Data Warehousing
資料倉儲

Social Network Analysis and Link Mining

Min-Yuh Day
戴敏育
Assistant Professor
專任助理教授

Dept. of Information Management, Tamkang University
淡江大學 資訊管理學系

http://mail.tku.edu.tw/myday/
2011-12-13
<table>
<thead>
<tr>
<th>Week</th>
<th>Date</th>
<th>Content</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>100/09/06</td>
<td>Introduction to Data Warehousing</td>
</tr>
<tr>
<td>2</td>
<td>100/09/13</td>
<td>Data Warehousing, Data Mining, and Business Intelligence</td>
</tr>
<tr>
<td>3</td>
<td>100/09/20</td>
<td>Data Preprocessing: Integration and the ETL process</td>
</tr>
<tr>
<td>4</td>
<td>100/09/27</td>
<td>Data Warehouse and OLAP Technology</td>
</tr>
<tr>
<td>5</td>
<td>100/10/04</td>
<td>Data Warehouse and OLAP Technology</td>
</tr>
<tr>
<td>6</td>
<td>100/10/11</td>
<td>Data Cube Computation and Data Generation</td>
</tr>
<tr>
<td>7</td>
<td>100/10/18</td>
<td>Data Cube Computation and Data Generation</td>
</tr>
<tr>
<td>8</td>
<td>100/10/25</td>
<td>Project Proposal</td>
</tr>
<tr>
<td>9</td>
<td>100/11/01</td>
<td>Midterm Exam Week</td>
</tr>
<tr>
<td>週次</td>
<td>日期</td>
<td>內容（Subject/Topics）</td>
</tr>
<tr>
<td>------</td>
<td>----------</td>
<td>-----------------------------------------------</td>
</tr>
<tr>
<td>10</td>
<td>100/11/08</td>
<td>Association Analysis</td>
</tr>
<tr>
<td>11</td>
<td>100/11/15</td>
<td>Association Analysis</td>
</tr>
<tr>
<td>12</td>
<td>100/11/22</td>
<td>Classification and Prediction</td>
</tr>
<tr>
<td>13</td>
<td>100/11/29</td>
<td>Classification and Prediction</td>
</tr>
<tr>
<td>14</td>
<td>100/12/06</td>
<td>Cluster Analysis</td>
</tr>
<tr>
<td>15</td>
<td>100/12/13</td>
<td>Social Network Analysis and Link Mining</td>
</tr>
<tr>
<td>16</td>
<td>100/12/20</td>
<td>Text Mining and Web Mining</td>
</tr>
<tr>
<td>17</td>
<td>100/12/27</td>
<td>Project Presentation</td>
</tr>
<tr>
<td>18</td>
<td>101/01/03</td>
<td>期末考試週</td>
</tr>
</tbody>
</table>
Outline

• Social Network Analysis
• Link Mining
Social Network Analysis

• A **social network** is a social structure of people, related (directly or indirectly) to each other through a common relation or interest

• **Social network analysis (SNA)** is the study of social networks to understand their structure and behavior

Source: (c) Jaideep Srivastava, srivasta@cs.umn.edu, Data Mining for Social Network Analysis
Social Network Analysis

• Using Social Network Analysis, you can get answers to questions like:
  – How highly connected is an entity within a network?
  – What is an entity's overall importance in a network?
  – How central is an entity within a network?
  – How does information flow within a network?

Source: http://www.fmsasg.com/SocialNetworkAnalysis/
Social Network Analysis: Degree Centrality

Alice has the highest degree centrality, which means that she is quite active in the network. However, she is not necessarily the most powerful person because she is only directly connected within one degree to people in her clique—she has to go through Rafael to get to other cliques.

Social Network Analysis: Degree Centrality

• Degree centrality is simply the number of direct relationships that an entity has.
• An entity with high degree centrality:
  – Is generally an active player in the network.
  – Is often a connector or hub in the network.
  – Is not necessarily the most connected entity in the network (an entity may have a large number of relationships, the majority of which point to low-level entities).
  – May be in an advantaged position in the network.
  – May have alternative avenues to satisfy organizational needs, and consequently may be less dependent on other individuals.
  – Can often be identified as third parties or deal makers.

Source: http://www.fmsasg.com/SocialNetworkAnalysis/
Rafael has the highest betweenness because he is between Alice and Aldo, who are between other entities. Alice and Aldo have a slightly lower betweenness because they are essentially only between their own cliques. Therefore, although Alice has a higher degree centrality, Rafael has more importance in the network in certain respects.
Social Network Analysis: Betweenness Centrality

• Betweenness centrality identifies an entity's position within a network in terms of its ability to make connections to other pairs or groups in a network.

• An entity with a high betweenness centrality generally:
  – Holds a favored or powerful position in the network.
  – Represents a single point of failure—take the single betweenness spanner out of a network and you sever ties between cliques.
  – Has a greater amount of influence over what happens in a network.

Source: http://www.fmsasg.com/SocialNetworkAnalysis/
Social Network Analysis: Closeness Centrality

Rafael has the highest closeness centrality because he can reach more entities through shorter paths. As such, Rafael's placement allows him to connect to entities in his own clique, and to entities that span cliques.
Social Network Analysis: Closeness Centrality

- Closeness centrality measures how quickly an entity can access more entities in a network.
- An entity with a high closeness centrality generally:
  - Has quick access to other entities in a network.
  - Has a short path to other entities.
  - Is close to other entities.
  - Has high visibility as to what is happening in the network.

Source: http://www.fmsasg.com/SocialNetworkAnalysis/
Alice and Rafael are closer to other highly close entities in the network. Bob and Frederica are also highly close, but to a lesser value.

Source: http://www.fmsasg.com/SocialNetworkAnalysis/
Social Network Analysis: Eigenvalue

• Eigenvalue measures how close an entity is to other highly close entities within a network. In other words, Eigenvalue identifies the most central entities in terms of the global or overall makeup of the network.

• A high Eigenvalue generally:
  – Indicates an actor that is more central to the main pattern of distances among all entities.
  – Is a reasonable measure of one aspect of centrality in terms of positional advantage.

Source: http://www.fmsasg.com/SocialNetworkAnalysis/
Hubs are entities that point to a relatively large number of authorities. They are essentially the mutually reinforcing analogues to authorities. Authorities point to high hubs. Hubs point to high authorities. You cannot have one without the other.

Source: http://www.fmsasg.com/SocialNetworkAnalysis/
Social Network Analysis: Hub and Authority

- Entities that many other entities point to are called Authorities. In Sentinel Visualizer, relationships are directional—they point from one entity to another.
- If an entity has a high number of relationships pointing to it, it has a high authority value, and generally:
  - Is a knowledge or organizational authority within a domain.
  - Acts as definitive source of information.

Source: http://www.fmsasg.com/SocialNetworkAnalysis/
### Social Network Analysis

![Network Metrics Table]

<table>
<thead>
<tr>
<th>Name</th>
<th>Type</th>
<th>Degree</th>
<th>Betweenness</th>
<th>Closeness</th>
<th>Eigenvalue</th>
<th>Hub</th>
<th>Authority</th>
</tr>
</thead>
<tbody>
<tr>
<td>Osama bin Laden</td>
<td>Person</td>
<td>44</td>
<td>0.920492092358</td>
<td>1</td>
<td>0.0271</td>
<td>0</td>
<td>0.011</td>
</tr>
<tr>
<td>Abdallah Al-Halabi</td>
<td>Person</td>
<td>2</td>
<td>0</td>
<td>0.654867256637</td>
<td>0.0001</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Abu Mussab al-Zarqawi</td>
<td>Person</td>
<td>84</td>
<td>0.934887847326</td>
<td>0.7028</td>
<td>0.6572</td>
<td>0.1076</td>
<td>0.0166</td>
</tr>
<tr>
<td>Al Qaeda</td>
<td>Terrorist Organiz.</td>
<td>85</td>
<td>1</td>
<td>0.962427745664</td>
<td>0.0416</td>
<td>0.3341</td>
<td>0.0156</td>
</tr>
<tr>
<td>Ayman Al-Zawahiri</td>
<td>Person</td>
<td>14</td>
<td>0.045794908783</td>
<td>0.716129032258</td>
<td>0</td>
<td>0</td>
<td>0.0173</td>
</tr>
<tr>
<td>Enaam Arnaout</td>
<td>Person</td>
<td>4</td>
<td>0.031189325814</td>
<td>0.656804733727</td>
<td>0.0001</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Imad Eddin Barakat Yarbas</td>
<td>Person</td>
<td>11</td>
<td>0.065049589038</td>
<td>0.704016913319</td>
<td>0.0015</td>
<td>0</td>
<td>0.0025</td>
</tr>
<tr>
<td>Khalid Shaikh Mohammed</td>
<td>Person</td>
<td>32</td>
<td>0.339916464724</td>
<td>0.666059817945</td>
<td>0.002</td>
<td>0</td>
<td>0.1528</td>
</tr>
<tr>
<td>Mohamed Atta</td>
<td>Person</td>
<td>61</td>
<td>0.666268740074</td>
<td>0.820197044334</td>
<td>0.0015</td>
<td>0</td>
<td>0.6316</td>
</tr>
</tbody>
</table>

Social Network Analysis

Source: http://www.fmsasg.com/SocialNetworkAnalysis/
Link Mining

Link Mining
(Getoor & Diehl, 2005)

• Link Mining
  – Data Mining techniques that take into account the links between objects and entities while building predictive or descriptive models.

• Link based object ranking, Group Detection, Entity Resolution, Link Prediction

• Application:
  – Hyperlink Mining
  – Relational Learning
  – Inductive Logic Programming
  – Graph Mining

Source: (c) Jaideep Srivastava, srivasta@cs.umn.edu, Data Mining for Social Network Analysis
Characteristics of Collaboration Networks
(Newman, 2001; 2003; 3004)

• Degree distribution follows a power-law
• Average separation decreases in time.
• Clustering coefficient decays with time
• Relative size of the largest cluster increases
• Average degree increases
• Node selection is governed by preferential attachment

Source: (c) Jaideep Srivastava, srivasta@cs.umn.edu, Data Mining for Social Network Analysis
Social Network Techniques

• Social network extraction/construction
• Link prediction
• Approximating large social networks
• Identifying prominent/trusted/expert actors in social networks
• Search in social networks
• Discovering communities in social network
• Knowledge discovery from social network

Source: (c) Jaideep Srivastava, srivasta@cs.umn.edu, Data Mining for Social Network Analysis
Social Network Extraction

• Mining a social network from data sources
• Three sources of social network (Hope et al., 2006)
  – Content available on web pages
    • E.g., user homepages, message threads
  – User interaction logs
    • E.g., email and messenger chat logs
  – Social interaction information provided by users
    • E.g., social network service websites (Facebook)

Source: (c) Jaideep Srivastava, srivasta@cs.umn.edu, Data Mining for Social Network Analysis
Social Network Extraction

- IR based extraction from web documents
  - Construct an “actor-by-term” matrix
  - The terms associated with an actor come from web pages/documents created by or associated with that actor
  - IR techniques (TF-IDF, LSI, cosine matching, intuitive heuristic measures) are used to quantify similarity between two actors’ term vectors
  - The similarity scores are the edge label in the network
    - Thresholds on the similarity measure can be used in order to work with binary or categorical edge labels
    - Include edges between an actor and its k-nearest neighbors

- Co-occurrence based extraction from web documents
Link Prediction

• Link Prediction using supervised learning (Hasan et al., 2006)
  – Citation Network (BIOBASE, DBLP)
  – Use machine learning algorithms to predict future co-authorship
    • Decision three, k-NN, multilayer perceptron, SVM, RBF network
  – Identify a group of features that are most helpful in prediction
  – Best Predictor Features
    • Keywork Match count, Sum of neighbors, Sum of Papers, Shortest distance
Identifying Prominent Actors in a Social Network

• Compute scores/ranking over the set (or a subset) of actors in the social network which indicate degree of importance / expertise / influence
  – E.g., Pagerank, HITS, centrality measures
• Various algorithms from the link analysis domain
  – PageRank and its many variants
  – HITS algorithm for determining authoritative sources
• Centrality measures exist in the social science domain for measuring importance of actors in a social network

Source: (c) Jaideep Srivastava, srivasta@cs.umn.edu, Data Mining for Social Network Analysis
Identifying Prominent Actors in a Social Network

• Brandes, 2011
• Prominence $\rightarrow$ high betweenness value
• Betweenness centrality requires computation of number of shortest paths passing through each node
• Compute shortest paths between all pairs of vertices

Source: (c) Jaideep Srivastava, srivasta@cs.umn.edu, Data Mining for Social Network Analysis
Summary

• Social Network Analysis
• Link Mining
References

- Jiawei Han and Micheline Kamber, Data Mining: Concepts and Techniques, Second Edition, 2006, Elsevier
- Michael W. Berry and Jacob Kogan, Text Mining: Applications and Theory, 2010, Wiley
- Guandong Xu, Yanchun Zhang, Lin Li, Web Mining and Social Networking: Techniques and Applications, 2011, Springer
- Matthew A. Russell, Mining the Social Web: Analyzing Data from Facebook, Twitter, LinkedIn, and Other Social Media Sites, 2011, O'Reilly Media
- Jaideep Srivastava, Nishith Pathak, Sandeep Mane, and Muhammad A. Ahmad, Data Mining for Social Network Analysis, Tutorial at IEEE ICDM 2006, Hong Kong, 2006