CMSC5733 Social Computing

Tutorial 3: Introduction to Project

Yuanyuan, Man
The Chinese University of Hong Kong
sophiaqhsw@gmail.com
Tutorial Overview

• Introduction to CLANS
  – System overview
  – Data acquisition
  – Data preprocessing
  – Modeling social network
  – Data management
  – Social network analysis
  – Visualization

• Proposed projects
Introduction to CLANS

• Objective of the Corporate Leaders Analytics and Network System (CLANS)
  – Identify and analyze social networks among corporations and business elites in China
Introduction to CLANS

• Why we identify business social network in China?
  – Social networks are essential for business in China, especially, relationship plays a crucial role in Chinese business model
  – Related researches indicate that social networks among US firms benefit the debt financing, firm performance and corporate governance.
Introduction to CLANS

• Who can benefit from the analysis of Chinese social network?
  – Investors
    • They can make investment decision according to the social connecting issues among Chinese firms.
  – Common businessman
    • They can do better or potential commercial activities by learning more about specific information for Chinese companies and senior executives and their social networks.
  – Researchers
    • They can do deeper research in this area.
CLANS System Overview
Data Acquisition

• CSMAR DB
  – A list of senior executives and directors of all Chinese listed companies between 1999 and 2011
  – Detailed information of Chinese listed companies

• Baidu Baike Data

• Hexun Renwu Data
Data Preprocessing

• Data Cleaning
• Information Extraction
• Name Disambiguation
• Data Integration
Data Cleaning

• Data quality problems
  – There are set of problems about how to purify, organize and condense the raw data so that be able to implement further high-level operations on them
  – After solving these problems, data should be cleaner, less error and more consistent
Data Cleaning

- Data quality problems types
  - Single sources
    - Text files, webs, databases
    - Misspelling, typos, redundant duplications and inconsistencies
  - Multiple sources
    - In data warehouses or global web-based information systems
    - Different representations among them
Data Cleaning

- **Duplicate Cleaning**
  - Reason: these data were collected every year, may include the same person every year.
Data Cleaning

- **Duplicate Cleaning**
  - Reason: these data were collected every year, may include the same person every year

<table>
<thead>
<tr>
<th>证券代码</th>
<th>统计截止日期</th>
<th>姓名</th>
<th>职务类别</th>
<th>具体职务</th>
<th>性别</th>
<th>年龄</th>
<th>教育背景</th>
<th>职称</th>
</tr>
</thead>
<tbody>
<tr>
<td>620</td>
<td>2011-12-31</td>
<td>龚鹏</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>620</td>
<td>2010-12-31</td>
<td>龚鹏</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>620</td>
<td>2009-12-31</td>
<td>龚鹏</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>620</td>
<td>2008-12-31</td>
<td>龚鹏</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>620</td>
<td>2007-12-31</td>
<td>龚鹏</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>620</td>
<td>2006-12-31</td>
<td>龚鹏</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>620</td>
<td>2005-12-31</td>
<td>龚鹏</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>526</td>
<td>2009-12-31</td>
<td>龚高</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>526</td>
<td>2008-12-31</td>
<td>龚高</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>526</td>
<td>2007-12-31</td>
<td>龚高</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>526</td>
<td>2006-12-31</td>
<td>龚高</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>526</td>
<td>2005-12-31</td>
<td>龚高</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>526</td>
<td>2004-12-31</td>
<td>龚淑媛</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>659</td>
<td>2003-12-31</td>
<td>龚淑媛</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Data Cleaning

• Duplicate Cleaning
  – Method
    • Rule-based approach to detect duplicate records based on the characteristic of our data
    • stock id, <collection date, age>, gender
Data Cleaning

• Duplicate Cleaning
  – Result
    • Find 84000 entries
    • 9 pairs of entries which have same name and same stock id, relatively very few for all 84000 entries
    • 60886 names which are owned by only one entry
    • 8773 names which are owned by more than one entries
Data Cleaning

- Data Correction
  - Missing data

<table>
<thead>
<tr>
<th>stock_id</th>
<th>id</th>
<th>finish_year</th>
<th>name</th>
<th>position_classi</th>
<th>position</th>
<th>gender</th>
<th>age</th>
<th>education</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>30</td>
<td>2010-12-31</td>
<td>男宝瑞</td>
<td>00390000N0</td>
<td>副行长(兼任)</td>
<td>男</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>34</td>
<td>2009-12-31</td>
<td>男宝瑞</td>
<td>10300000N0</td>
<td>“董事,副行长”</td>
<td>男</td>
<td>52</td>
<td>0</td>
</tr>
</tbody>
</table>

  - Contradictory data

<table>
<thead>
<tr>
<th>stock_id</th>
<th>id</th>
<th>finish_year</th>
<th>name</th>
<th>position_classi</th>
<th>position</th>
<th>gender</th>
<th>age</th>
<th>education</th>
</tr>
</thead>
<tbody>
<tr>
<td>11</td>
<td>561</td>
<td>2011-12-31</td>
<td>曹子扬</td>
<td>2300000000</td>
<td>监事会主席(兼任)</td>
<td>男</td>
<td>60</td>
<td>0</td>
</tr>
<tr>
<td>11</td>
<td>588</td>
<td>2000-12-31</td>
<td>曹子扬</td>
<td>1000000000</td>
<td>董事</td>
<td>男</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>11</td>
<td>563</td>
<td>2010-12-31</td>
<td>曹子扬</td>
<td>2100000000</td>
<td>监事会主席</td>
<td>男</td>
<td>60</td>
<td>2</td>
</tr>
</tbody>
</table>

  - Spelling error

<table>
<thead>
<tr>
<th>stock_id</th>
<th>id</th>
<th>finish_year</th>
<th>name</th>
<th>position_classi</th>
<th>position</th>
<th>gender</th>
<th>age</th>
<th>education</th>
</tr>
</thead>
<tbody>
<tr>
<td>998</td>
<td>10/03</td>
<td>2010-12-31</td>
<td>Emmanuel KOUKPER</td>
<td>1000000000</td>
<td>童争</td>
<td>男</td>
<td>30</td>
<td>3</td>
</tr>
<tr>
<td>2032</td>
<td>11876</td>
<td>2010-12-31</td>
<td>Francois LECLEIRE</td>
<td>2300000000</td>
<td>监事会主席(兼任)</td>
<td>男</td>
<td>32</td>
<td>0</td>
</tr>
<tr>
<td>2032</td>
<td>11880</td>
<td>2009-12-31</td>
<td>Francois LECLEIRE</td>
<td>2100000000</td>
<td>监事会主席</td>
<td>男</td>
<td>31</td>
<td>4</td>
</tr>
</tbody>
</table>
Data Cleaning

- Data Correction
  - Result

<table>
<thead>
<tr>
<th>Type</th>
<th>Error</th>
<th>Solution</th>
</tr>
</thead>
<tbody>
<tr>
<td>missing data</td>
<td>859 entries with age=0</td>
<td>have been marked or replaced with correct age from deleted duplicate entries</td>
</tr>
<tr>
<td>missing data</td>
<td>18221 entries don not have introduction</td>
<td>need data acquisition</td>
</tr>
<tr>
<td>contradictory data</td>
<td>2354 couples of entries with different age</td>
<td>the smaller group have been removed</td>
</tr>
<tr>
<td>spelling error</td>
<td>593 couple of entries with different name</td>
<td>checked manually and correct the errors if string-distance less than a threshold</td>
</tr>
</tbody>
</table>
Information Extraction

• Rule-learning method
  – According to the characteristic of the data, we find several pattern of expression
    • XXXXyear-XXXX year, XXperson in XXXXcompany, XXXX department, as XXXXposition
    • XXXXyear-XXXX year, XXperson in XXXXcompany, as XXXXposition and XXXXposition
    • XXXXyear-XXXX year, XXperson in XXXXcompany, XXXXcompany, as XXXXposition
Information Extraction

• Rule-learning method
  – Result
    • Precision rate is low
    • Because of expression’s diversity and complexity
    • Manually check and revise the result
Information Extraction

• HMM model
  – A general statistical modeling technique for ‘linear’ problems like sequences or time series
  – Widely used in NLP and speech recognition applications
  – Application:
    • In our project, given a segmented text files
    • Extract work information, education information, general information (age, gender, and etc.)
Information Extraction

• HMM model
  – Statistics:
    • Extract 50000 text files as learning samples
    • Extract 6000 text files as testing samples
  – Predefined tags

• 888: 中国平安，深圳发展银行...
• 999: 董事长，总经理...
• 777: 化学车间，经济学院，生产中心...
• 666: 辞去, 免去, 辞职, 退休, 解任...
• ...
Information Extraction

HMM model

- Words segmentation with word tags on learning samples
- Use company, position, department’s dictionary to replace the word tags with predefined tags
Information Extraction

- 2006年7月到2007年1月兼任中国平安副首席保险业务执行官

HMM model
- Words segmentation with word tags on testing samples, then remove the tags
Information Extraction

• HMM model
  – Calculate two matrixes: 1. Mat(p(x|x)) – state transition probabilities; 2. Mat(p(y|x)) – output probabilities
  – Based on the two matrixes, use Viterbi algorithm to determine the word tags of testing samples
  – Verify the tagging result, and store specific words of certain type of tag into database
Information Extraction

- HMM model
Modeling Social Network

• Individual social network
  – Alumni social network
  – Colleague social network

• Corporation social network
Alumni Social Network

- We define alumni relationship as the closeness of the relationship between two alumni
  - Four criteria: major, degree, time of enrollment, intersection school time
  - Deduce 13 types of relationships
    - The closest relationship means that two people are classmates (same major, same degree and same time of enrollment), weight is 0.9
    - The weight of farthest relationship is 0.1 (with different major, different degree and no intersection school time)
Colleague Social Network

• Let position rank (PS) denoted as a representation of job level by integer ranging from 0 to 9.
  – The higher position rank has a larger value
  – The PS of the board chairman is 9
  – The PS of the CEO is 8
  – The PS of the independent director is 1.
Colleague Social Network

• Let value relation between two colleagues denoted as the average position rank of the two people.

• Let close relation between two colleagues denoted as the intersection years that they work together.
Colleague Social Network

• Let colleague relationship denoted as a combination of value relation and close relation.

• The colleague weight between person $p_i$ and $p_j$ is defined as

$$
\omega_{p_i,p_j} = \sum_{t \in L(p_i,p_j)} \frac{PS_{t,p_i} + PS_{t,p_j}}{2},
$$

where $L(p_i,p_j)$ denotes a collection of the intersection years that person $p_i$ and $p_j$ used to work with each other, and $PS_{t,p_i}$ denotes the position rank of person $p_i$ in the year $t$. At the end, all the weights are normalized, which is also applied in the following weight calculation.
Individual Social Network

• We define the individual social network as an undirected graph $G(V,E)$.

• In $G(V, E)$, every edge (relationship) has weighted value, which is defined as

$$W_{i,j} = \alpha \omega_{i,j}^{al} + \beta \omega_{i,j}^{co}$$

$\omega_{i,j}^{al}$ is a weight for alumni relationship, $\omega_{i,j}^{co}$ for colleague relationship; $\alpha$ and $\beta$ denotes the corresponding percentage

• Will add family, friends, corporation social network to the whole individual social network
Definition 3 We define the corporation social network as an directed graph \( \hat{G}(\hat{V}, \hat{E}) \). In \( \hat{G}(\hat{V}, \hat{E}) \), every vertex (corporation) has feature set \( P_i = \{p_{i,1}, p_{i,2}, \ldots, p_{i,n}\} \) and every direct edge (relationship) has weighted value \( W_{i,j} = (\omega_{i,j}^{gp}, \omega_{i,j}^{nk}) \). \( n \) is the size of the set (total number of staffs); \( \omega_{i,j}^{gp} \) is a weight for group membership, \( \omega_{i,j}^{nk} \) for network relationship.
Corporation Social Network

Person a and Person b once worked at Company C.
So Person a and Person b have a colleague relation.

Person a and Person b contributes to link Company A and Company B.

Person c works in Company A and B at the same time.
So c contributes to link A and B.
\( \omega_{i,j}^{gp}, \omega_{i,j}^{nk} \) are defined as follows:

\[
\omega_{i,j}^{gp} = \sum_{p_i^k \in P_i \cap P_j} PS_{p_i^k} \cdot \omega_{p_i^k}^{gp}
\]

\[
\omega_{i,j}^{nk} = \sum_{(p_i^k, p_j^r) \in L_2(P_i, P_j)} PS_{p_i^k} \cdot \omega_{p_i^k, p_j^r}^{nk}
\]

\( PS_{p_i^k} \) denotes the position rank of person \( p_i^k \) in corporation \( i \); \( \omega_{p_i^k}^{gp} \) is a weight for \( p_i^k \) connecting \( P_i \) with \( P_j \); \( L_2(P_i, P_j) \) denotes a collection of connections between \( (P_i - P_i \cap P_j) \) and \( (P_j - P_i \cap P_j) \); \( \omega_{p_i^k, p_j^r}^{nk} \) denotes a weight between \( p_i^k \) and \( p_j^r \) calculated in the previous equation.
Corporation Social Network

Thus, the corporation weight from corporation $i$ to $j$ is defined as $W_{i,j} = \alpha \omega_{i,j}^{gp} + \beta \omega_{i,j}^{nk}$, where $\alpha$ and $\beta$ denotes the corresponding percentage that the two relations contribute to the corporation social network respectively.
Corporation Social Network
Data Management

• Define individual and corporation scheme
Data Management

- Use XML files to store individual and corporation entities
- Form a latest updated data
- Easily access
Data Management

• Why use XML?
  – Extensibility
    • Easy to add new features or modify selected fields
    • Like <birthplace>
      <gender src="CSMAR info" update="128900000"> Male </gender>
      <birthday src="CSMAR info" update="128900000"> 1981 − 06 − 18 </birthday>
  – Traceability
    • The src attribute indicates where the text value comes from
  – Distinguishability
    • Easily handle various properties with the same tag
      <name desc="Chinese" src="CSMAR info" update="128900000"> Tongming Wang </name>
      <name desc="English" src="Baidu info" update="134565800"> Tom Wang </name>
  – Version Control
    • Error positioning, difference checking and data recovering
Data Mining

• Link Analysis
  – Aim to find important individuals and corporations
  – For individual social network
    • A method takes into consideration of both personal and network information
    • The basic idea is that an important person knows someone a, then a is also important.
    • Steps: firstly, we assign every individual with initial score according to position rank; secondly, we distribute the score according to the weight of the out-link edge; third, the algorithm will stop if the change is less than a threshold.
  – For corporation social network
    • Use PageRank algorithm
Data Mining

- Relation Mining
  - Aim to find out important people’s link between two corporations’ link

![Image of data table]
公司详细关系(深圳发展银行股份有限公司 and 交通银行股份有限公司)

type(e:两人是教育关系, w:两人是工作关系, t:该人在两家公司就职过)

relation_rank(5代表最紧密关系，1代表最不紧密关系，0表示不改变当前值)

<table>
<thead>
<tr>
<th>pid1</th>
<th>name1</th>
<th>position1</th>
<th>posi_rank1</th>
<th>pid2</th>
<th>name2</th>
<th>position2</th>
<th>posi_rank2</th>
<th>type</th>
<th>weight</th>
<th>relation_rank/set new value</th>
<th>visited or not</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>肖道宁</td>
<td>执行董事,董事长</td>
<td>1</td>
<td>0</td>
<td>N/A</td>
<td>N/A</td>
<td>0</td>
<td>t</td>
<td>1.0</td>
<td>5 / 0</td>
<td>visited</td>
</tr>
<tr>
<td>3</td>
<td>王利平</td>
<td>非执行董事</td>
<td>9</td>
<td>341643</td>
<td>王冬胜</td>
<td>非执行董事</td>
<td>9</td>
<td>w</td>
<td>0.12269592</td>
<td>3 / 0</td>
<td>visited</td>
</tr>
<tr>
<td>4</td>
<td>姚波</td>
<td>非执行董事</td>
<td>9</td>
<td>341643</td>
<td>王冬胜</td>
<td>非执行董事</td>
<td>9</td>
<td>w</td>
<td>0.12157635</td>
<td>3 / 0</td>
<td>visited</td>
</tr>
<tr>
<td>5</td>
<td>顾敏</td>
<td>非执行董事</td>
<td>9</td>
<td>341643</td>
<td>王冬胜</td>
<td>非执行董事</td>
<td>9</td>
<td>w</td>
<td>0.1047828</td>
<td>3 / 0</td>
<td>visited</td>
</tr>
<tr>
<td>6</td>
<td>叶素兰</td>
<td>非执行董事</td>
<td>9</td>
<td>341643</td>
<td>王冬胜</td>
<td>非执行董事</td>
<td>9</td>
<td>w</td>
<td>0.11373936</td>
<td>3 / 0</td>
<td>visited</td>
</tr>
<tr>
<td>8</td>
<td>王开国</td>
<td>非执行董事</td>
<td>9</td>
<td>341843</td>
<td>郭宇</td>
<td>监事</td>
<td>0</td>
<td>w</td>
<td>0.13165249</td>
<td>3 / 0</td>
<td>visited</td>
</tr>
<tr>
<td>15</td>
<td>储一的</td>
<td>独立董事</td>
<td>9</td>
<td>341855</td>
<td>朱鹤新</td>
<td>业务总监</td>
<td>4</td>
<td>e</td>
<td>0.1</td>
<td>2 / 0</td>
<td>visited</td>
</tr>
<tr>
<td>20</td>
<td>罗康平</td>
<td>外部监事</td>
<td>0</td>
<td>341643</td>
<td>王冬胜</td>
<td>非执行董事</td>
<td>9</td>
<td>w</td>
<td>0.00793538</td>
<td>1 / 0</td>
<td>visited</td>
</tr>
<tr>
<td>20</td>
<td>罗康平</td>
<td>外部监事</td>
<td>0</td>
<td>341803</td>
<td>冯婉霞</td>
<td>非执行董事</td>
<td>9</td>
<td>w</td>
<td>0.00793538</td>
<td>1 / 0</td>
<td>visited</td>
</tr>
<tr>
<td>20</td>
<td>罗康平</td>
<td>外部监事</td>
<td>0</td>
<td>341856</td>
<td>叶迪奇</td>
<td>副行长</td>
<td>3</td>
<td>w</td>
<td>0.00793538</td>
<td>1 / 0</td>
<td>visited</td>
</tr>
</tbody>
</table>
个人信息

pid1：20
姓名：罗康平
当前职位：外部监事
简历：罗康平先生，外部监事。1954年出生，理学（经济学）硕士，现任香港信和集团租赁及特殊项目总经理。2010年12月至今，任深圳发展银行监事会外部监事。罗康平先生曾任中电集团经济师，香港上海汇丰银行总部战略规划执行官，管理会计，财务控制管理信息资源管理、营销经理，区域高管，抵押业务主管，银行服务主管，中国银行（香港）零售银行总经理。

pid2：341830
姓名：冯婉眉
当前职位：非执行董事

工作关系详情：

<table>
<thead>
<tr>
<th>company</th>
<th>start date</th>
<th>end date</th>
<th>weight</th>
<th>right</th>
<th>wrong</th>
</tr>
</thead>
<tbody>
<tr>
<td>香港上海汇丰银行</td>
<td>N/A</td>
<td>N/A</td>
<td>0.00793538</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

确认修改

校友关系详情：
Data Mining

• **Shortest path finding**
  – **People-to-people**
    • Direct connection: schoolmate, family, friend or colleague
    • Indirect connection between them through closest connected intermediate nodes
  – **People-to-company**
    • Direct connection: employment relationship
    • Indirect connection: find out the possible link to the people who worked in the company
  – **company-to-company**
    • Direct connection: cooperative relationship
    • Indirect connection
Data Mining

• Temporal Relation Comparison
  – Compare two people’s timeline
Visualization

• Website overview
Visualization

- First page
肖遂宁

基本信息

姓名：肖遂宁
单位：深圳发展银行
职务：前董事长
生日期：1948年02月
毕业院校：水电部重庆电力学校
学历：大专

简介：
现年50多岁的深发展董事长肖遂宁，曾是交行深圳分行原行长，在交行工作期间，除负责分行全面管理，还曾负责分管分行的公司银行、个人银行、信贷、人力资本、房地产和证券事务，曾经兼任交通银行监事会监事，是股份制银行高管中的实干派。
腾讯公司（腾讯控股有限公司），成立于1998年11月，是目前中国最大的互联网综合服务提供商之一，也是中国服务用户最多的互联网企业之一。成立以来，腾讯一直秉承一切以用户价值为依归的经营理念，始终处于稳健、高速发展的状态。腾讯打造了中国最大的网络社区，满足互联网用户沟通、资讯、娱乐和电子商务等方面的需求。
Proposed Project

- Identify and analyze business social network in Sina Weibo
- Identify Chinese politician social network and analyze their influence on Chinese business social network

Please talk to me if you are interested in these two projects