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

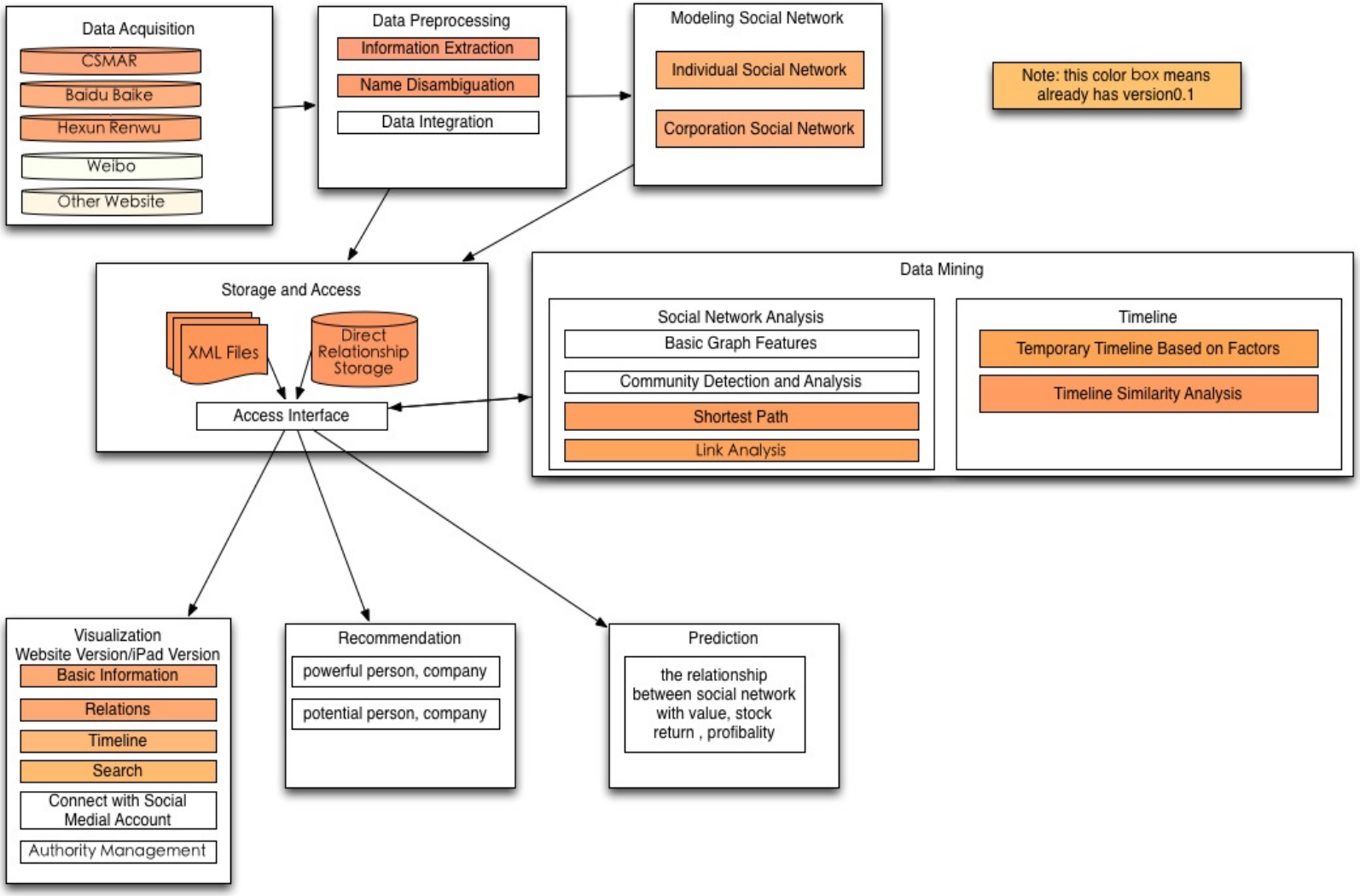
Introduction to CLANS

- 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

证券代码	统计截止日期	姓名	职务类别	具体职务	性别	年龄	教育背景	职称
620	2011-12-31	龚鹏	1400000000	董事(离任)	男	44	0	
620	2010-12-31	龚鹏	1000000000	董事	男	43	3	高级工程师
620	2009-12-31	龚鹏	1000000000	董事	男	42	3	高级工程师
620	2008-12-31	龚鹏	1000000000	董事	男	41	3	高级工程师
620	2007-12-31	龚鹏	1000000000	董事	男	40	3	高级工程师
620	2006-12-31	龚鹏	1000000000	董事	男	39	3	高级工程师
620	2005-12-31	龚鹏	1000000000	董事	男	38	3	高级工程师
526	2009-12-31	龚高	2100000000	监事会主席	男	60	0	
526	2008-12-31	龚高	2100000000	监事会主席	男	59	0	
526	2007-12-31	龚高	2100000000	监事会主席	男	58	0	
526	2002-12-31	龚高	1100000000	副董事长	男	53	0	
659	2004-12-31	龚滢媛	2100000000	监事会主席	女	52	0	
659	2003-12-31	龚滢媛	2100000000	监事会主席	女	51	2	高级政工师

Data Cleaning

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

证券代码	统计截止日期	姓名	职务类别	具体职务	性别	年龄	教育背景	职称
620	2011-12-31	龚鹏	1400000000	董事(离任)	男	44	0	
620	2010-12-31	龚鹏	1000000000	董事	男	43	3	高级工程师
620	2009-12-31	龚鹏	1000000000	董事	男	42	3	高级工程师
620	2008-12-31	龚鹏	1000000000	董事	男	41	3	高级工程师
620	2007-12-31	龚鹏	1000000000	董事	男	40	3	高级工程师
620	2006-12-31	龚鹏	1000000000	董事	男	39	3	高级工程师
620	2005-12-31	龚鹏	1000000000	董事	男	38	3	高级工程师
526	2009-12-31	龚高	2100000000	监事会主席	男	60	0	
526	2008-12-31	龚高	2100000000	监事会主席	男	59	0	
526	2007-12-31	龚高	2100000000	监事会主席	男	58	0	
526	2002-12-31	龚高	1100000000	副董事长	男	53	0	
659	2004-12-31	龚滢媛	2100000000	监事会主席	女	52	0	
659	2003-12-31	龚滢媛	2100000000	监事会主席	女	51	2	高级政工师

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

证券代码	统计截止日期	姓名	职务类别	具体职务	性别	年龄	教育背景	职称
620	2011-12-31	龚鹏	1400000000	董事(离任)	男	44	0	
620	2010-12-31	龚鹏	1000000000	董事	男	43	3	高级工程师
620	2009-12-31	龚鹏	1000000000	董事	男	42	3	高级工程师
620	2008-12-31	龚鹏	1000000000	董事	男	41	3	高级工程师
620	2007-12-31	龚鹏	1000000000	董事	男	40	3	高级工程师
620	2006-12-31	龚鹏	1000000000	董事	男	39	3	高级工程师
620	2005-12-31	龚鹏	1000000000	董事	男	38	3	高级工程师
526	2009-12-31	龚高	2100000000	监事会主席	男	60	0	
526	2008-12-31	龚高	2100000000	监事会主席	男	59	0	
526	2007-12-31	龚高	2100000000	监事会主席	男	58	0	
526	2002-12-31	龚高	1100000000	副董事长	男	53	0	
659	2004-12-31	龚淑媛	2100000000	监事会主席	女	52	0	
659	2003-12-31	龚淑媛	2100000000	监事会主席	女	51	2	高级政工师

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

stock_id	id	finish_year	name	position_classi	position	gender	age	education
1	30	2010-12-31	刘宝瑞	00390000N0	副行长(离任)	男	0	0
1	34	2009-12-31	刘宝瑞	10300000N0	"董事,副行长"	男	52	0

  - Contradictory data

stock_id	id	finish_year	name	position_classi	position	gender	age	education
11	561	2011-12-31	曹子扬	2300000000	监事会主席(离任)	男	60	0
11	588	2000-12-31	曹子扬	1000000000	董事	男	0	0
11	563	2010-12-31	曹子扬	2100000000	监事会主席	男	60	2

  - Spelling error

stock_id	id	finish_year	name	position_classi	position	gender	age	education	title
998	10703	2010-12-31	Emmanuel KOUGLEK	1000000000	董事	男	58	3	
2032	11876	2010-12-31	Fran?ois LECLEIRE	2300000000	监事(离任)	男	32	0	
2032	11880	2009-12-31	Francois LECLEIRE	2100000000	监事会主席	男	31	4	

Data Cleaning

- Data Correction
  - Result

Type	Error	Solution
missing data	859 entries with age=0	have been marked or replaced with correct age from deleted duplicate entries
missing data	18221 entries don not have introduction	need data acquisition
contradictory data	2354 couples of entries with different age	the smaller group have been removed
spelling error	593 couple of entries with different name	checked manually and correct the errors if string-distandce less than a threshold

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

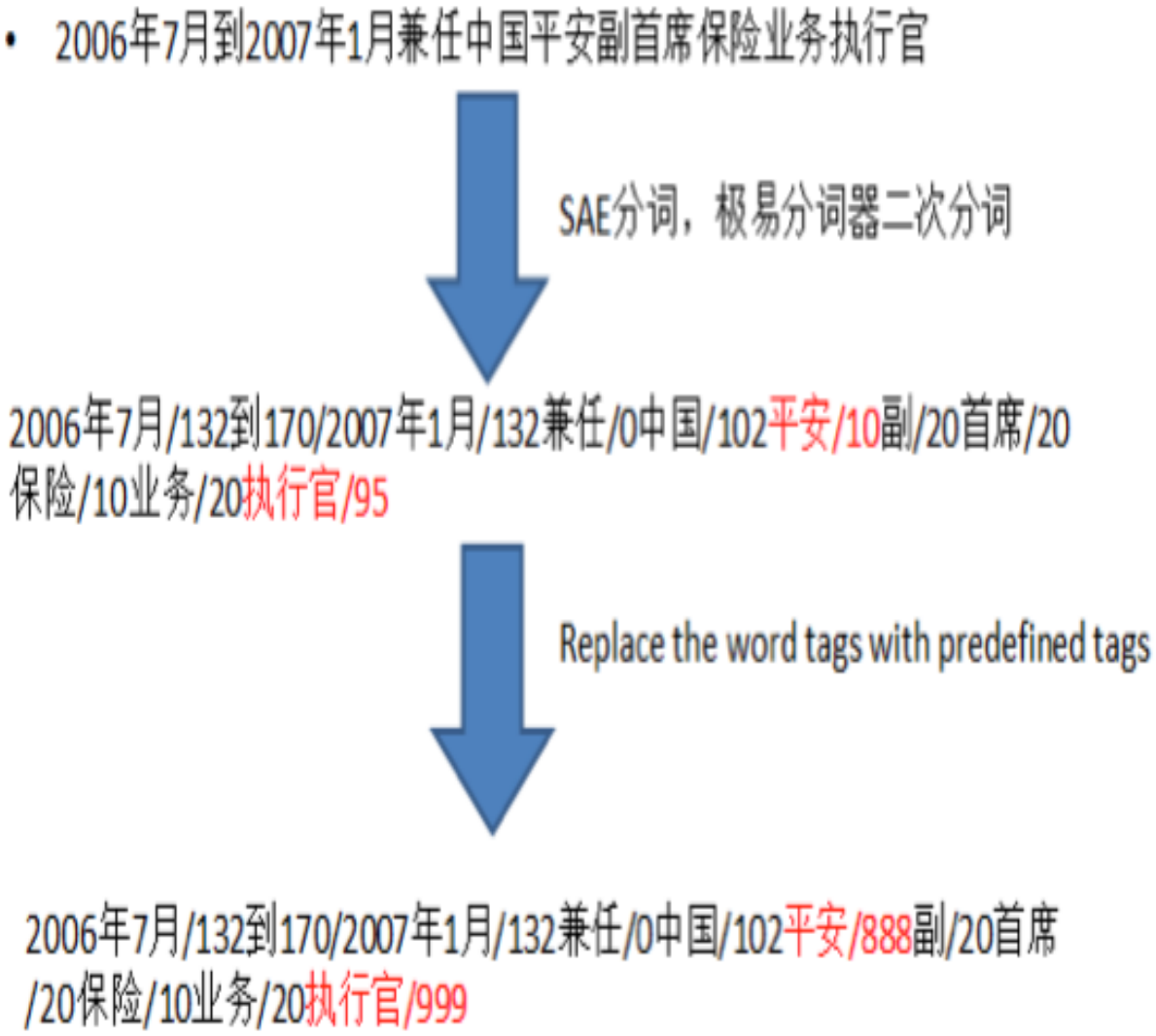
Information Extraction

- 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...
    - ...

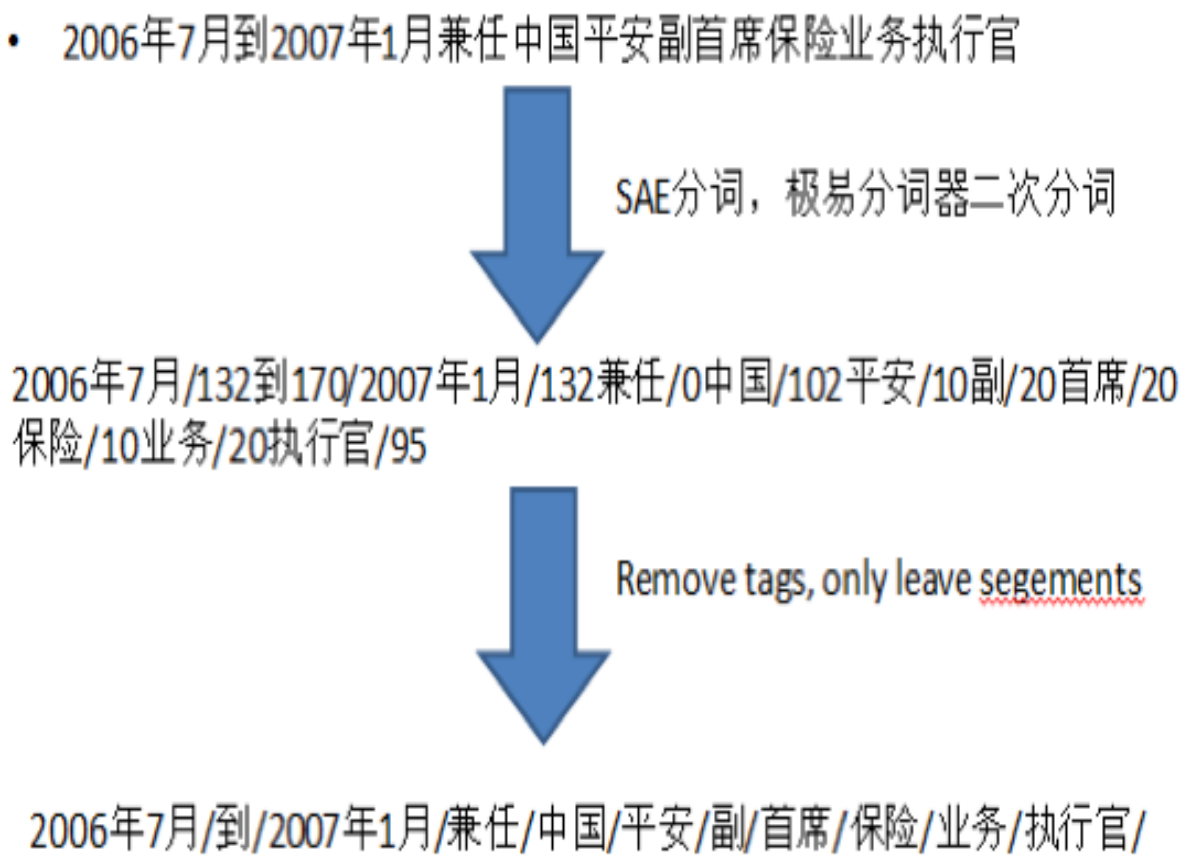
Information Extraction



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



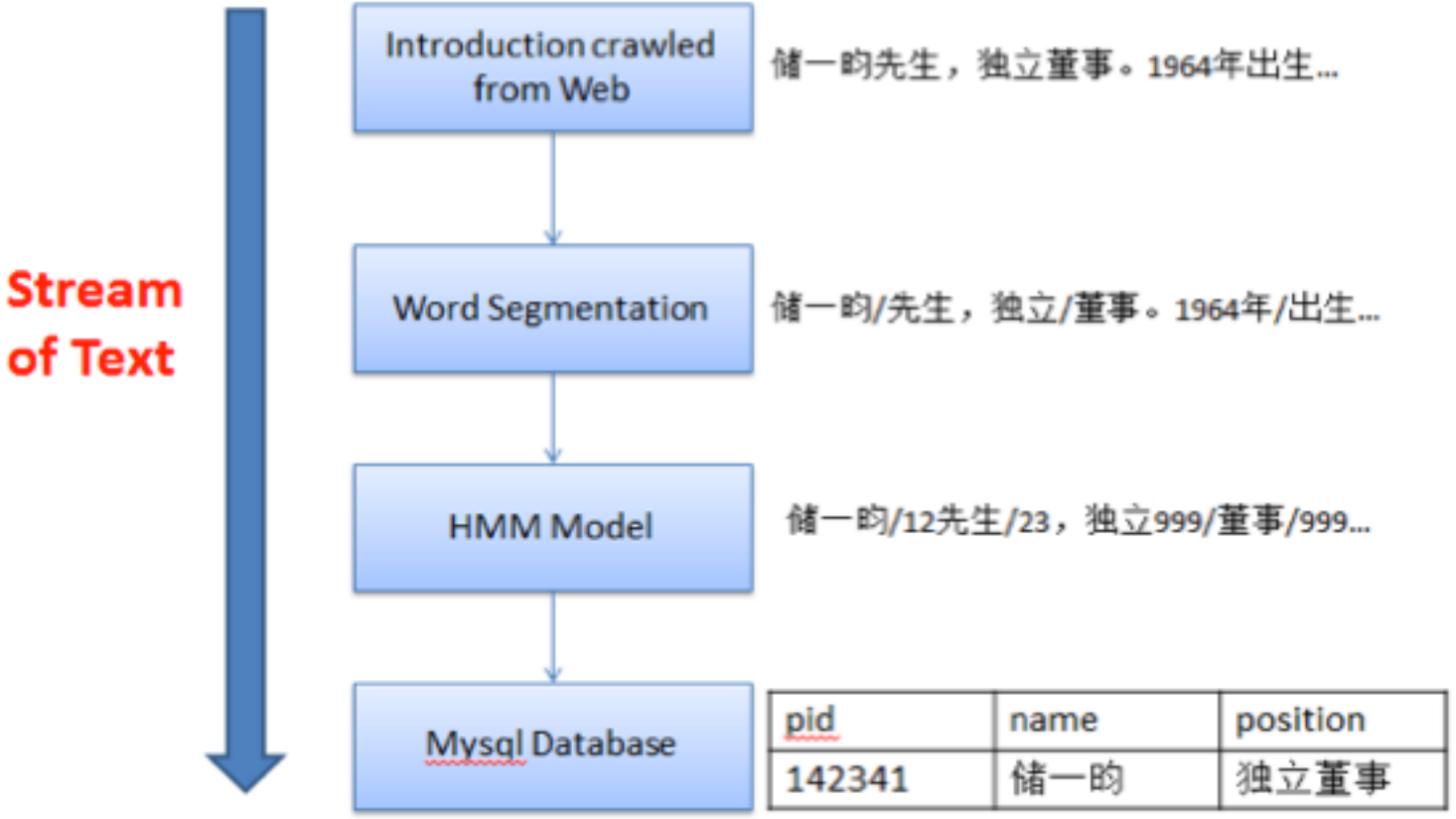
- HMM model
  - Words segmentation with word tags on testing samples, then remove the tags

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

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

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.

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

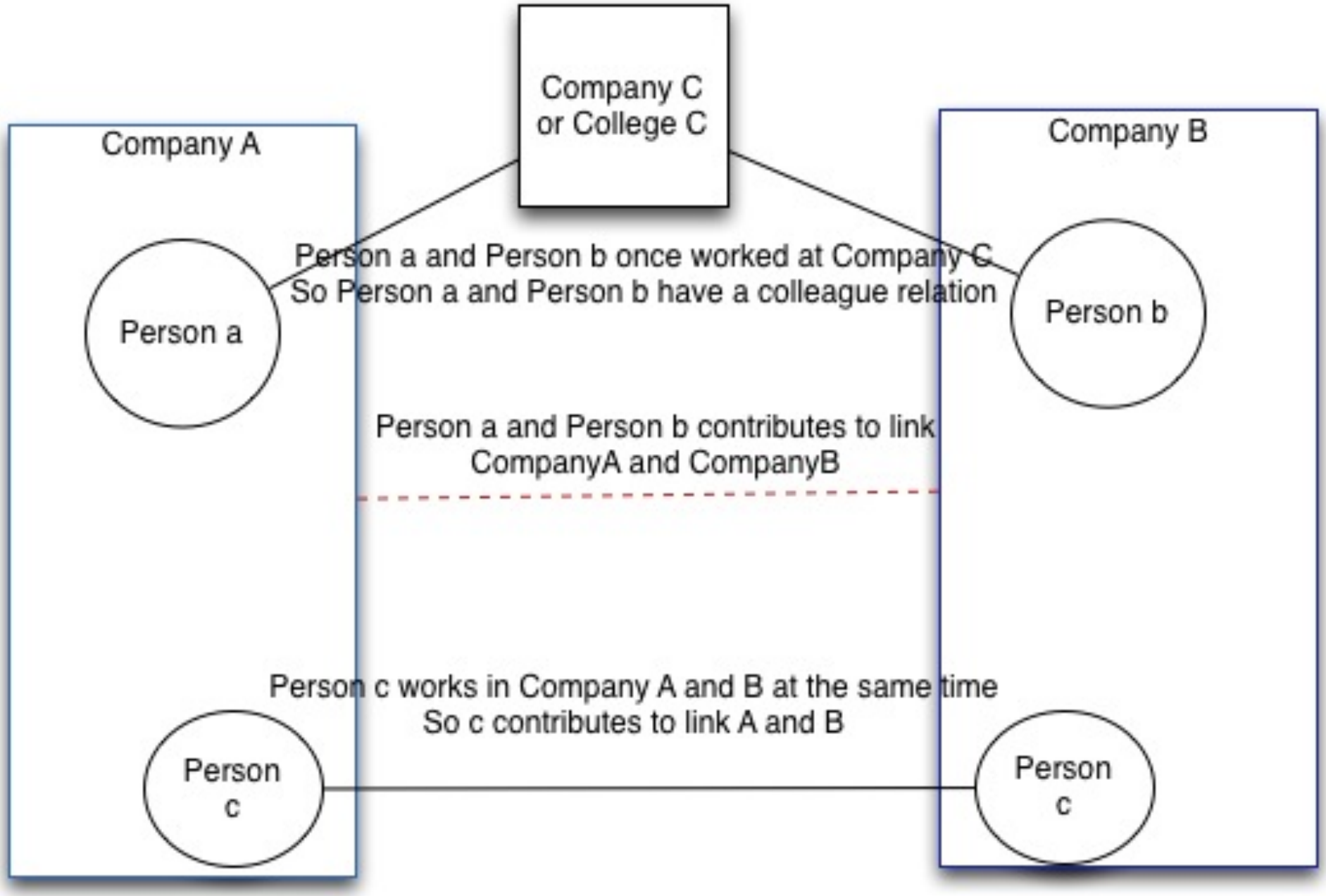
$$W_{i,j} = \alpha\omega_{i,j}^{al} + \beta\omega_{i,j}^{co}$$

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

- Will add family, friends, corporation social network to the whole individual social network

Corporation 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* $P_i = \{p_i^1, p_i^2, \cdots, p_i^n\}$ *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.*


Company C
or College C
Company A
Company B
Person a and Person b once worked at Company C
So Person a and Person b have a colleague relation
Person a
Person b
Person a and Person b contributes to link
CompanyA and CompanyB
Person c works in Company A and B at the same time
So c contributes to link A and B
Person
c
Person
c

Corporation Social Network

$\omega^{gp}_{i,j}$, $\omega^{nk}_{i,j}$ *are defined as follows:*

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

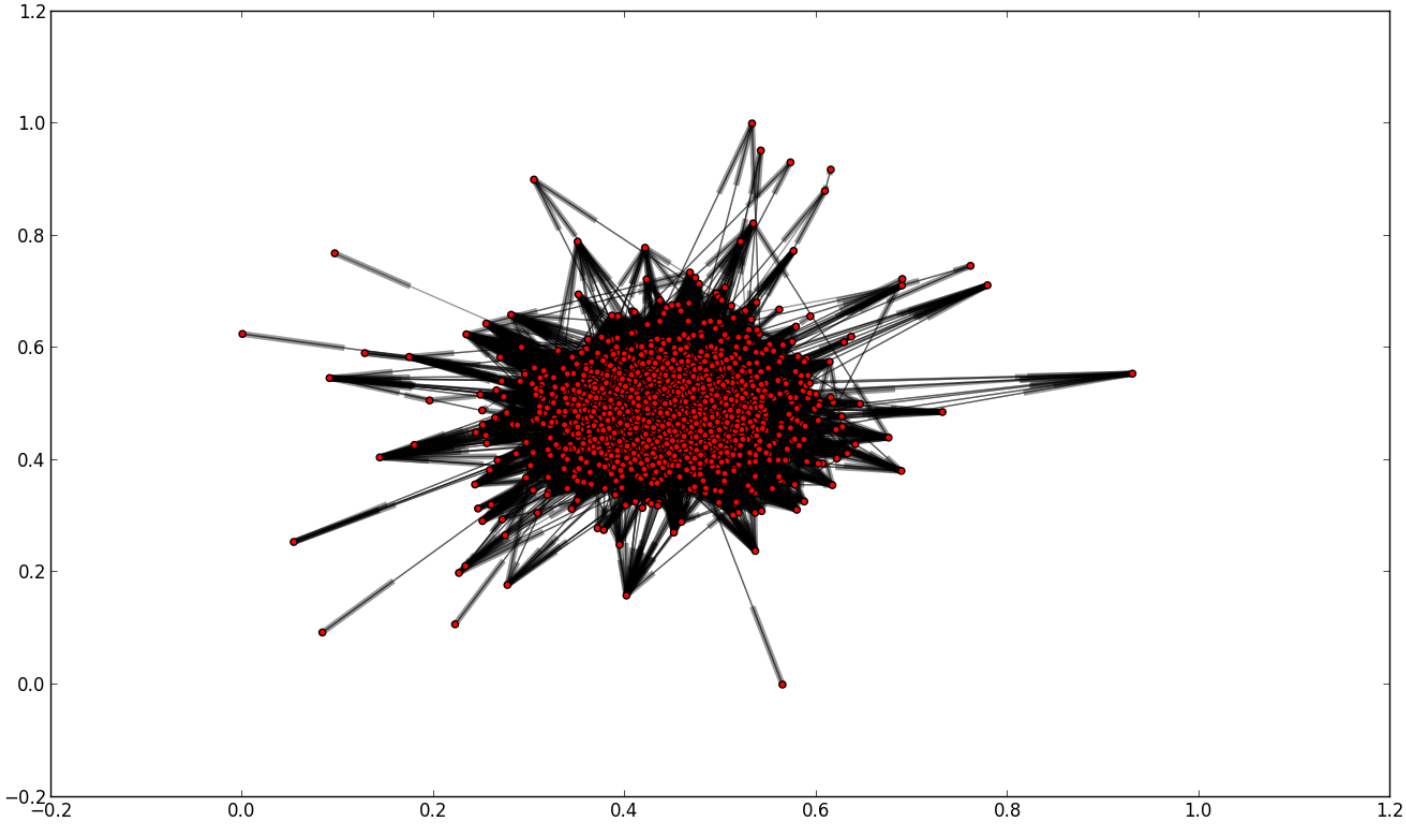
$$\omega^{nk}_{i,j} = \sum_{(p^k_i, p^r_j) \in L_2(P_i, P_j)} PS_{p^k_i} * \omega^{nk}_{p^k_i, p^r_j} \tag{3}$$

$PS_{p^k_i}$ *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.*

Corporation Social Network

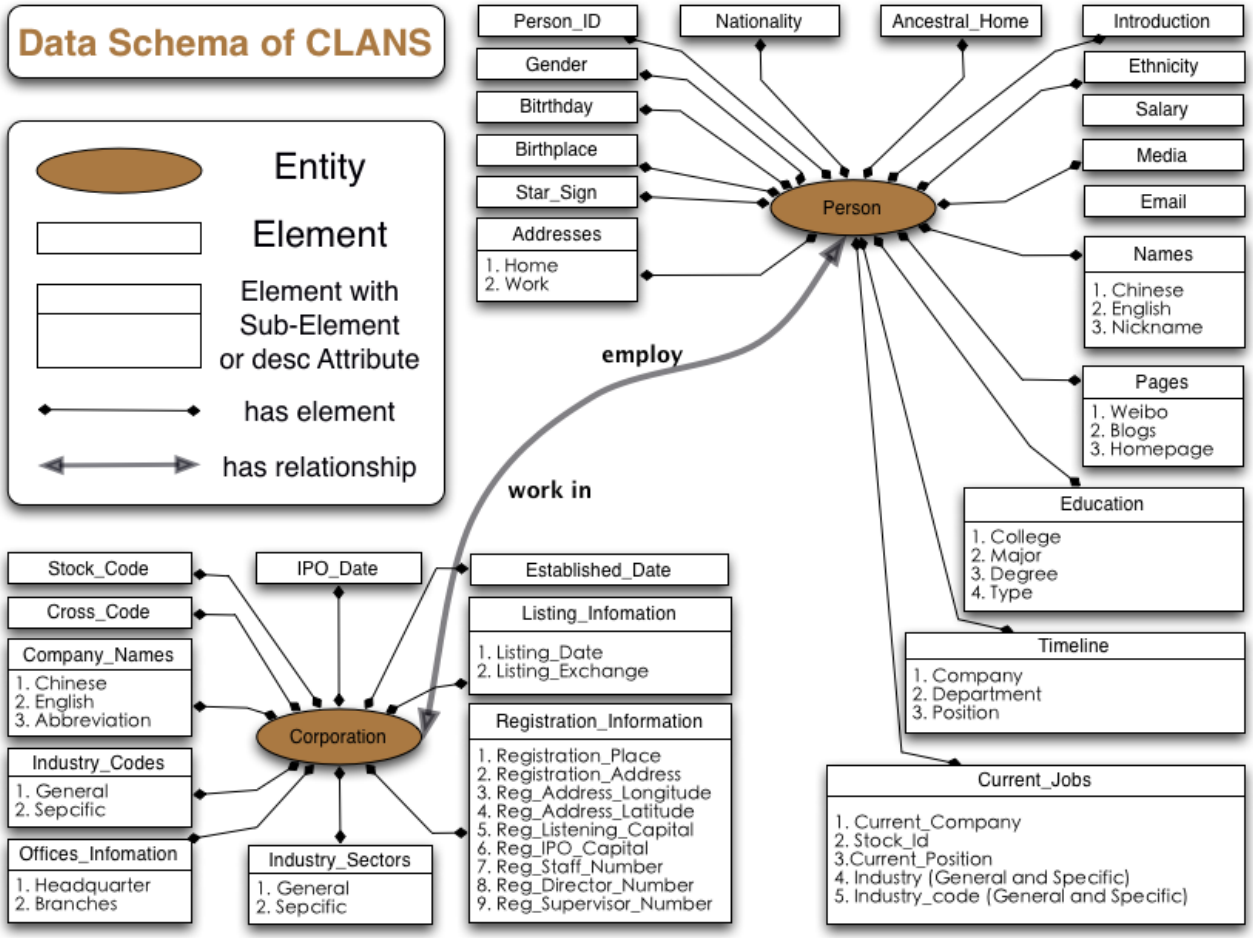
Thus, the corporation weight from corporation i to j is defined as $W_{i,j} = \alpha\omega_{i,j}^{gp} + \beta\omega_{i,j}^{nk}$, where α and β denotes the corresponding percentage that the two relations contribute to the corporation social network respectively.

Corporation Social Network



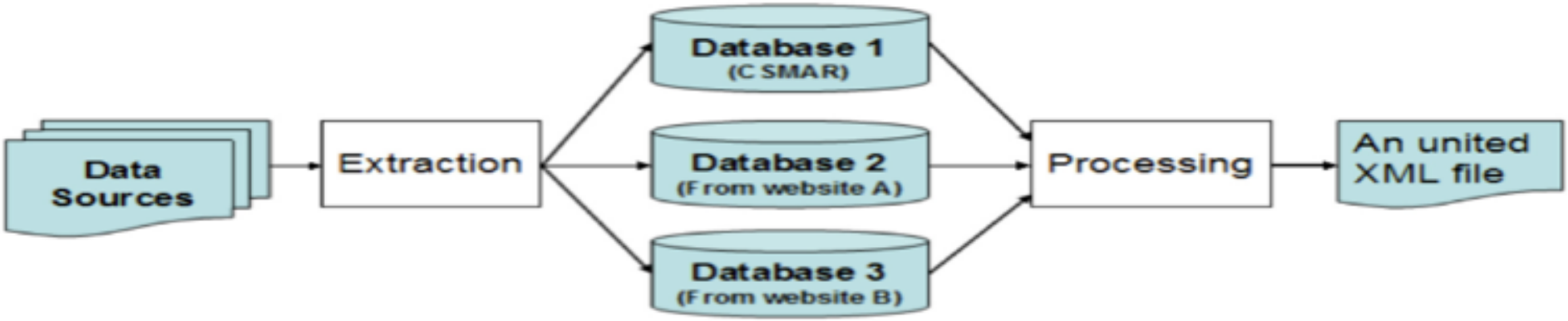
Data Management

- 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

证券代码(stock_id=1)的公司关系(按weight排序，最高前20)

	Stkcd1	Name of company1	Stkcd2	Name of company2	weight	relation_rank/set new value	visited or not
点击查看详细信息	1	深圳发展银行股份有限公司	601328	交通银行股份有限公司	9.696827534	4 / 0	visited
点击查看详细信息	1	深圳发展银行股份有限公司	600999	招商证券股份有限公司	9.61491685	4 / 0	visited
点击查看详细信息	1	深圳发展银行股份有限公司	600036	招商银行股份有限公司	7.54981625	4 / 0	visited
点击查看详细信息	1	深圳发展银行股份有限公司	601998	中信银行股份有限公司	7.22696608	4 / 0	visited
点击查看详细信息	1	深圳发展银行股份有限公司	601318	中国平安保险(集团)股份有限公司	4.915842336	3 / 0	visited
点击查看详细信息	1	深圳发展银行股份有限公司	601818	中国光大银行股份有限公司	4.89615494	3 / 0	visited
点击查看详细信息	1	深圳发展银行股份有限公司	601939	中国建设银行股份有限公司	4.74133307	3 / 0	visited
点击查看详细信息	1	深圳发展银行股份有限公司	601398	中国工商银行股份有限公司	4.40156822	3 / 0	visited
点击查看详细信息	1	深圳发展银行股份有限公司	600011	华能国际电力股份有限公司	3.44539185	3 / 0	visited
点击查看详细信息	1	深圳发展银行股份有限公司	600875	东方电机股份有限公司	3.28828442	3 / 0	visited
点击查看详细信息	1	深圳发展银行股份有限公司	600031	三一重工股份有限公司	3.1442184	3 / 0	visited

Data Mining

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)

	pid1	name1	position1	posi_rank1	pid2	name2	position2	posi_rank2	type	weight	relation_rank/set new value	visited or not
点击查看详细信息	1	肖遂宁	"执行董事,董事长"	1	0	N/A	N/A	0	t	1.0	5 / 0	visited
点击查看详细信息	3	王利平	非执行董事	9	341643	王冬胜	非执行董事	9	w	0.12269592	3 / 0	visited
点击查看详细信息	4	姚波	非执行董事	9	341643	王冬胜	非执行董事	9	w	0.12157635	3 / 0	visited
点击查看详细信息	5	顾敏	非执行董事	9	341643	王冬胜	非执行董事	9	w	0.1047828	3 / 0	visited
点击查看详细信息	6	叶素兰	非执行董事	9	341643	王冬胜	非执行董事	9	w	0.11373936	3 / 0	visited
点击查看详细信息	8	王开国	非执行董事	9	341843	郭宇	监事	0	w	0.13165249	3 / 0	visited
点击查看详细信息	15	储一昀	独立董事	9	341855	朱鹤新	业务总监	4	e	0.1	2 / 0	visited
点击查看详细信息	20	罗康平	外部监事	0	341643	王冬胜	非执行董事	9	w	0.00793538	1 / 0	visited
点击查看详细信息	20	罗康平	外部监事	0	341830	冯婉眉	非执行董事	9	w	0.00793538	1 / 0	visited
点击查看详细信息	20	罗康平	外部监事	0	341856	叶迪奇	副行长	3	w	0.00793538	1 / 0	visited
确认修改												

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、财务控制管理信息系统经理、财务控制部经理、零售银行区域经理、零售风险管理经理、区域经理、特别项目营销经理、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
简历：冯婉眉女士，51岁，本行非执行董事。冯女士现任汇丰集团总经理兼汇丰银行香港区总裁，以及汇丰环球投资管理（香港）有限公司主席及董事、汇丰银行（中国）有限公司副董事长及恒生银行有限公司非执行董事、HSBCMarkets(Asia) Limited董事、HSBC Securities(Japan) Limited董事等职务。冯女士自2010年1月至2011年9月任汇丰集团总经理兼汇丰银行环球银行及资本市场亚太区主管，自2008年5月至2010年1月任汇丰集团总经理兼汇丰银行环球资本市场亚太区主管兼司库，自1996年9月至2008年4月历任汇丰银行港币债券市场主管、亚洲固定收益交易主管、亚太区交易主管、环球资本市场亚太区联席主管兼司库。冯女士1995年于澳大利亚麦考瑞大学获应用财务硕士学位。冯女士自2010年11月起任本行非执行董事。

工作关系详情：

company	start_date	end_date	weight	right	wrong
香港上海汇丰银行	N/A	N/A	0.00793538	◉	○
确认修改					

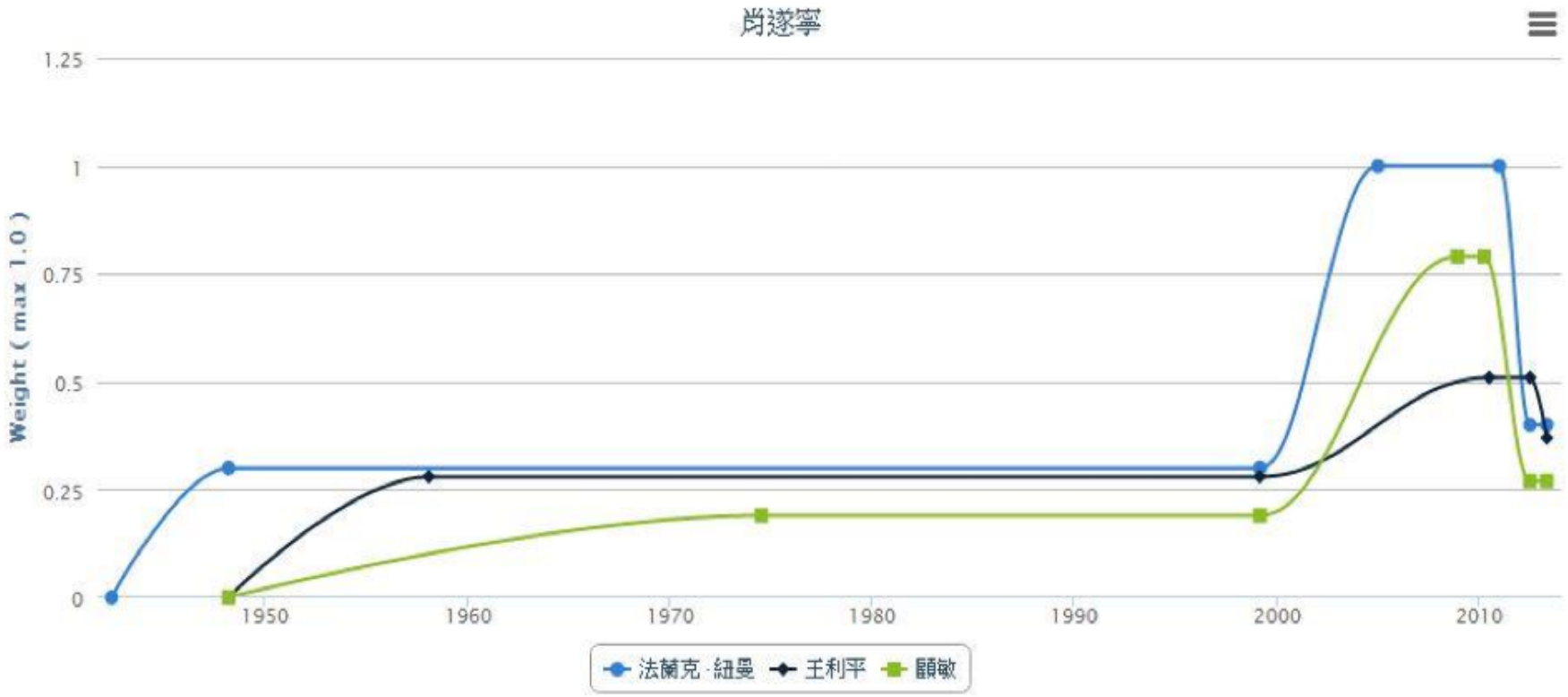
校友关系详情：

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

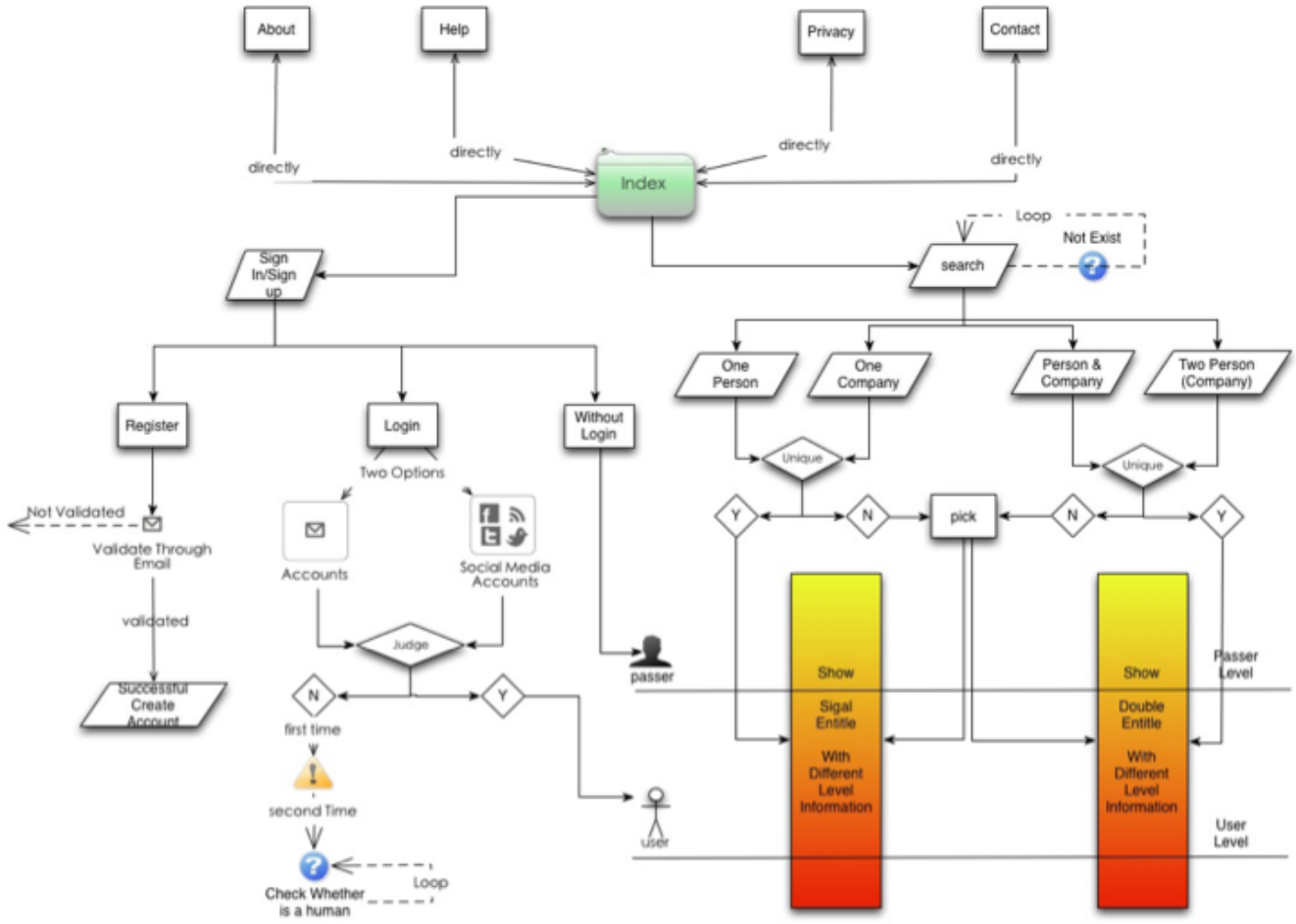
Data Mining

- Temporal Relation Comparison
  - Compare two people's timeline



Visualization

- Website overview



Visualization

- First page

联系我们

localhost:8080/fdmV3.5/SearchAction!search

Google

南京

登录 | 注册

您要找到是不是:

姓名： 李南京
公司： 广州金发科技股份有限公司
职位： 董事,总经理

公司： 南京中北(集团)股份有限公司
地址：江苏省南京市

公司： 南京天龙股份有限公司
地址：江苏省南京市

公司： 南京华东电子集团股份有限公司
地址：江苏省南京市

公司： 南京港股份有限公司
地址：江苏省南京市

公司： 南京红宝丽股份有限公司
地址：南京市

公司： 南京云海特种金属股份有限公司
地址：南京市

公司： 南京科远自动化集团股份有限公司
地址：南京市

公司： 南京新联电子股份有限公司
地址：江苏省南京市

公司： 南京中电联环保股份有限公司
地址：南京市

南京

江苏省省会、副省级城市 更多含义>>

别名:	金陵、建康等
所属地区：	华东地区
电话区号：	025
邮政区码：	210000-213000
地理位置：	江苏省西南部，长江下游

肖遂宁

基本信息 同学圈 同行圈 同事圈

姓　　名:	肖遂宁	**性　　别:**	男
工作单位:	深圳发展银行	**职　　务:**	前董事长
出生日期:	1948年02月	**所属行业:**	银行业
毕业院校:	水电部重庆电力学校	**最高学历:**	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、人力资源、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。

情况总概 关系图 时间轴 最新咨询

腾讯

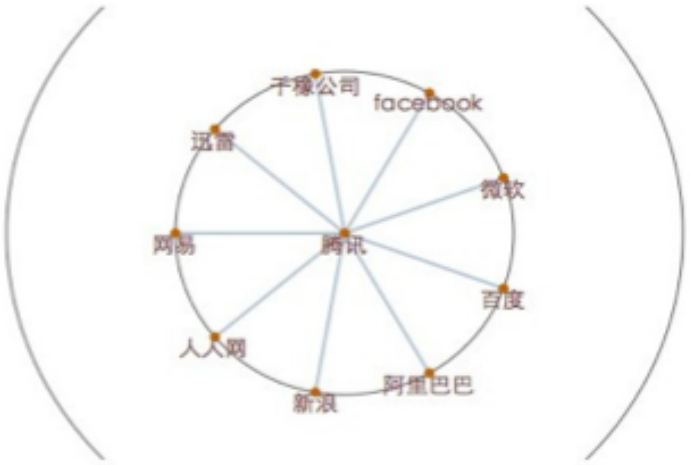
公司概况 | 同行圈

公司名称:	腾讯控股有限公司	**英文名称:**	Tencent
总部地点:	中国·深圳	**证券代码:**	000070
成立日期:	1998年11月	**所属行业:**	互联网
上市日期:	2004年6月	**注册资本:**	100万

简介:

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。成立十多年以来，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。

情况总概 | 关系图 | 时间轴 | 最新咨询



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