance from a given starting node to all other nodes in the network.

2.1.4.Eccentricity

The distance from a given starting node to the farthest node from it in the network.

2.1.5.Hits

It computes two separate values for each node. The first value (called Authority); it measures how valuable information stored at that node is. The second value (called Hub); it measures the quality of the node links.

2.1.6.Average Clustering Coefficient

The clustering coefficient, along with the mean shortest path, can indicate a "small-world" effect. It indicates how nodes are embedded in their neighborhood. The average gives an overall indication of the clustering in the network.

2.1.7.Eigenvector Centrality

It is a measure of node importance in a network based on a node's connections.

2.2. NodeXL

NodeXL insights about a person's position within the network, helping to identify important or "central" people: analyst and managers can better know who to contact or influence or bring to the table when trying to implement new programs or gain broader understanding, it identify cliques or persistent social roles that show up in many communities. Clustering can help identify competing or complementary groups, potential allies to form a powerful group, and individuals who can connect to a new group.

2.2.1.Graph Metrics

- Degree: In an undirected graph, a vertex's degree is the number of edges incident to the vertex. A self-loop in an undirected graph is counted twice when a vertex's degree is calculated.
- In-degree: In an directed graph, a vertex's in-degree is the number of incoming edges incident to the vertex. In an undirected graph, in-degree is undefined and is not calculated. A self-loop in a directed graph is counted once as an incoming edge (in-degree) and once as an outgoing edge (out-degree).
- Out-degree: In a directed graph, a vertex's out-degree is the number of outgoing edges incident to the vertex. In an undirected graph, out-degree is undefined and is not calculated. A self-loop is a directed graph is counted once as an outgoing edge (out-degree) and once as an incoming edge (in-degree).

2.2.2.Betweenness and Closeness Centrality

- Group Metrics: In a directed graph, an edge from vertex A to vertex B is reciprocated, if the graph also has an edge from B to A. In an undirected graph, edge reciprocation is undefined and not calculated.
- Overall Graph Metrics: It includes vertex counts, edge counts, geodesic distances, graph density and modularity.
- Subgraph Images: It creates an image of each vertex's subgraphs. The image can be saved as files in a folder or inserted as thumbnails into vertices worksheet.
- Groups: It is grouped by vertex attribute.

2.2.3.Fruchterman-Reingold Algorithm

The Fruchterman - Reingold Algorithm is a force-directed layout algorithm. The idea of a force directed layout algorithm is to consider a force between any two nodes. In this algorithm, the nodes are represented by steel rings and the edges are springs between them. The attractive force is analogous to the spring force and the repulsive force is analogous to the electrical force. The basic idea is to minimize the nodes and changing the forces between them.

In this algorithm, the sum of the force vectors determines which direction a node should move. The step width, which is a constant determine how far a node moves in a single step. When the energy of the system is minimized, the nodes stop moving and the system reaches its equilibrium state. The drawback of this is that if it defines a constant step width, there is no guarantee that the system will reach equilibrium at all. T.M.J. Fruchterman and E.M. Reingold introduced a "global temperature" that controls the step width of node movements and the algorithm's termination.

According to this layout algorithm, it performs layout of unweighted graph. Unlike the Kamada-Kawai layout algorithm, this algorithm directly supports the layout of disconnected graphs. In force-directed algorithm, the vertex layout is determined by the forces pulling vertices together and pushing them apart. Attractive forces occur between adjacent vertices only, whereas repulsive forces occur between every pair of vertices. Each iteration computes the

sum of the forces on each vertex, then moves the vertices to their new positions.

2.2.3. Data Set

The data et is collected from social network sites facebook and twitter. Here, it collected nearly three megabyte of data of individual node network from facebook and twitter. The twitter data set represents the followed and following node; the facebook data set represents the friend's connections between the friends of friends.

The NodeXL tool is used for twitter data collection. It collects followed and following nodes in depth of five levels. The facebook online API tool is to retrieve friend network data set. The instances of twitter contains 16 rows and 16 columns, and facebook contains 254 rows and 254 columns. The data has been collected from particular user ID: thottarayaswamy@gmail.com.

2.3. Predicted Network Link Result

The calculated summary values of twitter and facebook:

Table 1: Measurement of twitter and facebook values

	Twitter	Facebook
Total Number of Nodes	**49**	**243**
Total Number of Edges	**455**	**2055**
Average degree (IN degree and OUT degree)	**9.286**	**80457**
Betweenness centrality (BC)	**16.367**	**2.509**

Here, the betweenness centrality measures of twitter and facebook results 90286 and 8.457. It calculates the balanced betweenness centrality value to predict the common actor (node's) link strength by multiplying total number of nodes.

It is manipulated using the following formula:

Predicted Betweenness Centrality = (predicted BC/100)*$T_{NO.}$

Here T_{NO} is the total number of nodes. Therefore predicted twitter betweenness centrality = (predicted BC/100)*$T_{NO.}$

= (16.367/.100)*49

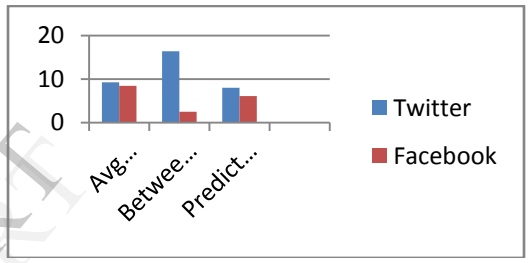
=8.01483%

Therefore predicted Facebook betweenness centrality = (predicted BC/100)*T_{NO}

= (2.509/100)*243

= 6.09689%

Here, BC denotes Betweenness Centrality and T_{NO} denote Total Number of Nodes.



Thus the comparison measures for betweenness centrality results that the particular user's account in Twitter predicts strong links between the nodes compare to that of Facebook network based on exchange of messages.

3.Conclusion

In this research paper, it examines a frame work to provide friend recommendations in Open Social Networks. It defines a new node similarity measure that exploits local and global characteristics of a network. It shows a significant accuracy improvement can be gained by using information about both positive and negative edges. It performed extensive experimental comparison of the proposed algorithmic method against two existing link prediction algorithms, using real data sets (Twitter, Facebook). It have shown that the two algorithms provides more accurate and faster friend recommendations compared to existing approaches. The proposed algorithm also outperforms the existing global-based friend recommendation algorithms in terms of time complexity.

4.Scope For Future Work

In future work, it indent to examine ways of improving friend recommendations based on other features that Open Social Networks offer. Except the friendship network, users in Open Social Networks can also form several implicit social networks through their daily interactions like co-commenting on people's post, co-rating similar products, and co-tagging people's photos. It collect real-time data in different social network websites like YouTube, LinkedIn. It can apply social network mining algorithm and community mining algorithm.

5. References

[1] Allan, Jr., E. G.; Turkett, Jr., W. H.; and Fulp, E. W. 2009. Using network motifs to identify application protocols. In Proceedings of the 28th IEEE Conference on Global Telecommunications, GLOBECOM'09, 42664272. Piscataway,NJ, USA: IEEE Press.

[2] Becchetti, L.; Boldi, P.; Castillo, C.; and Gionis, A. 2008. Efficient semi-streaming algorithms for local triangle counting in massive graphs. In Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '08, 16-24. New York, NY,USA:ACM.

[3] Benevenuto, F.; Duarte, F.; Rodrigues, T.; Almeida, V. A.;Almeida, J. M.; and Ross, K. W. 2008a. Understanding Video Interactions in YouTube In Proceedings of the 16th ACM International Conference on Multimedia, MM '08,761-764. New York,NY,USA: ACM.

[4] Benevenuto, F.; Rodrigues, T.; Almeida,V.; Almeida, J.;Zhang, C.; and Ross, K.2008b. Identifying video spammers in online social networks. In Proceedings of the 4th International Workshop on Adversarial Information Retrieval on the Web, AIRWeb '08, 45-52. New York,NY, USA: ACM.

[5] Benevenuto, F.; Rodrigues, T.; Almeida, V.; Almeida, J.; and Gonc, alves, M.2009. Detecting spammers and content promoters in online video social networks. In Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '09, 620-627. New York,NY, USA: ACM.

[6] Benevenuto, F.; Magno, G.; Rodrigues, T.; and Almeida, V.2010. Detecting Spammers on Twitter. In Proceedings of the 7th Annual Collaboration, Electronic Messaging, Anti-Abuse and Spam Conference (CEAS).

[7] Boykin, P., and Roychowdhury, V. 2005. Leveaging social networks to fight spam. Computer 38(4):61 – 68.

[8] Cheng, X.; Dale, C.; and Liu, J. 2008. Statistics and Social Network of YouTube Videos. In the 16th International Workshop on Quality of Service (IWQoS '08), 229- 238.

[9] Chhabra, S.; Aggarwal, A.; Benevenuto, F.; and Kumaraguru,P. 2011. Phi.sh/Social: The phishing landscape through short urls. In Proceedings of the 8th Annual Collaboration, Electronic Messaging, Anti-Abuse and Spam Conference, CEAS '11,92-101. New York, NY, USA: ACM.

[10] Gao, H.; Hu, J.;Wilson, C.; Li, Z.; Chen, Y.; and Zhao, B. Y. 2010. Detecting and characterizing social spam campaigns. In Proceedings of the 10th Annual Conference on Internet Meaurement, IMC '10, 35-47. New York, NY, USA: ACM.

[11] Grier, C.; Thomas, K.; Paxson, V.; and Zhang, M. 2012.@spam: The underground on 140 characters or less. In Proceedings of the 17th ACM Conference on Computer and Communications Security, CCS '10, 27-37. New York, NY, USA: ACM.

[12] Kamaliha, E.; Riahi, F.; Qazvinian, V.; and Adibi, J.2008. Characterizing Network Motifs to Identify Spam Comments. In Proceedings of the 2008 IEEE International Conference on Data Mining Workshops, 919-928. Washington, DC, USA: IEEE Computer Society.