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

Yuanyuan Man[⋆], Shuai Wang[⋆], 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 collect data from multiple sources and do the preprocessing to make it available to use; 2) we construct a business social network and formulate the similarity relations among individuals and corporations; 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 [12], firm performance [6], and corporate governance [5]. 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.

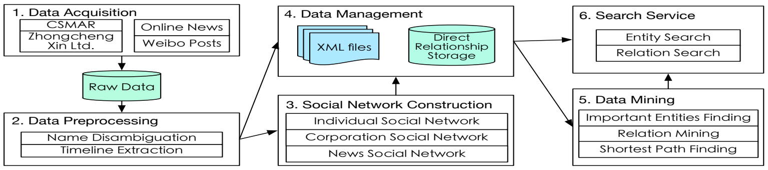


Fig. 1: Architecture of CLANS

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 [4] [9].

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 we collect data and make it available to use; 2) how to construct and quantify business social network; 3) how to mine more implicit information from social network.

We address the problems with our novel approaches: 1) we collect data from multiple sources and do the preprocessing to make it available to use; 2) we construct a business social network and formulate similarity relations among individuals and corporations; 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 details in addressing the three major issues. We present our system in website version in Section 3. 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 multiple sources (Section 2.2). Then we conduct Data Preprocessing (Section 2.3) and Social Network Construction (Section 2.4) to create entities and relations respectively. Then, all entities are stored in XML files for Data Management, with an auxiliary database to store relations [13]. After that, CLANS conduct Data Mining (Section 2.5) and provide Search Services (Section 2.6) with the latest data.

2.2 Data

We collect raw data related to listed firms and senior managers from two major sources. The first source is the biography of each of these corporate leaders among all the listed firms reported in the annual reports, which are available electronically in China Securities Market and Accounting Research Database (CSMAR) and a personnel database compiled by Zhongcheng Xin Ltd. Each record in CSMAR contains company's stock code, report year, individual name, position, gender, age, title, education background, biography and payment, from year 1999 to 2012, with totally 399,216 records and about 2,500 companies. The data provided by Zhongcheng Xin Ltd. contains the structured work experiences, like company, position, start year and end year, and education experiences, like college, major, degree, start date and end date. It covers 84,859 records and 68,289 people. But it do not cover every people in CSMAR.

The second source is the online news and Sina Weibo posts related to all senior managers and listed firms. We crawled all news in `"http://news.baidu.com/"` by searching the name of all corporate leaders and firms. In total, we get 1,126,299 company related news and 16,374,279 people news. As Sina Weibo has become the most important micro-blogging platform in China, and news agents are more likely to utilize it to public news, we also crawled Weibo posts related to our target names. We get 2,367,619 company related posts and 19,445,929 individual posts.

2.3 Data Preprocessing

We conduct data preprocessing to create individual entities, identify related online news and Weibo posts, and extract individual detail structured timeline information.

Name Disambiguation. In this stage we encounter and tackle two major issues. Problem one is that a certain person matches multiple records, so we need to extract unique individual entities. In CSMAR, people may stay in one company for several years so he appears in annual records repeatedly, or he works in many companies so he appears in multiple companies' annual records. Meanwhile, although the data in Zhongcheng Xin Ltd. already has unique identifiers for each people, we need to map people to CSMAR. To figure out this problem, our solution is that, if two records share a high similarity of cognizable features (like name, age, gender, company and birthplace) over a defined threshold, we consider them as the same person. In this way, we identify 87,458 individual entities in CSMAR and find the common 46,130 people in two datasets.

Problem two is that a popular name in our database matches multiple people from Internet sources. A name may map multiple individual entities in our database or a name in our database may map multiple people in online news and Weibo posts. As Weibo posts are shorts and using a nonstandard language with similarities to SMS style, online news is much longer than Weibo posts, and the language styles used in news are much more standard. Then we use a two-round labeling method to identify Weibo posts and a third-round labeling method to identify related online news to a certain person. For the first round, we use the related key words (like company stock name, positions, and college) to the certain person in our database to do the filtering. Then we get a small size of reliable data set in which most posts are truly related to the people in our dataset. In the second round, we use the labeled result from the first round as training set, then transform the text into vector expression based on tf-idf [8] and calculate the cosine similarity [11] of the unlabeled posts to it. After selecting above a threshold, we get the final related Weibo posts. Same as the first two rounds in Weibo posts, in the third round for online news, we apply the Latent Semantic Allocation [2] to the labeled result from the second round, map all the documents to vectors in the lower dimensional latent semantic space, calculate documents similarities, set the threshold and label the rest news. We select a random sample of 1000 posts and news, and it shows that the problem is solved by a precision rate of 98% for Weibo posts and 86% for online news.

Timeline Extraction. For people in CSMAR, who are not covered in Zhongcheng Xin Ltd. data, 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 [3], owning to expression's diversity and complexity, and we achieve a precision rate of 83.1%.

2.4 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. Then we construct the news social network for individual and corporation based on Weibo posts and online news.

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 present the alumni, colleague and integrated individual social network in the website, shown in Fig. 2b.

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

$$W_{i,j} = \alpha\omega^{al}_{i,j} + \beta\omega^{co}_{i,j} \ , \tag{2}$$

where $\omega^{al}_{i,j}$ 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. We present the corporation social network in the website, shown in Fig. 2c.

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:

$$\omega^{gp}_{i,j} = \sum_{p^k_i \in P_i \cap P_j} PS_{p^k_i} * \omega^{gp}_{p^k_i} \ , \tag{3}$$

$$\omega^{nk}_{i,j} = \sum_{(p^k_i, p^r_j) \in L_2(P_i,P_j)} PS_{p^k_i} * \omega^{nk}_{p^k_i,p^r_j} \ . \tag{4}$$

$PS_{p^k_i}$ 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.

News Social Network Construction. We construct the news social network for individual and corporation based on Weibo posts and online news. As Weibo posts and online news publish multiple news mentioned corporate leaders and firms every day, it is important to identify how they are connected with each other and provide these information to investors. We have already identified each Weibo post or online news to an individual entity (Section 2.3), so we construct their social network by identifying two individuals or corporations share the same posts or news. Using this way, we find all the relations among corporate leaders and corporations, and present their relations in the website, as shown in Fig. 2b and 2c.

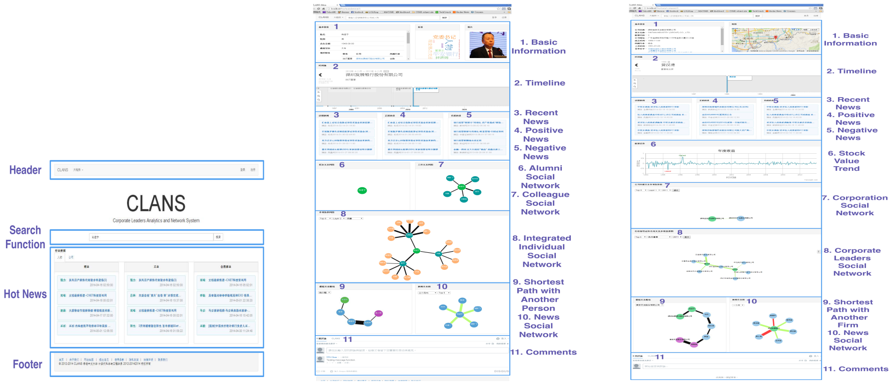
2.5 Data Mining

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. We integrate this with search result, so users could find important entities when they search. For individuals finding, our algorithm is refer to the work of [14], 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. For corporations finding, we apply PageRank [10] algorithm, which only take account of the corporations' relationships.

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 $\hat{e}_{i,j} = \{\tilde{e}^{nk}_{p_i^k,p_j^r}, \cdots, \tilde{e}^{gp}_{p_i^d}, \cdots\}$, where $\tilde{e}^{nk}_{p_i^k,p_j^r}$ denotes a connection between person p_i^k in corporation i and p_j^r in j, and $\tilde{e}^{gp}_{p_i^d}$ denotes person p_i^d 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. As shown in Figure 2b and 2c, users could check the shortest path among people and companies. 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,



(a) Homepage

(b) Person page

(c) Company page

Fig. 2: Sample pages of websites

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 three 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 timeline and relations, shortest path, recent, positive and negative new (we use the sentiment analysis tools in [7]) for both companies and individuals. Fig. 2a shows the homepage, in which shows popular corporations and individuals with their news. Fig. 2b and 2c show the basic information, timeline, temporal relations, news and shortest path of a senior executive and a corporation, which are the results of Section 2.3, Section 2.4 and Section 2.5.

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 language processing, network detection and construction, etc. Thus, we develop CLANS to solve some of these problems. 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.

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. Scott C. Deerwester, Susan T Dumais, Thomas K. Landauer, George W. Furnas, and Richard A. Harshman. Indexing by latent semantic analysis. *JASIS*, 41(6):391–407, 1990.
3. Sean R. Eddy. Profile hidden markov models. *Bioinformatics*, 14(9):755–763, 1998.
4. Irwin King, Jiexing Li, and Kam Tong Chan. A brief survey of computational approaches in social computing. In *IJCNN*, pages 1625–1632. IEEE, 2009.
5. 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.
6. David F Larcker, Eric C So, and Charles CY Wang. *Boardroom centrality and stock returns.* Citeseer, 2010.
7. Tak Pang Lau, Shuai Wang, Yuanyuan Man, Chi Fai Yuen, and Irwin King. Language technologies for enhancement of teaching and learning in writing. In *Proceedings of the companion publication of the 23rd international conference on World wide web companion*, pages 1097–1102. International World Wide Web Conferences Steering Committee, 2014.
8. Christopher D Manning, Prabhakar Raghavan, and Hinrich Schütze. *Introduction to information retrieval*, volume 1. Cambridge university press Cambridge, 2008.
9. 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.
10. Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. The pagerank citation ranking: bringing order to the web. 1999.
11. Amit Singhal. Modern information retrieval: A brief overview. *IEEE Data Eng. Bull.*, 24(4):35–43, 2001.
12. 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.
13. Shuai Wang, Yuanyuan Man, Tianyu Zhang, TJ Wong, and Irwin King. Data management with flexible and extensible data schema in clans. *Procedia Computer Science*, 24:268–273, 2013.
14. 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.