Corporate Leaders Analytics and Network System (CLANS): Constructing and Mining Social Networks among Corporations and Business Elites in China

Yuanyuan Man*, Shuai Wang*, Yi Li*, Yong Zhang*, Long Cheng*, Lixin Liu*, Tianyu Zhang**, T. J. Wong**, and Irwin King*

*Department of Computer Science and Engineering
and**School of Accountancy
The Chinese University of Hong Kong
Shatin, N.T., Hong Kong
{yyman,wangs}@cse.cuhk.edu.hk

Abstract. Social network plays a vital role in Chinese business and is highly valued by business people. However, social network analysis is difficult due to issues in data collection, natural language processing, social network detection and construction, relationship mining, etc. Thus, we develop the Corporate Leaders Analytics and Network System (CLANS) to tackle some of these problems. Our contributions are in three aspects: 1) we construct a business social network and formulate the similarity relations among individuals and corporations; 2) we utilize XML files with our defined schema to attain extensibility, traceability, distinguishability, and version control for data management; 3) we conduct further data mining to discover more implicit information, including important entities finding, relation mining and shortest path finding. In this paper, we present the overview of CLANS and specifically address these three major issues. We have made an operational system and achieved basic functionalities.

Keywords: social network, business analytics, data mining, China market, business elites, corporations

1 Introduction

Social networks are essential for business in China and many other emerging economies. Especially, relationship plays a crucial role in Chinese business model [1]. Related researches indicate that social networks among US firms benefit the debt financing [8], firm performance [5], and corporate governance [4]. However, few studies focus on corporations and elites in China. Hence, it is important to collect, investigate and analyze these business relations for corporation and elites in China.

Further, the analysis of Chinese social network is significant for business people. Investors take into account and highly value the social connecting issues among Chinese firms for their investment decision. Besides, common businessman would also like to learn more about specific information for Chinese companies, senior executives and their social networks, for better or potential commercial activities.

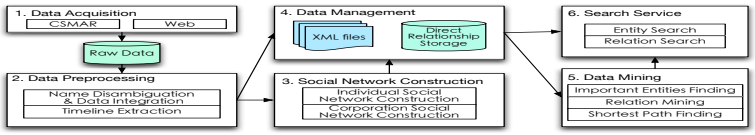


Fig. 1: Architecture of CLANS

Although the analysis of business social networks in China is important, there are a number of difficulties in data collection, natural language processing, network detection and construction, relationship mining, etc. Thus, we design and implement the Corporate Leaders Analytics and Network System (CLANS) to tackle some of these proposed problems, with the help of available computational approaches in social computing [3, 6].

The objective of CLANS is to identify and analyze social networks among corporations and business elites. Specifically, we currently focus on 2,500 Chinese listed firms and their senior managers. In this paper, we introduce the system overview of CLANS and mainly focus on addressing three issues: 1) how to construct and quantify business social network; 2) how to efficiently manage the constantly updated inconsistent data; 3) how to mine social network more implicit information.

We address the problems with our novel approaches: 1) we construct a business social network and formulate similarity relations among individuals and corporations; 2) we utilize XML files with our defined schema to attain extensibility, traceability, distinguishability and version control for data management; 3) we conduct data mining to discover more implicit information, including important entities finding, relation mining and shortest path finding.

The organization of the paper is as follows. We present the CLANS system in Section 2. Specifically Section 2.3, 2.4, and 2.5 describe more detail in addressing the three major issues. We present our system in website version in Section 3. And Section 4 gives a conclusion.

2 CLANS System

2.1 System Overview

The architecture of CLANS consists of six components, shown in Fig. 1. For Data Acquisition, we collect raw data from China Securities Market and Accounting Research Database (CSMAR) and the web. Then we conduct Data Preprocessing (Section 2.2) and Social Network Construction (Section 2.3) to create entities and relations respectively. Then, all entities are stored in XML files for Data Management (Section 2.4), with an auxiliary database to store relations. Table. 1 gives detail statistics of our dataset. After that, CLANS conduct Data Mining (Section 2.5) and provide Search Services (Section 2.6) with the latest data.

Table 1: Statistics of the Dataset

Dataset	Entities	Relations
Individual	83,929	2,600,000
Corporation	2,551	270,000

2.2 Data Preprocessing

We conduct data preprocessing to create individual entities and extract individual detail structured timeline information.

Name Disambiguation and Data Integration. In this stage we encounter and tackle two major issues. Problem one is that a certain person matches multiple records, and problem two is that a popular name matches multiple people. Our solution is that, if two records share a high similarity of cognizable features (like name, age, gender, and birthplace) over a defined threshold, we consider them as the same person. For problem one, the original name repetition rate in our database is 12.6%, and it is solved by a precision rate of 97.6%. For problem two, about 47% of our target names have this problem, and it is finally solved by a precision rate of 81%.

Timeline Extraction. We analyze personal unstructured profiles and extract structured timeline information, like education and work experience. We adopt different strategies for different parts. For education timeline, we employ rule-learning algorithm with precision rate 95.1%. For working timeline, we combine rule-learning algorithm with HMM model [2], owning to expression's diversity and complexity, and we achieve a precision rate of 83.1%.

2.3 Social Network Construction

We construct a business social network, which contains two parts of individual and corporation, and formulate similarity relations among individuals and corporations.

Individual Social Network Construction. We construct alumni and colleague social network respectively and formulate similarity relations among them, and then integrate them with weighting coefficients to construct the individual social network.

We define the alumni relationship as the closeness of the relationship between two alumni based on the combination of four criteria, including major, degree, time of enrollment, and intersection school time. We deduce 13 types of relationships between two alumni and assign the corresponding weight empirically. For example, the closest relationship (same major, same degree and same time of enrollment) means that the two people are classmates, with a high possibility that they know each other well, so we assign the weight between them to 0.9, while the farthest relationship is 0.1 (with different major, different degree and no intersection school time).

Definition 1 *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 a board chairman and a CEO are assigned to be 9 and 8 respectively, while we assign the independent director to be 1. 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.*

Definition 2 *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*

$$\omega_{p_i,p_j} = \sum_{t \in L(p_i,p_j)} \frac{PS_{t,p_i} + PS_{t,p_j}}{2}, \tag{1}$$

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.

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 $W_{i,j} = \alpha\omega^{al}_{i,j} + \beta\omega^{co}_{i,j}$. $\omega^{al}_{i,j}$ is a weight for alumni relationship, $\omega^{co}_{i,j}$ for colleague relationship; α and β denotes the corresponding percentage that the alumni and colleague relationship contribute to the individual relation respectively. We can construct the specific individual social network according to personalized requirements by specifying different weighting coefficients.

Corporation Social Network Construction. We construct the corporation social network based on individual relations and formulate the similarity relation among corporations.

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 $P_i = \{p^1_i, p^2_i, \cdots, p^n_i\}$ and every direct edge (relationship) has weighted value $W_{i,j} = (\omega^{gp}_{i,j}, \omega^{nk}_{i,j})$. n is the size of the set (total number of staffs); $\omega^{gp}_{i,j}$ is a weight for group membership, $\omega^{nk}_{i,j}$ for network relationship. $\omega^{gp}_{i,j}$, $\omega^{nk}_{i,j}$ are defined as follows:*

$$\omega^{gp}_{i,j} = \sum_{p^k_i \in P_i \cap P_j} PS_{p^k_i} * \omega^{gp}_{p^k_i} \tag{2}$$

$$\omega^{nk}_{i,j} = \sum_{(p^k_i, p^r_j) \in L_2(P_i, P_j)} PS_{p^k_i} * \omega^{nk}_{p^k_i, p^r_j} \tag{3}$$

$PS_{p^k_i}$ denotes the position rank of person p^k_i in corporation i; $\omega^{gp}_{p^k_i}$ is a weight for p^k_i 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 $(P_j - P_i \cap P_j)$; $\omega^{nk}_{p^k_i, p^r_j}$ denotes a weight between p^k_i and p^r_j calculated in the previous equation.

Thus, the corporation weight from corporation i to j is defined as $W_{i,j} = \alpha\omega^{gp}_{i,j} + \beta\omega^{nk}_{i,j}$, where α and β denotes the corresponding percentage that the two relations contribute to the corporation social network respectively.

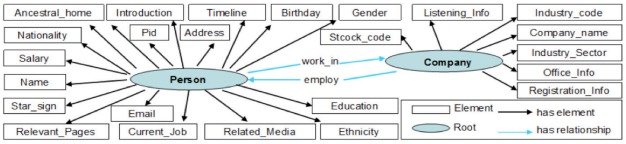


Fig. 2: The person and company entity schema

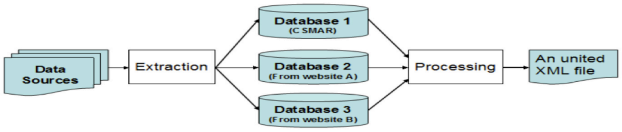


Fig. 3: Form an united XML file from multiple data sources.

2.4 Data Management

We utilize XML files with defined schema to achieve extensibility, traceability, distinguishability and version control, for data management. In CLANS, our defined person and company schema is shown in Fig. 2. For conciseness, some subelements are not displayed.

In data management, our target is to form a united and latest XML file from diverse and expanding databases. Specifically, after extracting and processing data from different sources, we utilize a united representation format (XML file) for the system, so that it can easily access to the latest updated data without complicated queries, as shown in Fig. 3. It is an implementation of MVC (Model-View-Control) model. In this way, the backend handler can keep crawling new data, and data management controller just need to produce latest XML files, and subsequently the front-end can access latest united data, all of them decoupling.

We demonstrate the advantages as well as the reasons for applying our approach, with our specific implementation, by illustrating the following XML sample. The sample is a segment of the XML content. Besides the defined elements in our schema, we also propose meaningful attributes in our XML files.

```
<person pid = "27435">
  <names>
    <name desc="Chinese" src="CSMAR_info" update="128900000"> Tongming Wang </name>
    <name desc="English" src="Baidu_info" update="134565800">Tom Wang</name>
  </names>
  <gender src="CSMAR_info" update="128900000"> Male </gender>
  <birthday src="CSMAR_info" update="128900000"> 1981 − 06 − 18 </birthday>
  ......
</person>
```

Extensibility. It is a lightweight operation to extend the XML content. For routine maintenance, we keep updating our databases and adding new acquired data. Instead of reconstructing or building up a database, we just need to modify the existing XML files, adding some new features or just modifying selected

fields. For instance, if we extract a new feature from Web, e.g., <*birthplace*>, we just need to add one line in a XML file.

Traceability. With defined meaningful attributes, we make the modification of XML files traceable. For example, the *src* attribute indicates where the text value comes from. In the example, the birthday element is from the table *CSMAR_info*, while the English name is from *Baidu_info*. The *update* attribute records the timestamp we update an element. They play important roles in version control.

Distinguishability. We can easily handle various properties with the same tag. Sometimes different elements share a same tag name, since they belong to a same general idea but different specific meaning. For example, an individual might own *Chinese Name* and *English Name*. Thus we define the *desc* (description) attribute to distinguish different types of same tag. This attribute also contributes to the extensibility. If we extract a new kind of name that was not pre-defined in our former schema, like *Nickname*, we just need to add a new line <*name desc="Nickname"*>*Tommy*</*name*>.

Version Control. Combined with version control, our approach achieves error positioning, difference checking and data recovering. With the help of meaningful attributes, we can easily check the data source and latest update time of every element in XML files. Thus, if an error were found, what leads to it and when it happens will be directly discovered. Further, we can find out what the specific false modification is, since the version control provides difference checking between two versions. Moreover, after the error detection, we can handle the emergency immediately. What we need to do is just checking out the updated time for that error, and then turn back to the previous version. Particularly, since our approach is based on updating existing XML files, our data management can be much easily combined with version control. In our implementation, we utilize Subversion (SVN) [1] to establish our own SVN server.

2.5 Data Mining

We conduct data mining to discover implicit information in these three aspects: important entities finding, relation mining, and shortest path finding.

Important Entities Finding. We utilize two algorithms to discover important individual and corporation entities in social network respectively. For individuals finding, our algorithm is refer to the work of [9], which takes into consideration of both personal and network information. The basic idea is that commonly the person with high position level plays an important role in business social network, and if he knows someone with close relation, then that person is also important. There are two steps in the algorithm: 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. For corporations finding, we apply our modified PageRank [7] algorithm, which only take account of the corporations' relationships.

[1] http://subversion.apache.org

Relation Mining. For any specific corporation, relation mining uses a method to find out its important correlated corporations and its staffs who support those links. The corporation relations are defined as a sequence of relationships $\{\hat{e}_{i,1}, \hat{e}_{i,2}, \cdots, \hat{e}_{i,j}\}$, where i and j represents the source corporation and target corporation respectively. A clustering algorithm is utilized to group the relationships by weight, and a pre-defined threshold is used to select the relations in the group. Then we identify its important correlated corporations. Each corporation relation is defined as $\hat{e}_{i,j} = \{\tilde{e}^{nk}_{p^k_i,p^r_j}, \cdots, \tilde{e}^{gp}_{p^d_i}, \cdots\}$, where $\tilde{e}^{nk}_{p^k_i,p^r_j}$ denotes a connection between person p^k_i in corporation i and p^r_j in j, and $\tilde{e}^{gp}_{p^d_i}$ denotes person p^d_i connecting corporation i with j. We use the same method to identify the important staffs that support those corporation links.

Shortest Path Finding. We utilize the state-of-the-art tools to identify the shortest path between three components: people-to-people, people-to-company and company-to-company. If they have direct connection, two people have direct connection like schoolmate, family, friend or colleague, or people and company have the employment relationship, or two corporations have the cooperative relationship, the system returns the direct relation between them. For two people or two corporations, who do not have direct connection with each other, shortest path aims to find out the indirect connection between them through closest connected intermediate nodes. For people and company, shortest path aims to find out the possible link to the people who worked in the company and have a high position level. We use the state-of-the-art tools to compute the shortest path for any input person or corporation within 3 seconds.

2.6 Search Service

In CLANS, we provide two types of services: entity search and relation search.

Entity Search. Given any keyword, system returns a list of ranked persons and companies. Chosen a person/corporation, the system returns related information about the person/corporation.

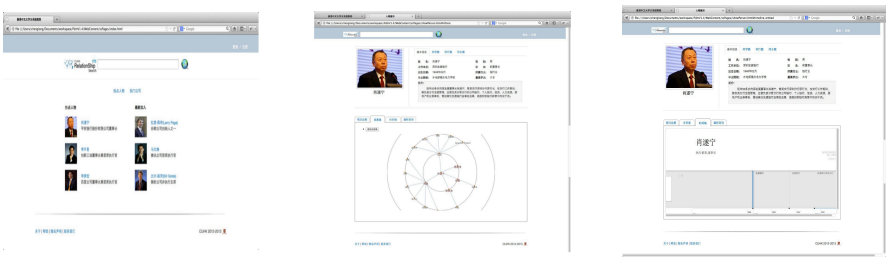
Relation Search. Given any two keywords, the system returns shortest path between them and the corresponding intermediate nodes and link information.

3 Website Illustration

We have been establishing a website to demonstrate CLANS. Though still first version, it now can visualize basic information, temporal relation and timeline for both companies and individuals. Figure. 4a shows the homepage, in which popular corporations and individuals come from the results of Section 2.5. Figure. 4b and 4c show the basic information, temporal relation, and timeline of a senior executive, which are the results of Section 2.2 and Section 2.3.

4 Conclusion

Social network is of great importance for Chinese business model and business people. However, the analysis is difficult due to issues in data collection, natural



(a) Homepage (b) Person page (c) Person page

Fig. 4: Sample pages of websites

language processing, network detection and construction, etc. Thus, we develop CLANS to solve some of these problems. And CLANS aims at constructing and mining social network among corporations and business elites. In this paper, we have described the system overview and specifically addressed three issues with our proposed novel solutions. We have established an operational system and achieved basic functionalities. We create a website to visualize information for both companies and individuals. However, it is just the first version and the development of CLANS with more powerful functions as well as a wider researched scope will be our long-term project.

5 Acknowledgement

This work was supported in part by a grant from the Research Grants Council of the Hong Kong Special Administrative Region, China (Project No. CUHK 413212).

References

1. Franklin Allen, Jun Qian, and Meijun Qian. Law, finance, and economic growth in china. *Journal of financial economics*, 77(1):57–116, 2005.
2. Sean R. Eddy. Profile hidden markov models. *Bioinformatics*, 14(9):755–763, 1998.
3. Irwin King, Jiexing Li, and Kam Tong Chan. A brief survey of computational approaches in social computing. In *Neural Networks, 2009. IJCNN 2009. International Joint Conference on*, pages 1625–1632. IEEE, 2009.
4. David Larcker, Scott Richardson, Andrew Seary, and Ayse Tuna. Back door links between directors and executive compensation. *Back Door Links Between Directors and Executive Compensation (February 2005)*, 2005.
5. David F Larcker, Eric C So, and Charles CY Wang. *Boardroom centrality and stock returns*. Citeseer, 2010.
6. Mingzhen Mo and Irwin King. Exploit of online social networks with community-based graph semi-supervised learning. In *Neural Information Processing. Theory and Algorithms*, pages 669–678. Springer, 2010.
7. Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. The pagerank citation ranking: bringing order to the web. 1999.
8. Brian Uzzi. Embeddedness in the making of financial capital: How social relations and networks benefit firms seeking financing. *American sociological review*, pages 481–505, 1999.
9. Jing Zhang, Jie Tang, and Juanzi Li. Expert finding in a social network. In *Advances in Databases: Concepts, Systems and Applications*, pages 1066–1069. Springer, 2007.