15CS065

Users' Public Privacy Mining Software

LIU Jiacheng (53057374)
Supervisor: Dr. LI Shuaicheng

Department of Computer Science
City University of Hong Kong

Introduction

User Needs

How can I know an online user which is totally strange to me?

How can I know my friend more comprehensively through his/her online activity?

How can I know this guy/girl who is actively interacting with me?

I need a user-study application and it must be:

1. Automatic — Least human labor
2. Objective — Voice of data
3. Easy to perceive — Better with graphs
4. Comprehensive — Various perspectives

Project Overview

This project is

A user study application, focusing on Sina's Weibo.com, that satisfies the described user requirements:

1. Automation — Only seed information needed
2. Objectiveness — Models and statistics
3. Ease to perceive — 14 graphs, 6 tables
4. Comprehensiveness — 7 different aspects

general information, friends interaction, interested topics and timeline, etc.

Statistics

Line Count:

- ~25000 lines

Component Count:

- Live modules: 9
- Graphs: 14
- Tables: 6

Outline

I. Introduction

II. System Design

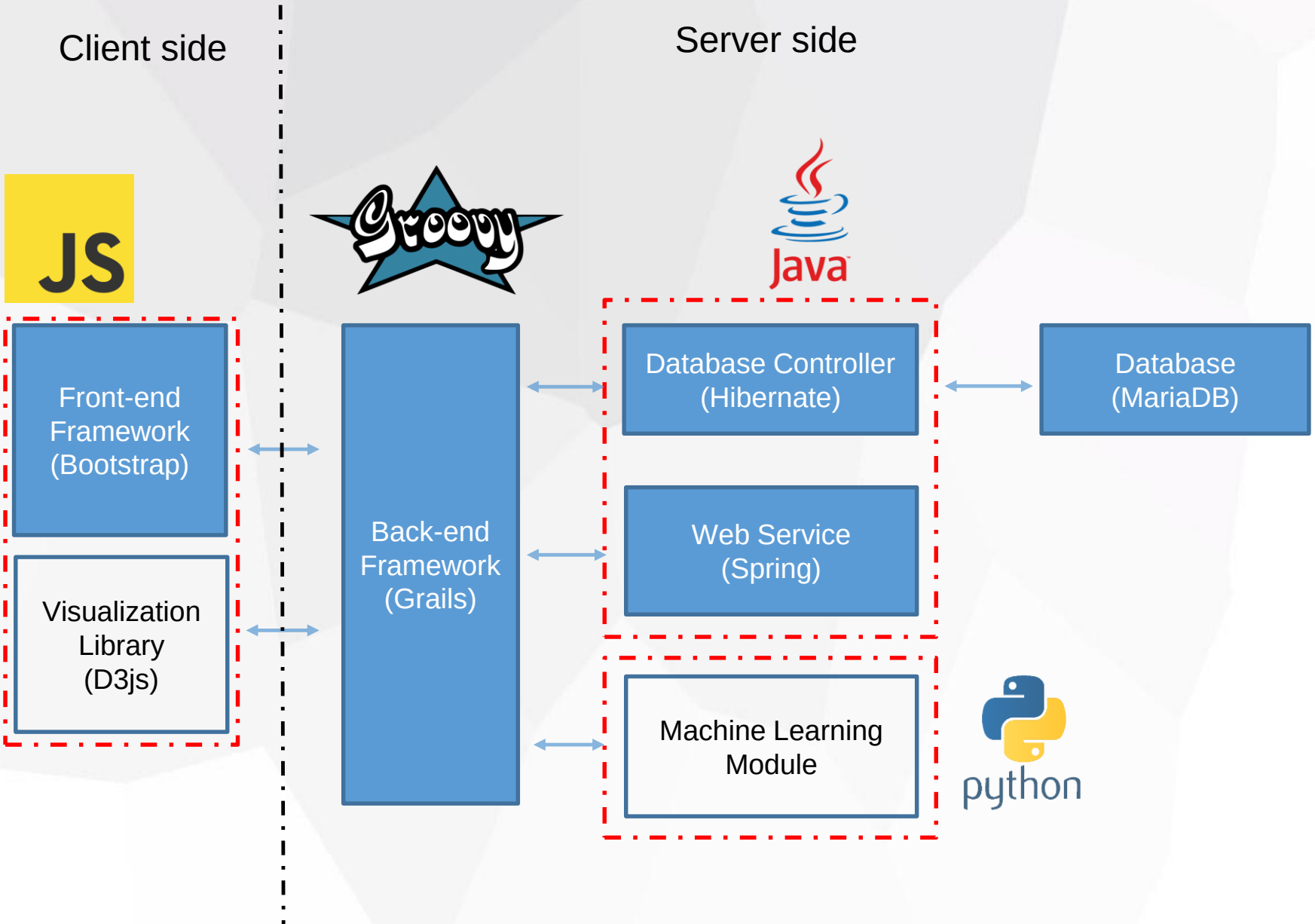
III. A Use Case

IV. Algorithms Applied

V. Conclusion

System Design

Structural Overview



Advantages

1. No dependency on hardware/software.
2. Stronger self-cohesion in front-end and back-end.
3. Exploiting the strengths of each language.

Work flow

User

Clicks

Browser
(Front-end)

Data in JSON

Server
(Back-end)

ID/Name/URL

Entrance Page

ID/Name/URL

A job and user session is initiated

Job status

Complete Notice

Crawler Module

Status Monitor Page (Main Page)

Complete Notice

Keyword Study Module

Complete Notice

General Info Study Module

Keyword Page

General Info Page

More pages

More modules

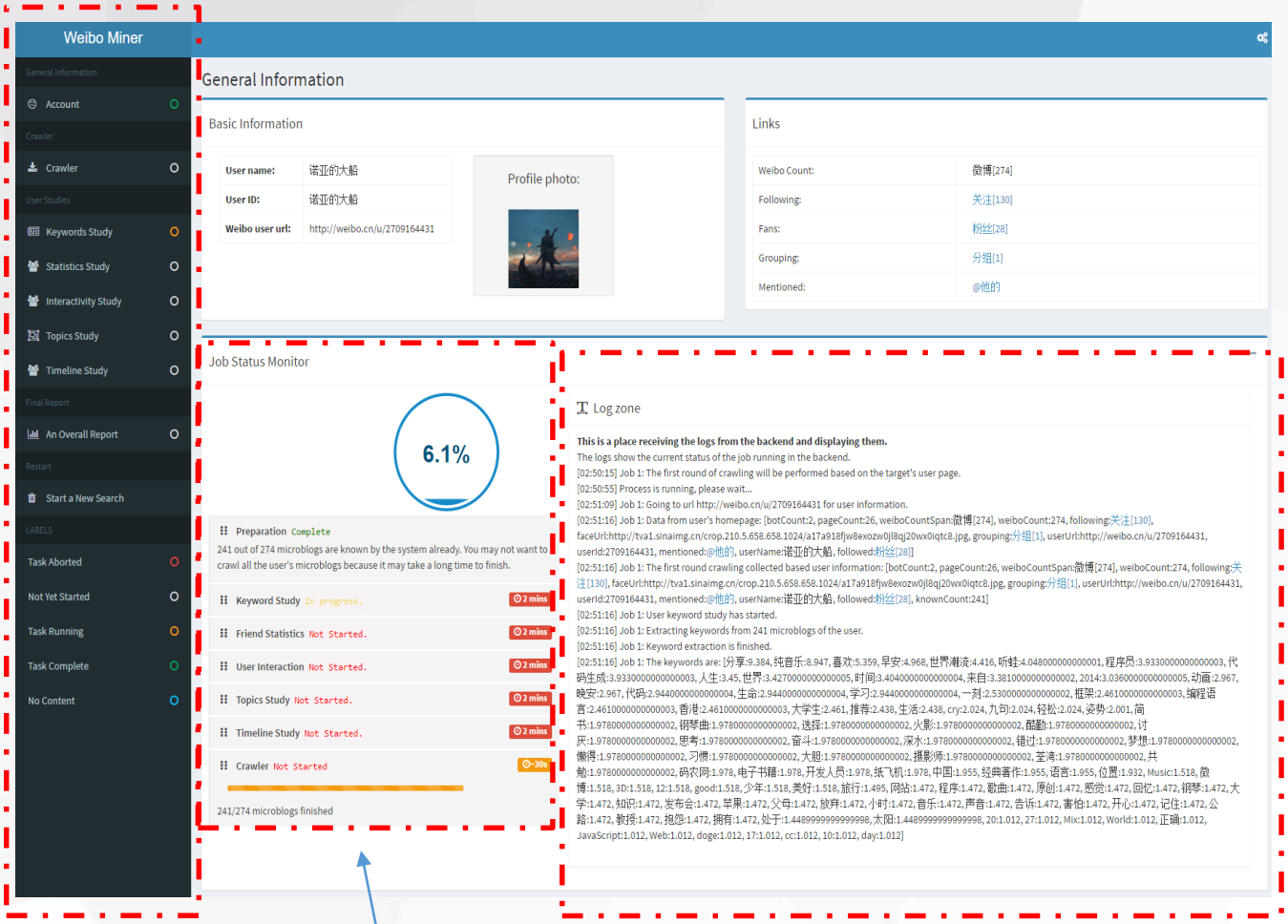
The wrapped web application

Asynchronism improves user experience

TIME

Sample Main Page

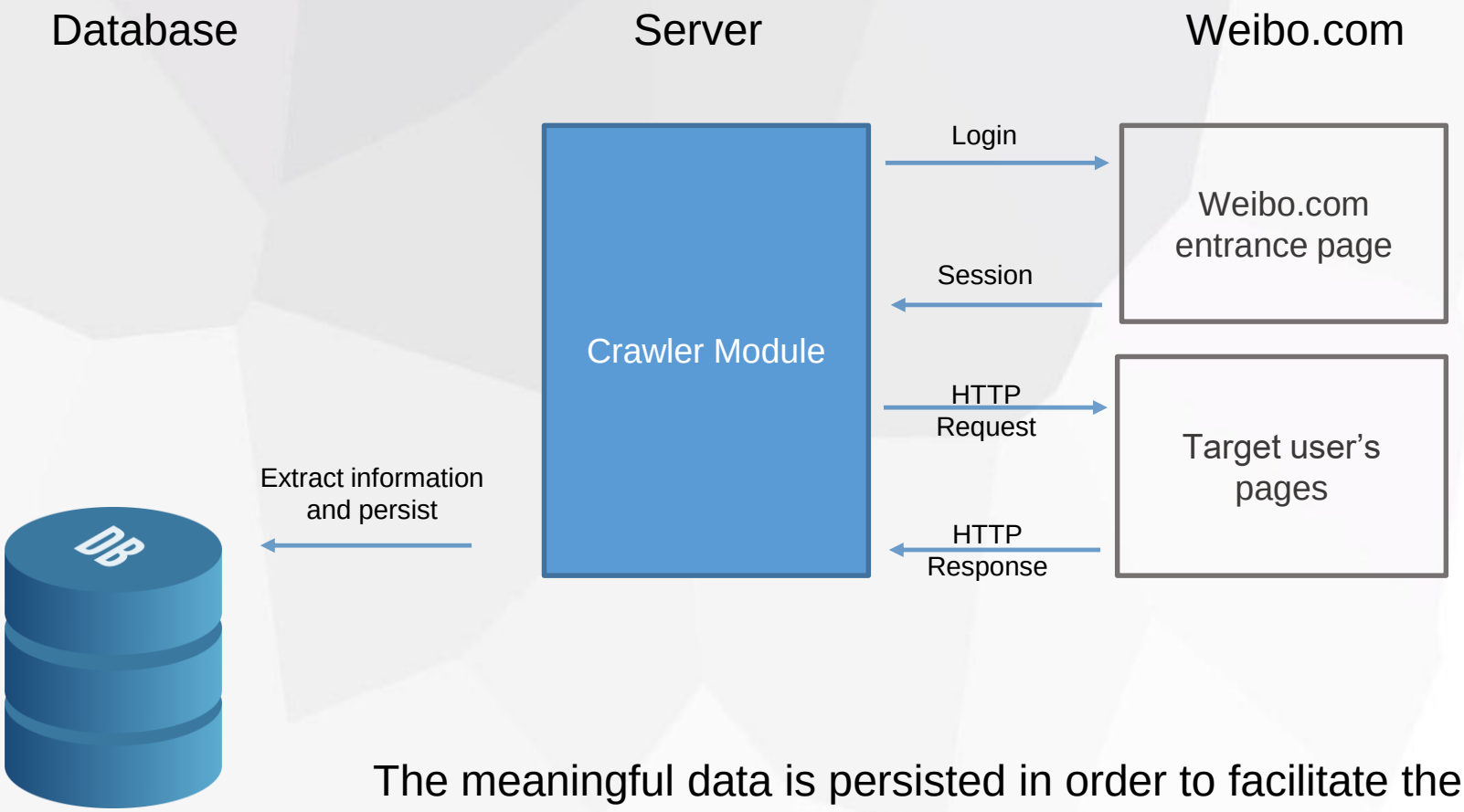
Status Monitoring Sidebar



Detailed module monitoring and time estimation

Logging

Crawler Mechanism

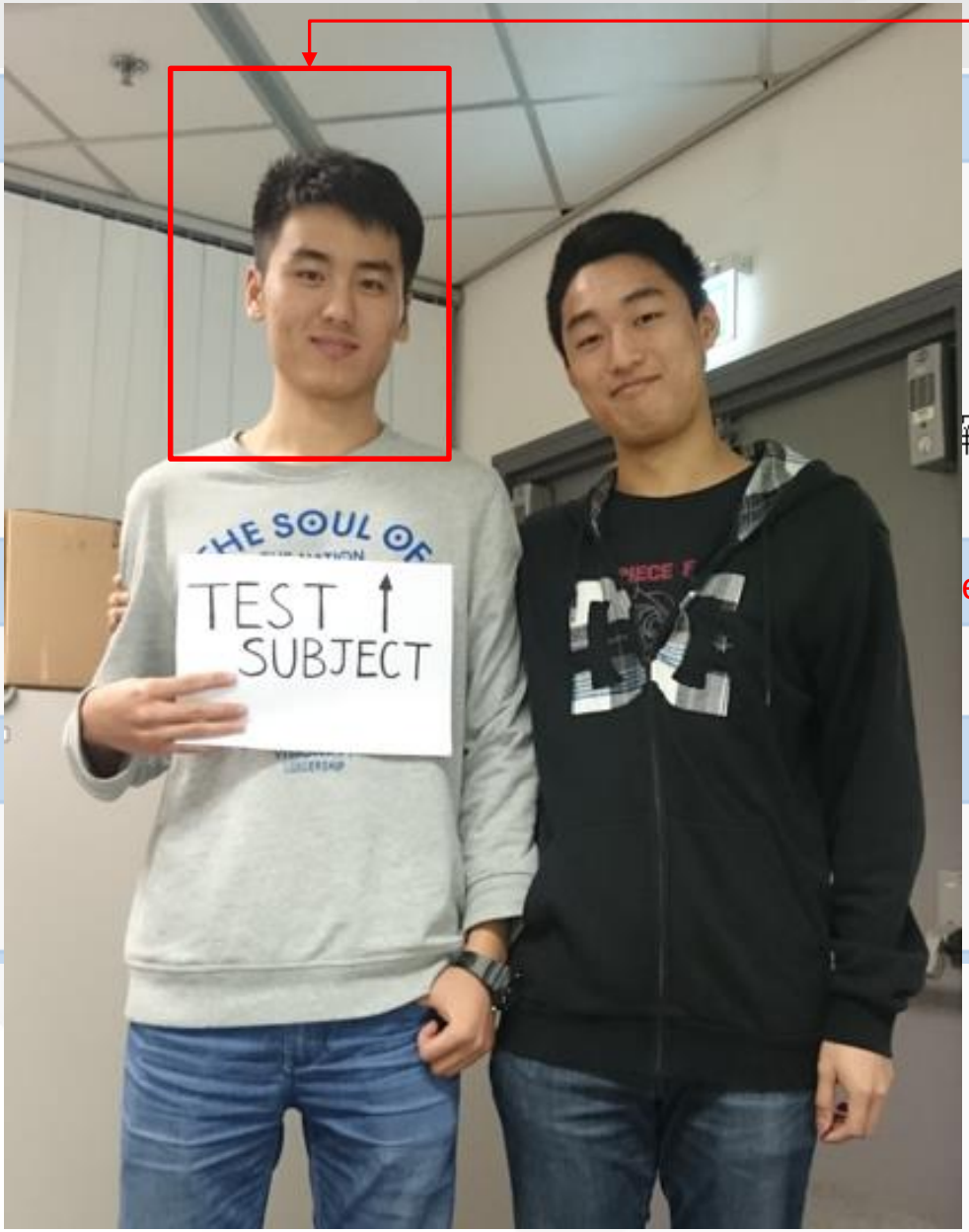


The meaningful data is persisted in order to facilitate the work next time.

A trade-off between speed and storage.

A Use Case

The person of test subject



Name: Mr. FANG

Gender: Male

Age: 22

Occupation:

Research Assistant in CityU

Hobbies:

Music, comics, programming, hiking

Others:

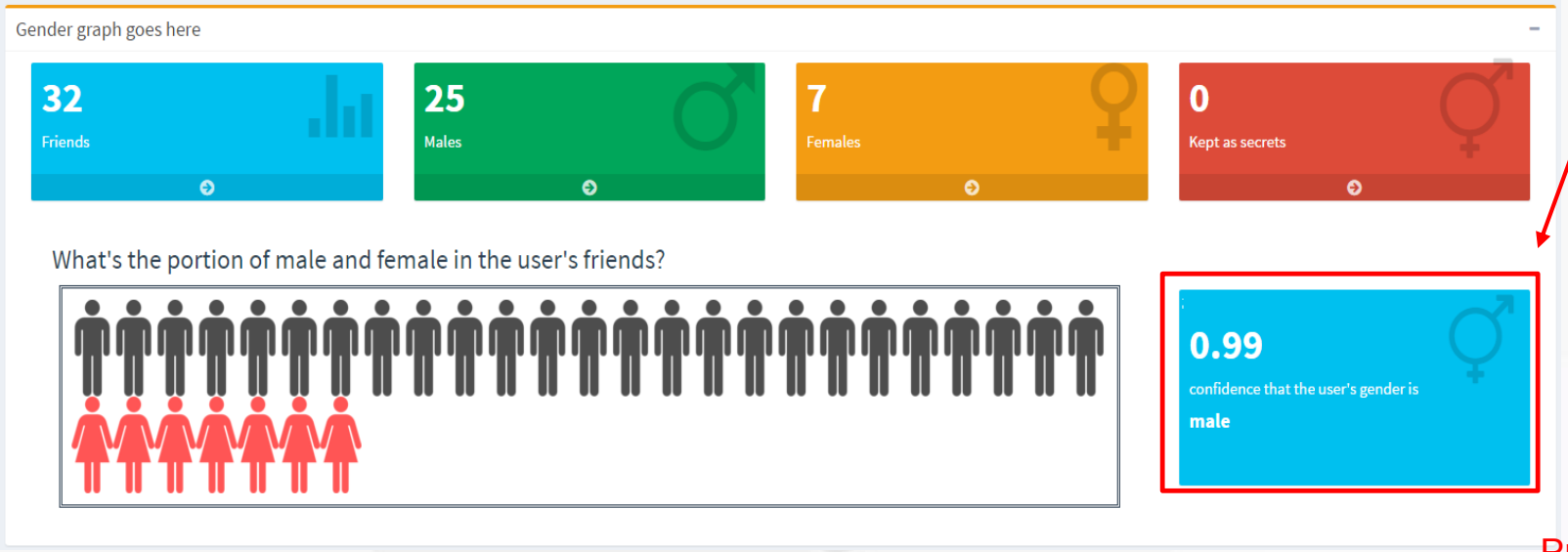
Placement in HSBC in 2014-2015

From Anhui province of China

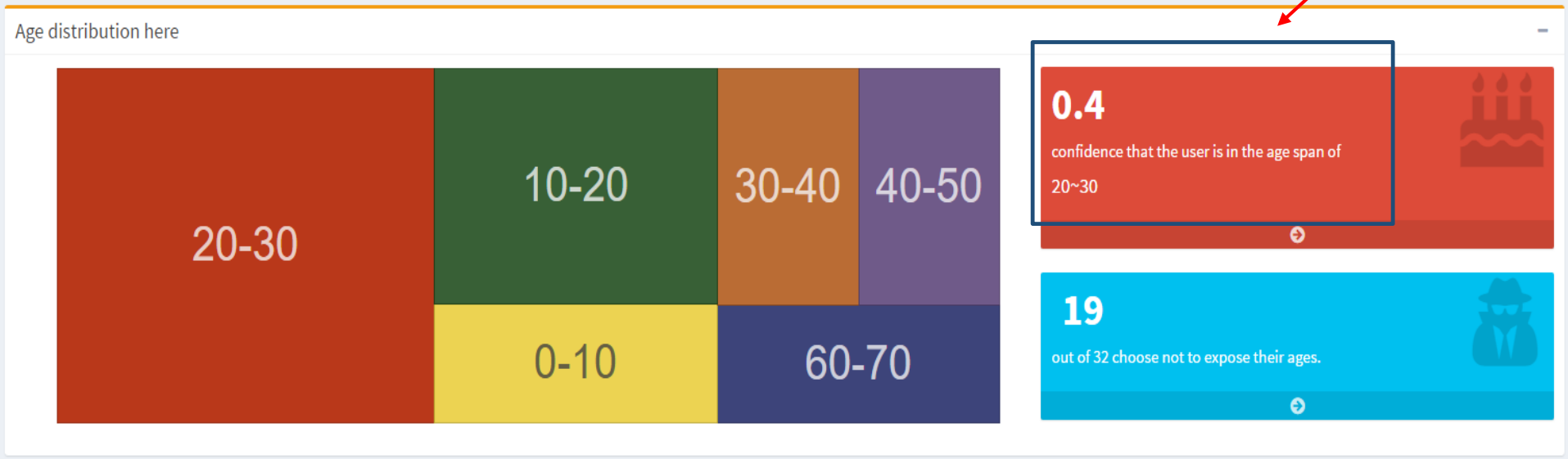
Special thanks to Mr. Fang for his generosity and pertinent feedback!

Basic Information Study

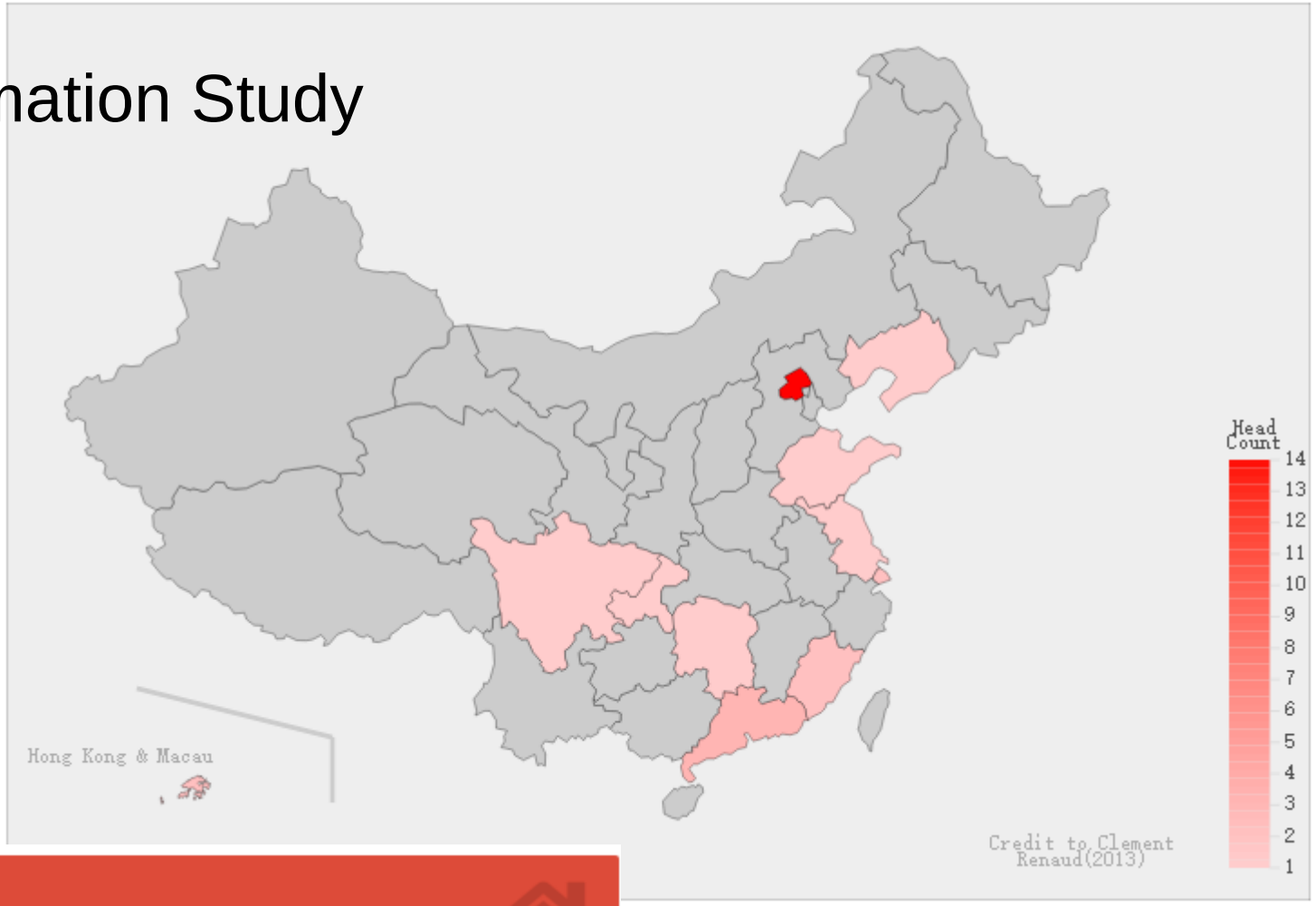
Report if public on user's information page



Predict if not



Basic Information Study



The user may lie so the max confidence can't reach 1.0.

14

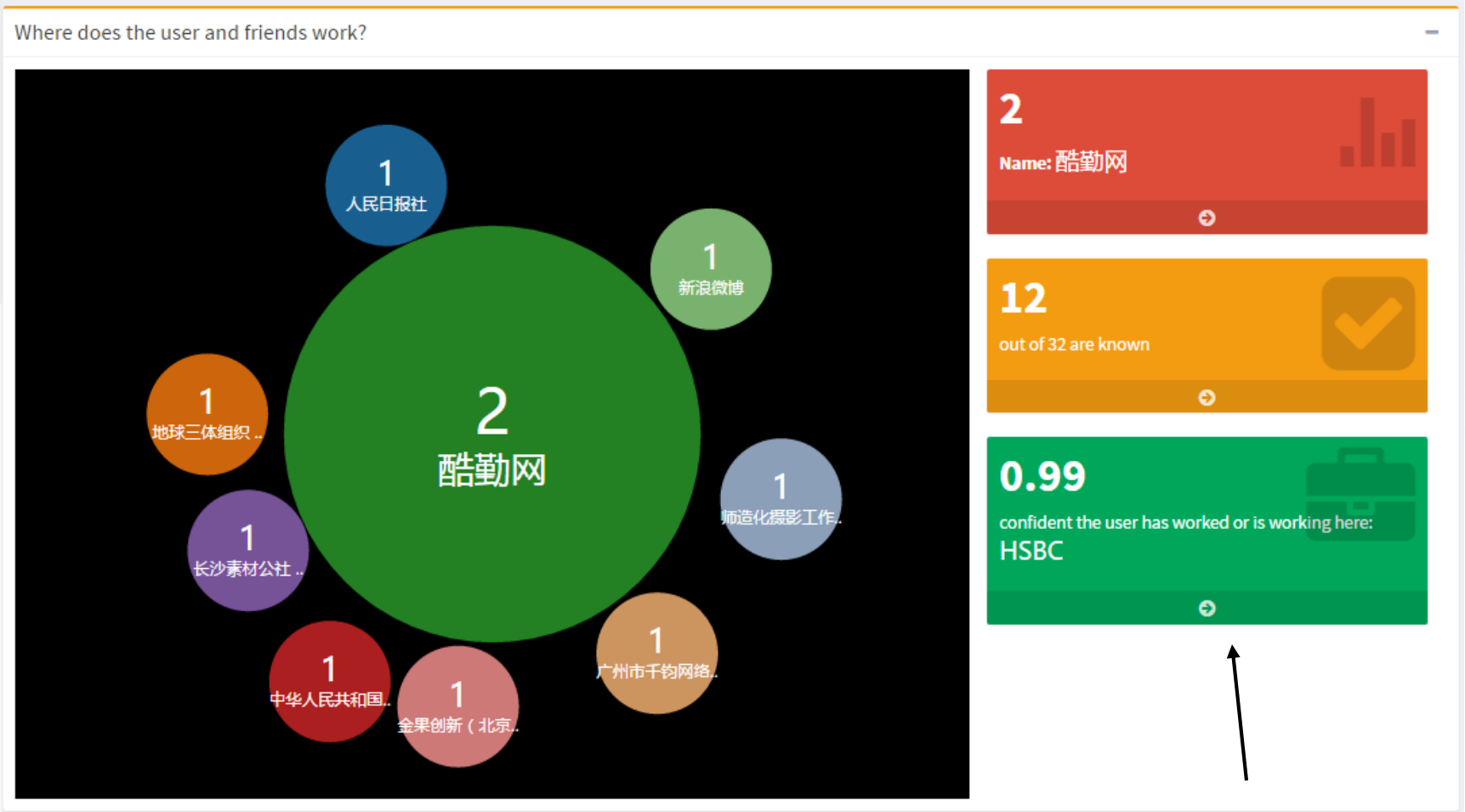
friends are from **Beijing** province, accounting for the largest portion among all places.

0.99

confident that the user is from Beijing province.

"I registered with wrong geographical information to protect myself.
Plus I have many friends in Beijing."

Basic Information Study



Same report/predict flow on user's work/education experience.

User Interest Study

"*"My friends on Weibo.com are mainly university peer students. And we are all CS students.*"

- User's interests can also be inferred from his/her friends' interests.

User Interest Study

- Keywords are displayed in a word cloud.

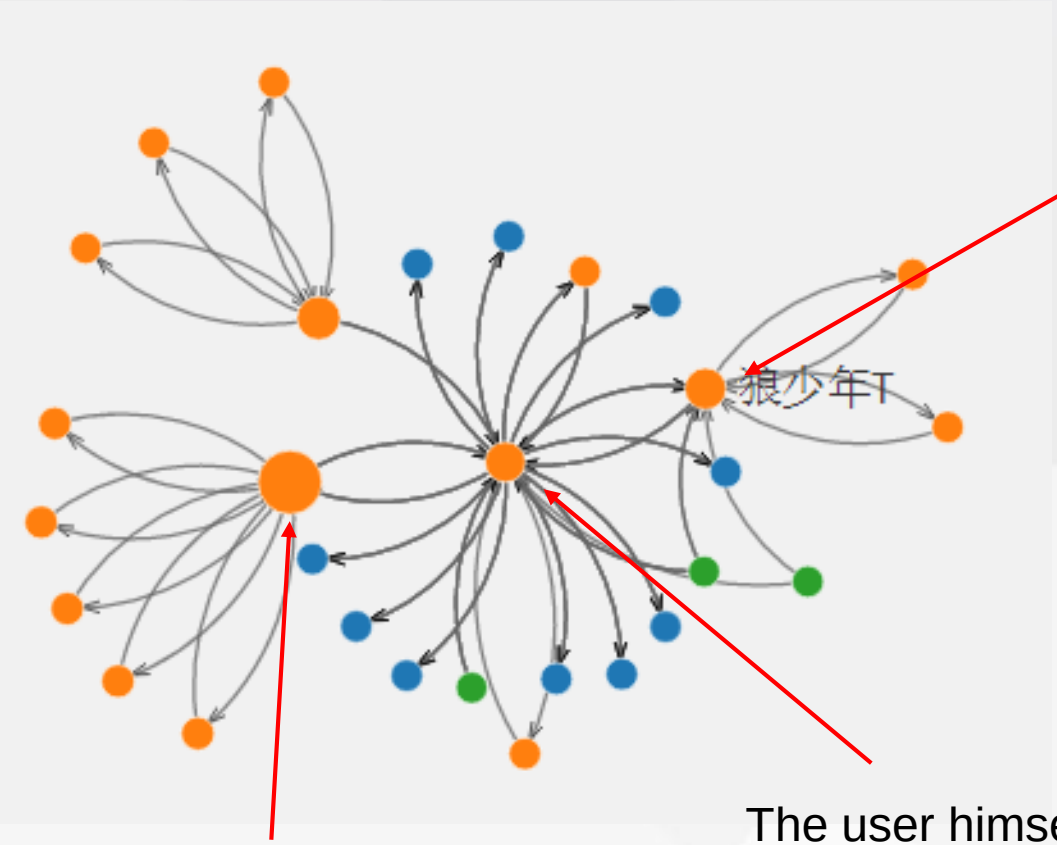
"I repost a lot from official accounts on music and programmer weeklies."

Music

Develop-ment, Program-ming

Comics and Animation

User Interactivity Study



"This is my best friend, peer student and co-worker."

The user himself

"This is my best friend in childhood."

- The users that have interacted with the target form a directed network.
- Only the important nodes are displayed.
- Nodes are displayed in different colors to distinguish.

User Interactivity Study

They have commented user's weibo

User	Count	Ratio
诺亚的大船	24	9.95%
狼少年T	8	3.31%
棟瞎紫	7	2.9%
-我要成为神-	3	1.24%
hackeryoung	3	1.24%
公馆教皇撒加1984	1	0.41%

The target himself

Best friend, co-worker

Girl friend

Roommate

Peer student, friend

They have reposted user's weibo

User	Count	Ratio
FOREVER全力以赴	1	0.41%
Rinka喵	1	0.41%
诺亚的大船	1	0.41%

Childhood best friend

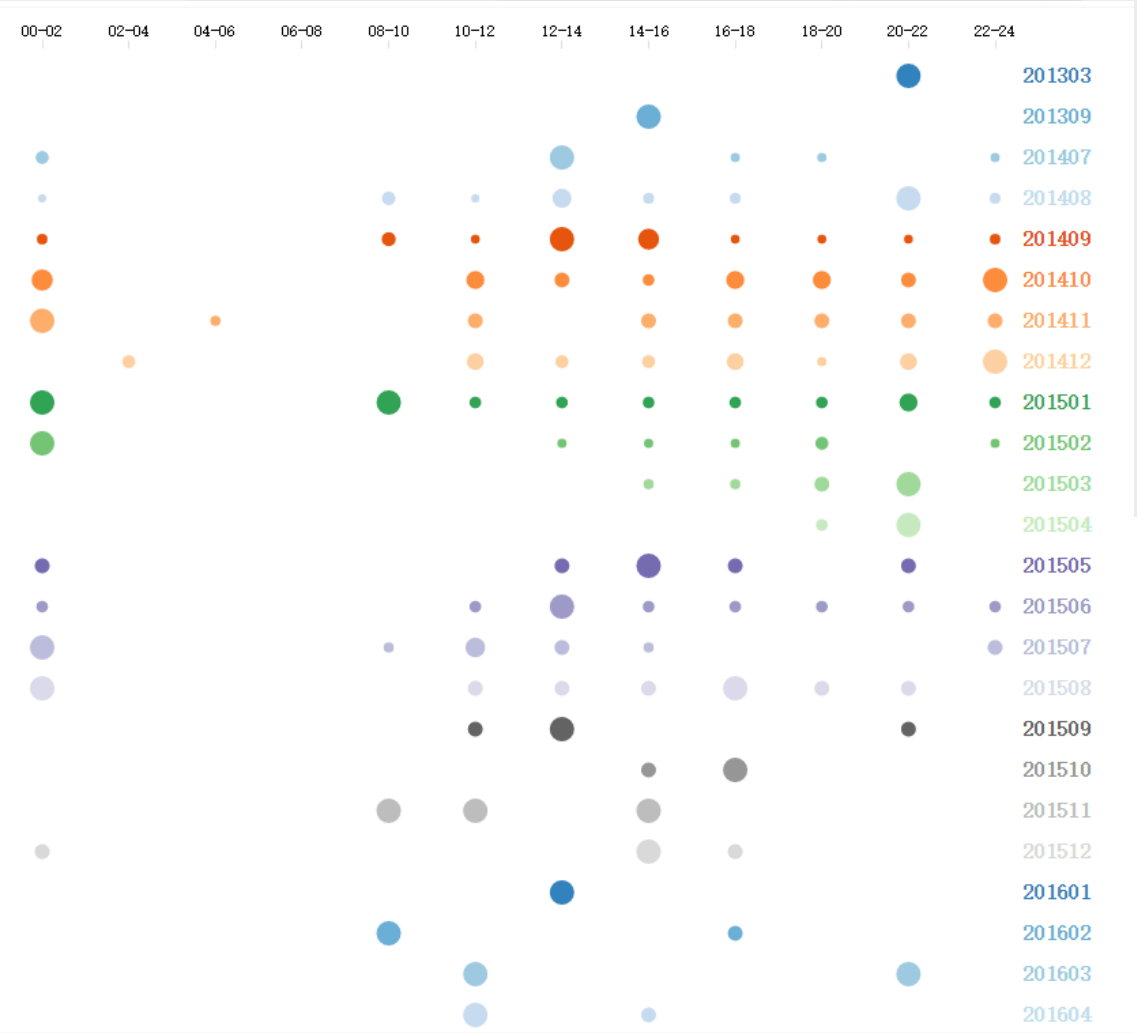
They have liked user's weibo

User	Count	Ratio
FOREVER全力以赴	4	1.65%
棟瞎紫	1	0.41%
hackeryoung	1	0.41%
狼少年T	1	0.41%
完蛋了的完蛋	1	0.41%

- The users who have most interacted with the target user are sorted and listed.

"”Those with top interaction are my the people that are closest to me.”

User Life Style Study



- The post time of each microblog is mapped to a 24-hour table.
- Predict life schedule.

Predictions on the user's daily schedule

- The user is active in these time slots:
  - 0:00~2:00
  - 12:00~14:00
  - 20:00~22:00
- The user is inactive in these time slots:
  - 2:00~4:00
  - 4:00~6:00
  - 8:00~10:00
- This is the predicted bedtime based on the activity of user:
  - 2:00~10:00

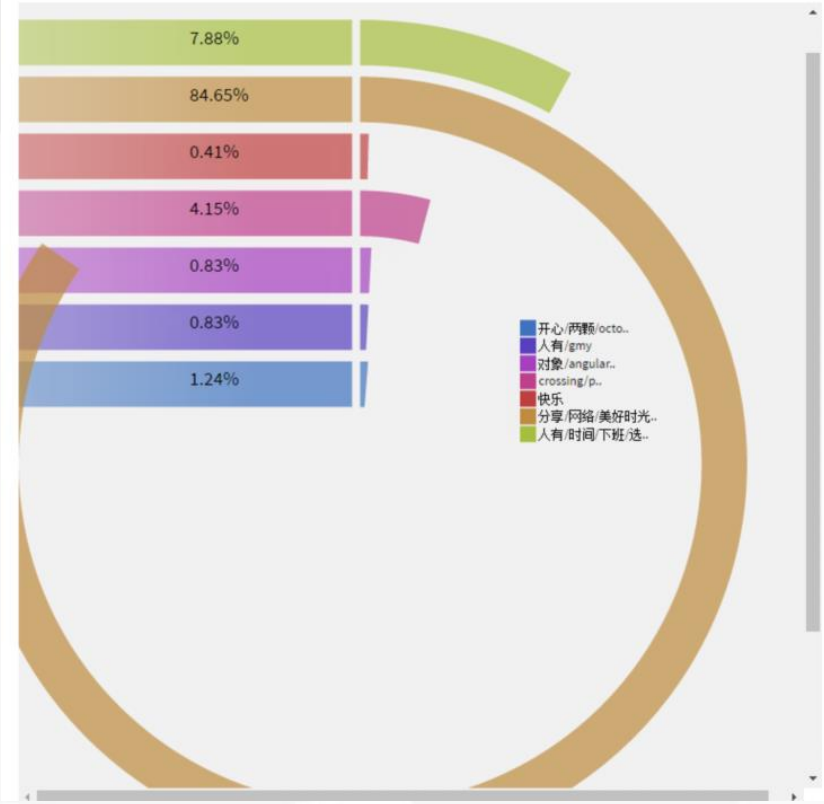
"The predictions are mostly correct . Better if it's finer grinded in 24 hours."

Topic Study

Reposts, music, moods, “chicken soup of the soul”.

K-means++:

- Clustering microblogs by content



“Yes it’s okay to perceive my microblogs as more of reposts about music or moods. But the correlation between the tags or keywords is not significant.”

Topic Study

Topic Modeling (Latent Dirichlet Allocation):

- Assume each piece of microblog is composed of words from multiple topics.
- Extract semantic clusters from the contents.

Topics

Index	Keywords
1	来自 架构 css 代码 框架 html 程序员 位置 选择 网站 香港 程序 人生 简单
2	公路 中国 推荐 大学 互联网 转发 学习 微博 流星雨 前端 浪漫 这是 值得
3	cry 父母 一刻 轻松 手机 世界 思考 哈哈哈 good 网友
4	分享 纯音乐 听蛙 喜欢 一首 music 音乐 hitomi 一生 文章 蜡烛
5	时间 生活 世界 努力 习惯 生命 早安 晚安 教授 九句

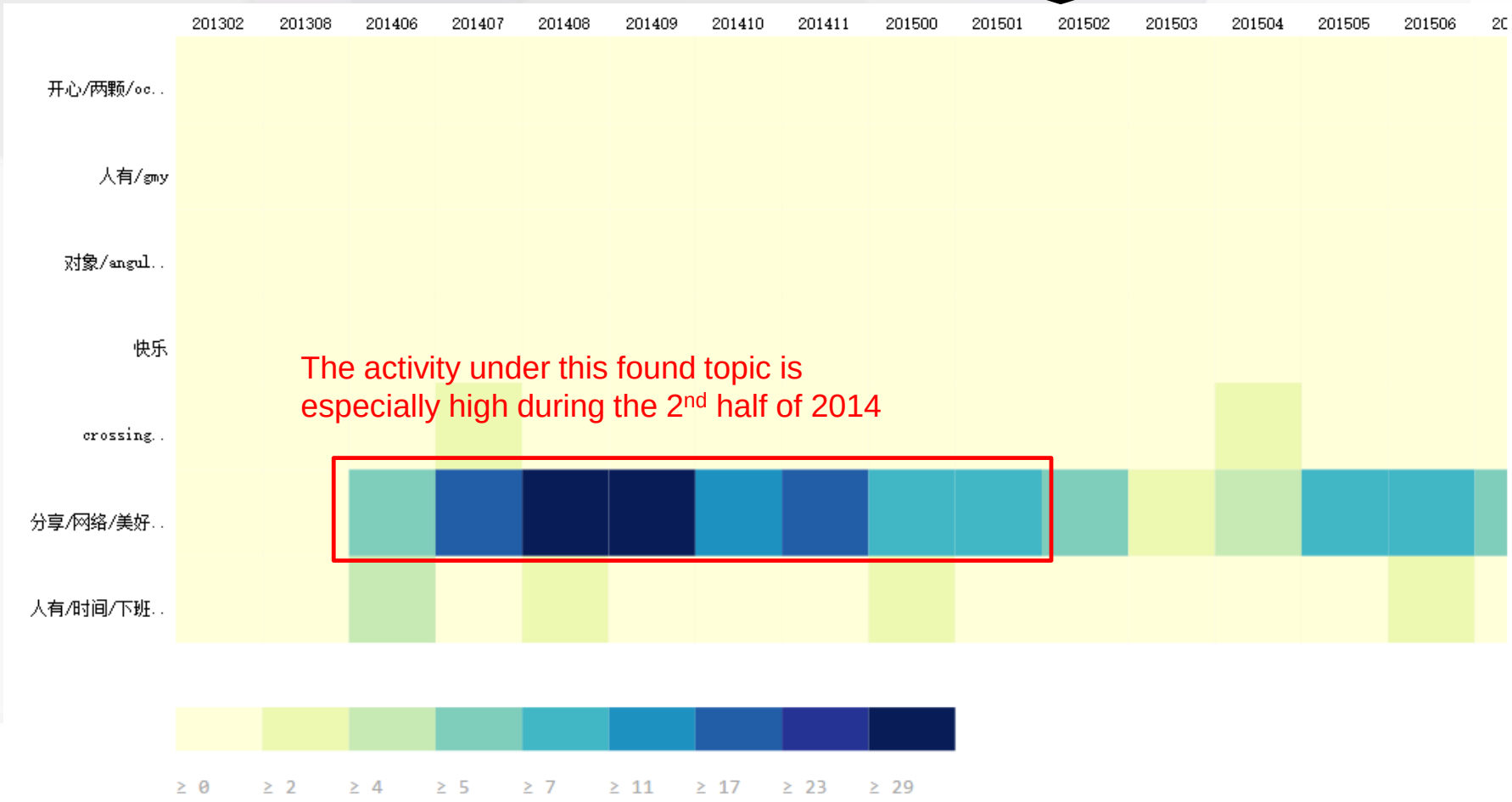
Programming

Music

"The semantic clusters are more obvious and indicative."

Timeline Study

"Yeah…I didn't have much to do at work…And I mostly interacted with girl friend so I was active in that period."



Algorithms Applied

Algorithms Applied

Network

PageRank

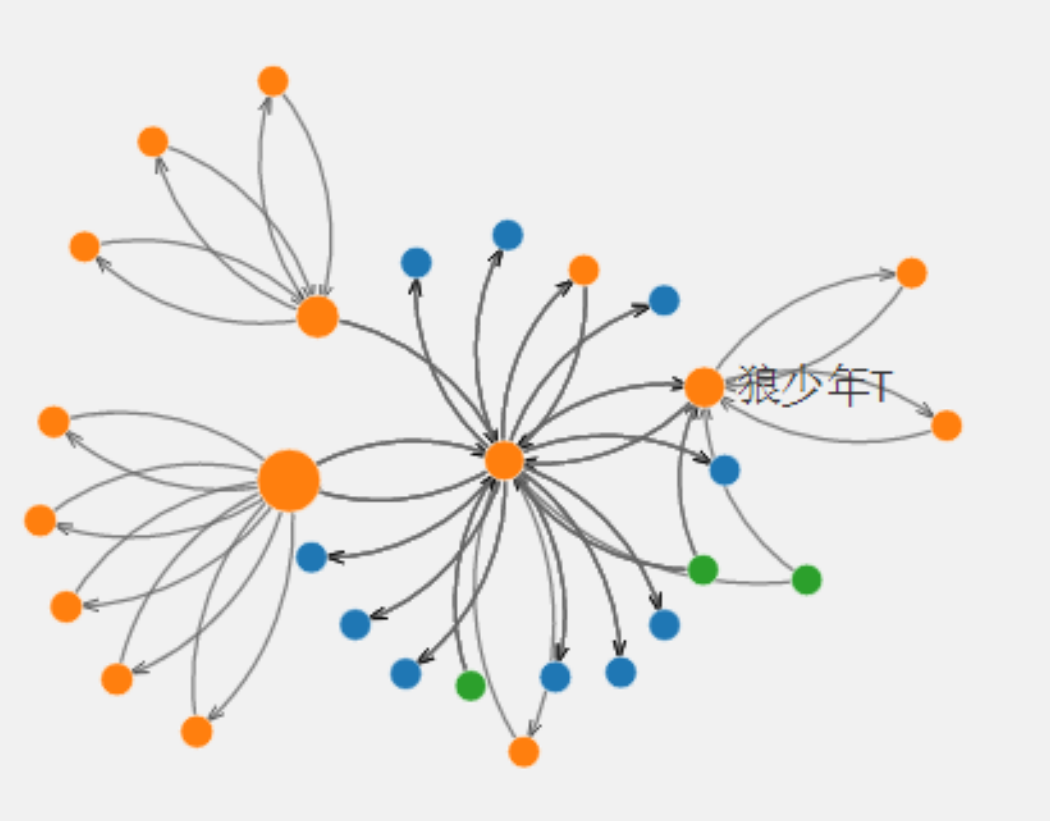
Classification

Logistic Regression, Random Forest, Gradient Boosting Trees (GBT), etc.

Clustering

K-means++, Latent Dirichlet Allocation (LDA)

Graph Definition



Node:

- User

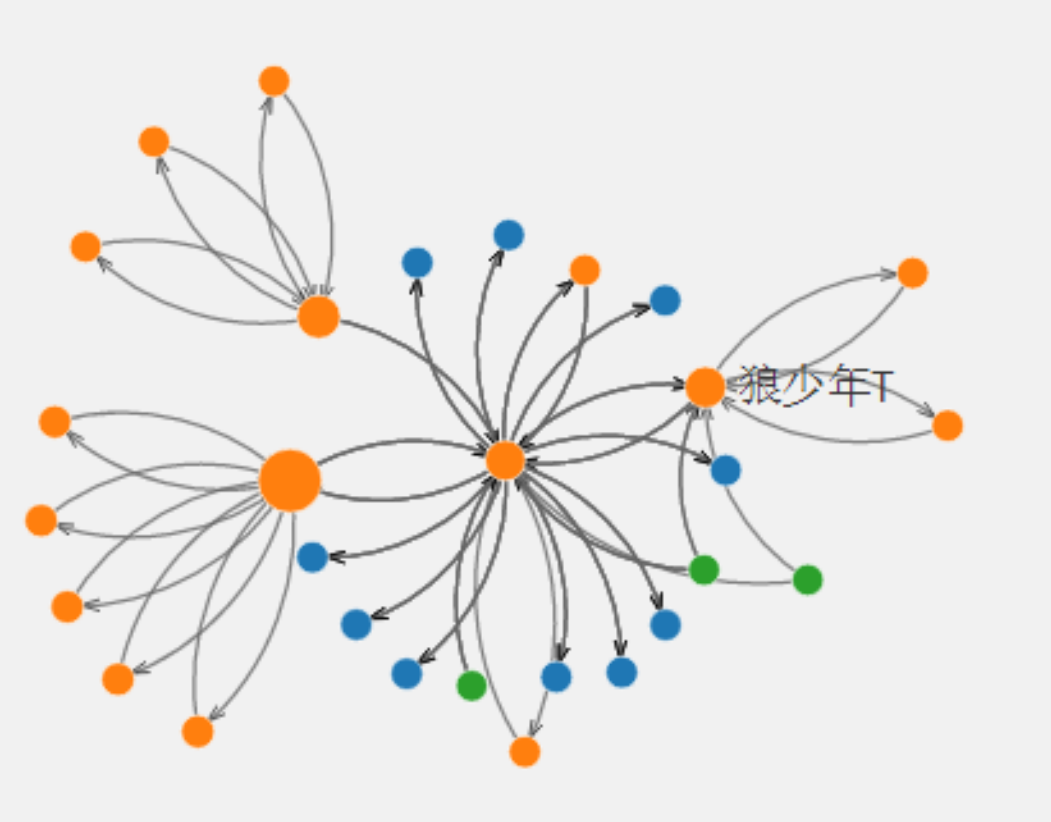
Link:

- There is a directed link from user A to B if A has commented / reposted / liked / mentioned B.

Weight:

- The count of interaction

Characteristic of Weibo.com



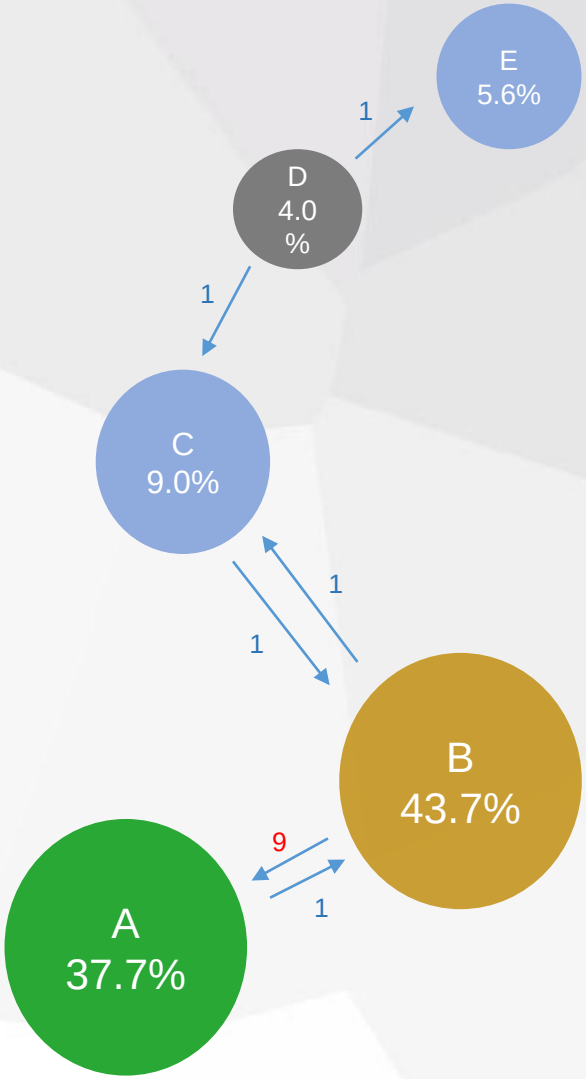
Celebrity Effect [1]

The correlation between inlinks and outlinks is low.

Base Assumption:

Outlinks outweigh inlinks.

PageRank



Step 1: Construct transition matrix [2,3]

$$H=\begin{array}{c} \\ i \\ \\ \\ \\ \end{array}\begin{array}{c} \begin{array}{ccccc} j & & & & \end{array} \\ \begin{bmatrix} 0 & 1 & 0 & 0 & 0 \\ 0.9 & 0 & 0.1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0.5 & 0 & 0.5 \\ 0.2 & 0.2 & 0.2 & 0.2 & 0.2 \end{bmatrix} \end{array}$$

Step 2: Initialize by evenly distribute weight to all the nodes:

$$\pi = \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{pmatrix}$$

π_A denotes importance of node A

Step 3: "Google Matrix" [4]

$$G = \theta H + (1-\theta)\frac{1}{N}11$$

θ is the dampening factor.

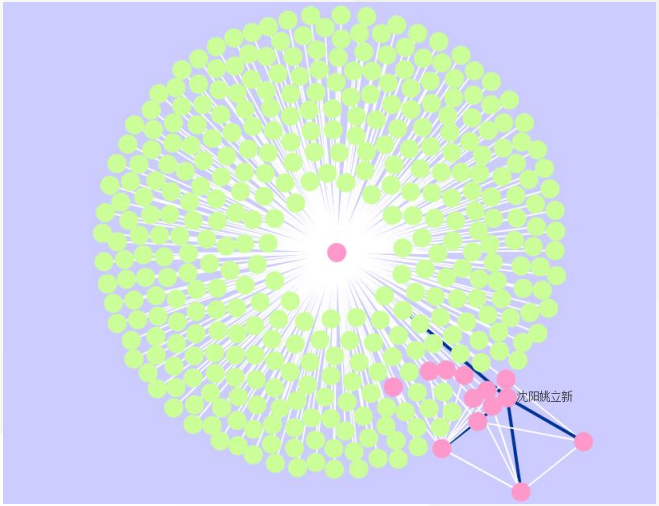
Step 4: Iteratively distribute the importance, until it converges:

$$\pi^T[k] = \pi^T[k-1]G$$

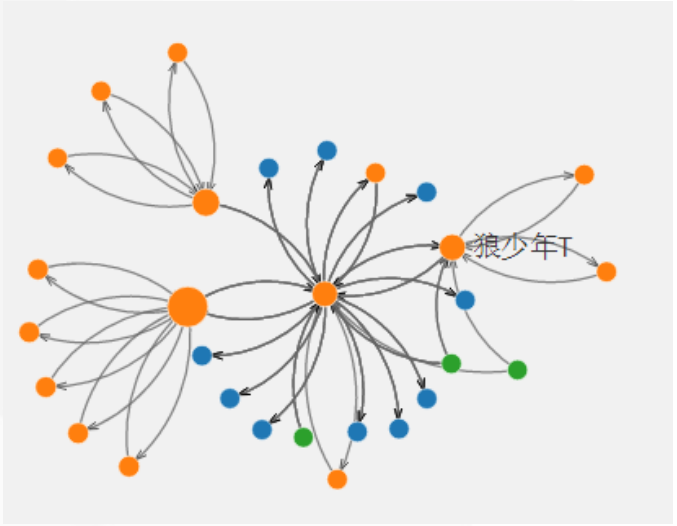
Benefits

1. Meaningful result

2. Better visualization.



Filter unimportant nodes



PART 4

Conclusion

The Value

Data Mining:

1. Models and algorithms
2. Automated process
3. Providing prior knowledge about the user

Data Visualization:

1. Raw data are extracted and structured.
2. Inspiration from graphs

Software Engineering:

1. Integration of frameworks
2. Strengths of languages
3. Flexibility

Reference

1. Fan, P., Li, P., Jiang, Z., Li, W., Wang, H. Measurement and analysis of topology and information propagation on Sina-Microblog, *Intelligence and Security Informatics (ISI), 2011 IEEE International Conference*
2. Page, L., Brin, S., Motwani, R. and Winograd, T. The PageRank citation ranking: bringing order to the web. 1999
3. Csendes, T., Antal, E. PageRank Based Network Algorithms for Weighted Graphs with Applications to Wine Tasting and Scientometrics, *Proceedings of the 8th International Conference on Applied Informatics, Vol. 2. pp. 209–216.* 2010
4. Chiang, M. Networked Life 20 Questions And Answers. *Cambridge University Press; 1 edition*, Sep. 2012, p. 71.

Q&A