R.Hema Latha, K.Sathiyakumari / International Journal of Engineering Research and Applications (IJERA) ISSN: 2248-9622 www.ijera.com
Vol. 2, Issue 6, November- December 2012, pp.703-707

Predicting Link Strength In Online Social Networks

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

K.Sathiyakumari
Assistant Professor
Department Of Computer Science
PSGR Krishnammal College For Women
Coimbatore

Abstract

Social Media is a term that encompasses the platforms of New Media, but also implies the inclusion of systems like Facebook, and other things typically thought of as social networking. The idea is that they are media platforms with social components and public communication channels. Social media are primarily Internet-based tools for sharing and discussing information among human beings. Data mining (the analysis step of the "Knowledge Discovery in Databases" process, or KDD), is the process that attempts to discover patterns in large data sets. 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. It involves database and data management aspects, data preprocessing, model and inference considerations, interestingness metrics, complexity considerations, post-processing of discovered structures, visualization, and online updating. 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 more likely between closer nodes. In this paper, Gephi tool is used to predict the link of Facebook.

Keywords:- Facebook, Gephi, Average Degree, Metrics, Page Rank, Centrality.

1. Introduction

Data mining is 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. Data mining tools predict future trends and behaviors, allowing business to make proactive, knowledge-driven decisions. The automated, prospective analyses offered by data mining move beyond the analyses of past events provided by retrospective tools typical of decision support systems. Data mining tools can answer business questions that traditionally were too time consuming to resolve. They scour databases for hidden patterns, finding predictive information that experts may miss because it lies outside the expectations.

Data mining inculdes six common classes of tasks:
Anomaly detection(Outlier/change/deviation detection) – The identification of unusual data records, that might be interesting or data errors and require further investigation.

Association rule learning (Dependency modeling) – Search for relationships between variables. For example a supermarket might gather data on customer purchasing habits. Using association rule learning, the supermarket can determine which products are frequently bought together and use this information for marketing purposes. This is sometimes referred to as market basket analysis.
Clustering – It is the task of discovering groups and structures in the data that are in some way or another "similar", without using known structures in the data.

Classification – It is the task of generalizing known structure to apply to new data. For example, an e-mail program might attempt to classify an e-mail as "legitimate" or as "spam".
Regression – Attempts to find a function which models the data with the least error.
Summarization – It provides a more compact representation of the data set, including visualization and report generation.

The network or graph is denoted is the collection of 'n' vertices connected by 'm' edges. A network can be directed, meaning the edgespointed in one direction or undirected, meaning the edges go in both the directions. The edges can join more than one vertices together. Such graph called as "hypergraph". Vertices represents node or people and edge represents link or tie. Social network is defined as a network of interactions or relationships where nodes consists of actors and the edges consists of relationship or interaction between these actors. Social network is a social structure made up of a set of actors (such as individuals or organizations) and the dyadic ties between these actors. The social network perspective provides a clear way of analyzing the structure of whole social entities. The study of these structures uses social network analysis to identify local and global patterns, locate influential entities, and examine network dynamics. A social network is based on certain structure that allow people to both express their individuality and meet people with similar

interest. This structure includes having profile, friends, blogs post, weights and usually something unique to that particular social networking websites such as the ability to 'poke' people on facebook . The "Profile" gives basic information about the particular user i.e., location, age, etc. The "Friends" are trusted members of the site that are allowed to post comments on the profile or send private messages. Friends are the heart and soul of social networking. The "Groups" are used to find people with similar interest or engage in discussions on certain topics. Sometimes groups are called by other names, such as 'networks' on Facebook. The "Discussion" is a primary focus of groups is to create interaction between users in the form of discussions. Most social networking websites supports discussion boards for the groups, and also allows members of the group to post picture, music, video clips, and other tidbits related to the group. The "Blog" is an another feature of some social networks, it is used to create own blog entries. Blogging through a social network is perfect for keeping people informed on what you are up to.

Facebook is more of a personal networking site as it promotes "reconnecting" with old high school, college, post graduate and former corporate friends and associates. Many have extended their networks to include commercial interests as well. Blogs, in and of itself, a form of social networking, is often linked through RSS (Really Simple Syndication) on Facebook pages so those within one's personal network can view the blogposts on the home page of the Facebookuser. It is a popular free social networking website that allows registered users to create profiles, upload photos and video, send messages and keep in touch with friends, family and colleagues. The site, which is available in 37 different languages, includes public features such as:

Groups- allows members who have common interests to find each other and interact.
Events- allows members to piblicize an event, invite guests and track who plans to attend.
Pages- allows members to create and promote a public page built around a specific topic.
Presence technology- allows members to see which contacts are online and chat.

Within each mamber's personal profile, there are several key networking components. The most popular is arguably the Wall, which is essentially a virtual bulletin board. Messages left on a member's Wall can be text, video or photos. Another popular component is the virtual Photo Album. Photos can be uploaded from the desktop or directly from a cell phone camera. There is no limitation on quantity, but facebook staff will remove inappropriate or copyrighted images. An interactive album feature allows the member's contacts (who are generically called "friends") to comment on each other's photos and identify (tag) people in the photos. Another popular profile components is Status Updates, a microblogging feature that allows members to broadcast short Twitter-like announcements to their friends. All interactions are published in newsfeed, which is distributed in real-time to the Member's friends.

Facebook offers a range of privacy options to its members. A member can make all communications visible to everyone, can block specific connections or can keep all communications private. Members can choose whether or not to be searchable, decide which parts of the profile are public, decide not to put in their newsfeed and determine exactly who can see their post. For members wished to use Facebook to communicate privately, there is a message feature, which closely resembles email.

Link prediction is spread throughout the work to uncover the missing links in the profile of real-world networks, which are usually obtained through different kinds of sampling methods. 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, node 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 hierarchial 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 an fastest graph visualization engine to speed-up understanding and pattern discovery in large graphs. Gephi is powered by 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.

1.1.1.Layout

Layout algorithm give the shape to the graph. Gephi provides state-of-art algorithms and layout algorithms, both for efficiency and quality. The Layout palette allows user to change layout settings while running, and therefore dramatically increase user feedback and experience.

- Force-based algorithms
- Multi-levelalgorithms(graph coarsening)

1.2 Metrics

The statistics and metrics framework offer the most common metrics for social network analysis (SNA) and scale-free networks.

- Betweenness, Closeness, Diameter, Clustering Coefficient, Average shortest path, PageRank, HITS
- Community detection (Modularity)
- Random generators

1.3 Create cartography

Use ranking or partition data to make meaningful the network representation. Customize colors, size or labels to bring sense to the network representation.

1.4 Clustering and Hierarchical Graph

Explore multi-level graphs with Gephi by facilitating exploration and edition of large, hierarchically structured graphs, e.g. , social communities, biochemical pathways or network traffic graphs. Aggregate networks using data attributes or use built-in clustering algorithm.

- Expand and contract subgraphs in Meta node.
- Link and attribute clustering

1.5 Modular

Gephi 0.7 architecture is modular and therefore allows developers to add and extend functionalities with ease. New features like Metrics, Layouts, Filters, Data sources and more can be easily packaged in plugins and shared.

2. Methodology

Gephi offers a set of powerful calculation tools that allow us to further explore the qualities of the Swamy network. One basic and useful calculation is the in-degree and out-degree of the network nodes. In a directed graph (one-way relationships), it has an immigrant-destination model, in-degree refers to the number of incoming edges and out-degree refers to the number of outgoing edges. In classic social network analysis a high in-degree is typicallycalled popularity and a high out-degree is referred to as gregariousness. For our purposes, Swami receives an high number of immigrants could likewise be called popular (as an immigrant destination).

2.1. Modularity

Modularity measures how well a network decomposes into modular communities. A high modularity score indicates sophisticated internal structure. This structure, often called a community structure, describes how the network is compartmentalized into sub-networks. These sub-networks (or communities) have been to have significant real-world meaning. The network modularity score is 0.444 and there are 11 distinct communities in a network with 243 nodes. A community can be easily grouped into a set of nodes with dense connections between them. The modularity invokes community-detection algorithm.

One of the most widely used methods for community detection is modularity maximization. Modularity is a benefit function that measures the quality of a particular division of a network into communities. The modularity maximization method detects communities by
searching over possible divisions of a network for one or more that have particularly high modularity.

2.2. Connected Component

The Connected Component feature calculates the number of strongly and weakly connected components in a network. The Swamy network has 243 strongly connected components and 4 weakly connected components. It detects that the network has many more strongly connected components than weakly connected components. The connected components invokes depth-first search and linear graph algorithm.

2.3. Graph Density

Graph Density measures how close the network is to complete. A complete graph has all possible edges and density equal to 1. Thus Graph Density results 0.035 (directed) – reports very closer graph.

2.4. Page Rank

Page Rank is a feature made with social networking. It measures the importance of each node within the network. The metric assigns each node a probability that is the probability of being at the page after many clicks. The standard adjacency matrix is normalized so that the columns of the matrix sum to 1. The Page Rank measures not only by how many other nodes are connected to it, but how many nodes "Those nodes" are connected to.

2.5. Average Degree

Average Degree is the sum of edges of a vertex.

2.6. Average Weighted Degree

Average of sum of weights of the edges of nodes. The graph is designed in such a way that, weight of an edges represents, how many times that edges is traversed between a pair of nodes. Thus, if weight of node is higher, it means it has been visited many times than any other low weight degree node.

2.7. Centrality

Within the scope of graph theory and network analysis, there are various types of measures of the centrality of a vertex within a graph that determine the relative importance of a vertex within the graph (i.e. how influential a person is within a social network, or, in the theory of space syntax, how well-used a road is within an urban network). Many of the centrality concepts were first developed in social network analysis, and many of the terms used to measure centrality reflect their sociological origin.

There are four measures of centrality that are widely used in network analysis:

- Degree Centrality
- Betweenness
- Closeness and
- Eigenvector

2.8. K-Core Algorithm

In social networks analysis one of the major concerns is identification of cohesive subgraphs of actors within a network. Friendship relation publications citation and many other more. Many studies and researches are focused on social network analysis, including in data mining. It is really important to find patterns in behavior of large online social networks, so the firms behind are able to create better mechanism to handle all that information with lower cost.

Online service such as Facebook, have millions of users using their services simultaneously and interacting with others. Even in different services, the behavior of the network is similar. People tend to interact in the same way as they do in real life, in a structure called "small world" where people in a social network can reach any other person with less than seven steps. Such behavior can be studied to prevent disease propagation or to predict how fast an information can flow in society. Several notions were introduced to formally describe cohesive groups: cliques, n-cliques, n-clans, n-clubs, k-plexes, k-cores, lambda sets. For most of them it turns out that they are algorithmically difficult, classified as NP hard. However for cores very efficient algorithm exists.

2.9. K-Core Technique

There are several tasks when handling with social networks.

Link-based object classification – It can classify the object based on its links.

Object type prediction – It can predict the object type based on the objects linked to it.

Link type prediction – Here it want to predict the link type based on the objects linked by it.

Predicting link existence – It predicts link existence between two objects.

Link cardinality estimation – There are several ways to estimate the cardinality of a link, such as counting the number of links of an object, or counting the numbers of smallest paths that pass through an object.

Object reconciliation – The task is to check whether two objects are the same, based on the links and attributes.

Group detection – It is a clustering task. Here it want to know when a group of objects belong to the same group.

Subgraph detection – Subgraph identification is to find characteristic subgraphs in a network. This is a form of search in graphs.

Metadata mining – Metadata is data about data.

This technique tells about the group detection in social networks based on the degree of each node in the network.

2.9.1 K-Core Algorithm

A k-core of a graph G is a maximal connected subgraph of G in which all vertices have degree al least k. Equivalently, it is one of the connected components of the subgraph of G formed by repeatedly deleting all vertices of degree less than k. If a non-empty k-core exists, then, clearly, G has degeneracy al least k, and the degeneracy of G is the largest k for which G has a k-core.

A vertex *u* has coreness *c* if it belongs to a C-core but not to any (c+1)-core.

The concept of a *k-core* was introduced to study the clustering structure of social networks and to describe the evolution of random graphs; it has also been applied in bioinformatics and network visualization.

As Batageli & Zaversnik (2003) describe, it is possible to find a vertex ordering of a finite graph G that optiomizes the coloring number of the ordering, in linear time, by repeatedly removing the vertex of smallest degree.

In more detail, the algorithm proceeds as follows:

- Initialize an output list L.
- Compute a number *dv* for each vertex v in G, the number of neighbors of v that are not already in L. Initially, these numbers are just the degrees of the vertices.
- Initialize an array D such that D[i] contains a list of the vertices v that are not already in L for which dv = i.
- Initialize k to 0.
- Repeat n times:
- ✓ Scan the array cells D[0], D[1],… until finding an I nonempty.
- ✓ Set k to max(k,i)
- ✓ Select a vertex v from D[i]. Add dv to the beginning of L and remove it from D[i].
- ✓ For each neighbor w of v, subtract one from dw and move w to the cell of D corresponding to the new value of dw.

At the end of the algorithm, k contains the degeneracy of G and L contains a list of vertices in an optimal ordering for the coloring number. The i-cores of G are the prefixes of L consisting of the vertices added to L after k first takes a value greater than or equal to i.

Initializing the variables L, dv, D, and k can easily be done in linear time. Finding each successively removed vertex v and adjusting the cells of D containing the neighbors of v take time proportional to the value of dv at that step; but the sum of these values is the number of edges of the graph (each edge contributes to the term in the sum for the later of its two endpoints) so the total time is linear.

3. Conclusion

Online social networking systems have become popular because they allow users to share content, such as videos and photos, and expand their social circle, by making new friendships. This paper introduced a framework to provide friend recommendations in OSNs. It defines a new node similarity measure that exploits local and global characteristics of a network. It shows that a significant accuracy improvement can be gained by using information about both positive and negative edges.

References

[1] Barab_asi, A.:Linked:How Everything Is Connected to EverythingElse and What It Means. Plume (Apr 2003)

[2] Bastian, M., Heymann, S., Jacomy, M.: Gephi: An open source software for exploring and manipulating networks. In: proceedings of the International AAAI Conference on Weblogs and Social Media (2009), http://WWW.aaai.org/ocs/index. php/ICWSM/09/paper/view/154

[3] Bizer, C., Heath, T., Berners-Lee, T., Heath, T., Hepp, M., Bizer, C.: Linked data- the story so far. International Journal on Semantic Web and Information Systems (IJSWIS) (2009)

[4] Crosby, A.W.: The Measure of Reality: Quanti cation in Western Europe, 1250-1600. Cambridge University Press (Dec 1997)

[5] Freeman, L.C.: Visualizing social networks. Journal of Social Structure 1(1), [np] (2000)

[6] Clark, T.N. (1968). The concept of power. In T.Clark (Ed.), Community structure and decision making: Comparative analysis. San Francisco: Chandler Publishing.

[7] Cohen, J. (1988). Statistical power analysis for the behavioral science (2nd ed.). Hillsdaie, NJ: Lawrence Eribaum Constant, D., Kieseler, S., & Sproull, L. (1994). What's mine is ours, or is it? A study of attitudes about information sharing. Information Systems Research, 5(4), 400-421

[8] Darr, E. D., & Kurtzberg, T.R. (2000). An investigation of partner similarity dimensions on knowledge transfer. Organizational Behavior and Human Decision Processes, 82(1), 28-44.

[9] Fredman, M. L. and Tarjan, R. E. (1987). Fibonacci heaps and their uses in improved network optimization algorithms. Journal of the Association for Computing Machinery, 34(3): 596{615.

[10] Freeman, L. C. (1977). A set of measures of centrality based on betweenness Sociometry, 40:35{41.

[11] Freeman, L. C. (1979). Centrality in social networks: Conceptual clari_cation.

[12] Freeman, L. C. (1980). The gatekeeper, pair-dependency and structural centrality. Quality and Quantity 14:585{592;