WP4-Big Data for Global System Science

An Zeng

Alibaba business school

Hangzhou normal university

The third meeting of the Growthcom project, Como, Italy
September 28th-29th 2014

Presentation outline

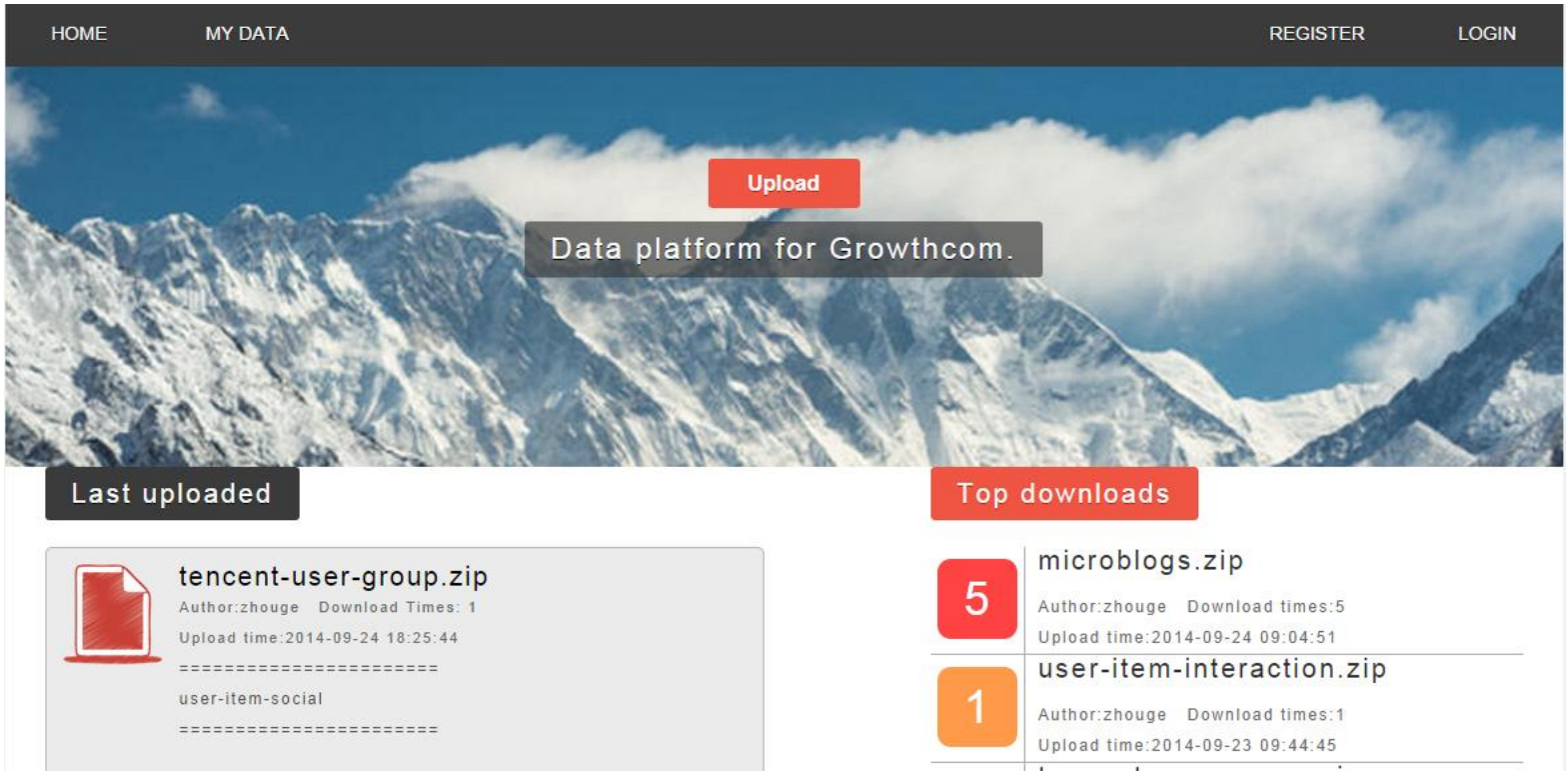
- 1. Hangzhou team
- 2. Data platform for Growthcom
- 3. Data from Alibaba
- 4. Data from other sources
- 5. Connections to Growthcom
- 6. Future work

Hangzhou team

- Prof. Zi-Ke Zhang (also doing a postdoc in Alibaba)
- Prof. Linyuan Lu
- Prof. Ming-Sheng Shang
- Prof. Xiaopu Han
- Prof. Runran Liu (also doing a postdoc in Alibaba)
- Dr. Peng Wang (also doing a postdoc in Alibaba)
- Dr. Chuang Liu
- Dr. Chunxiao Jia
- Dr. Yangxue Xiang

Data platform for Growthcom

- Address: http://suanzi.cn:8083//welcome/index



Data platform for Growthcom

- To download and upload data

1, Register.

2, Login.

3, Apply download and upload permission.

4, Download: the download button will only appear on the data page if you have obtained the download permission.

5. Upload: after obtaining the upload permission, you can upload data from your homepage.

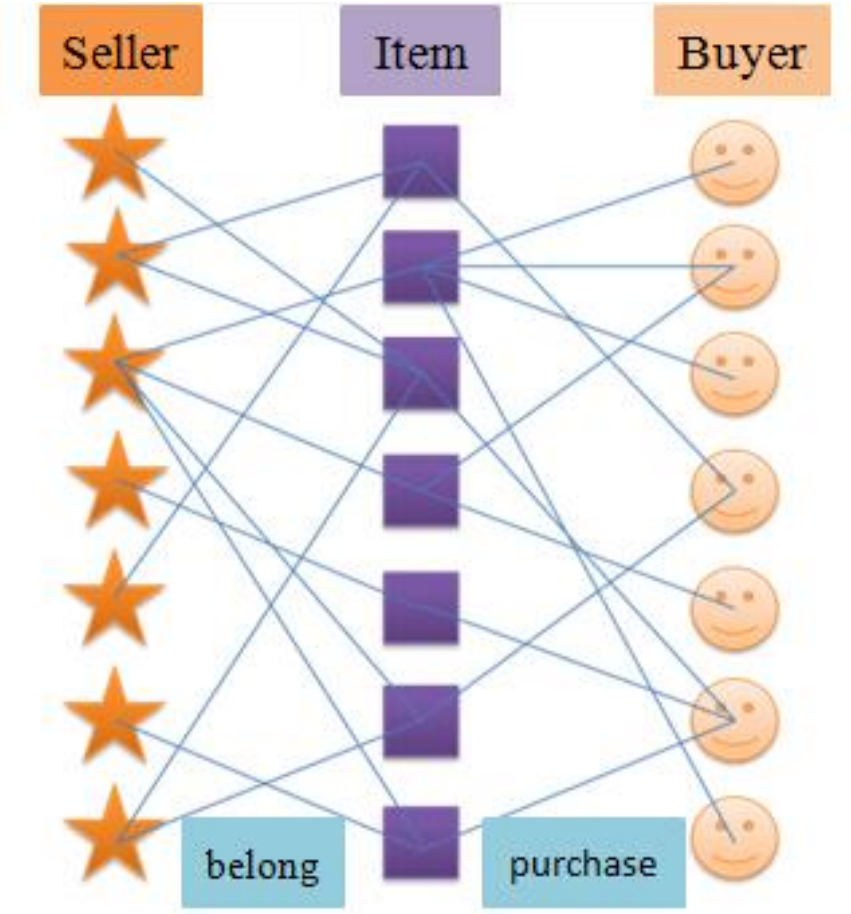
- Data already uploaded to the platform

About Alibaba: customer-product.zip, user-item-interaction.zip, shop-keyword_data_timestamp.zip, shop-keyword-time.zip, shop-category.zip

From other source:tencent-user-group.zip, douban-user-item-social.zip, microblogs.zip

Data from Alibaba

Structure of Taobao.com



Data from Alibaba

- The customer-product data (customer-product.zip) is the purchase records of users on *Taobao.com*. This data was provided by the Alibaba company for research purpose (originally for a modeling contest organized by Alibaba). In each record, it contains the user ID, items ID, and the number of the items in this transaction. Totally, there are 427,667 users, 112,812 items and 30,000,373 records in this data. This data has no time information.

- The user-item-interaction data (user-item-interaction.zip) is a more detailed interaction records (including both browse and purchase) between users and items on *Taobao.com*. This data was crawled via open API for research purposes. In each record of the data, it contains the user ID, item ID, action type (0=browse and 1=purchase). The data has no direct time information. However, the records are sorted in time order (from old to new records). In total, there are 477,138 user, 580,544 items, and 9,269,384 records.

Data from Alibaba

- Keyword data from the biggest Chinese online shopping website *taobao.com* were crawled via open API for research purposes. In the Taobao e-commerce platform, vendors can use keywords to describe their products and well-chosen keywords can contribute to their products being ranked at the top of customers' search results. At the same time, vendors have to pay a price for using keywords and the price of a keyword depends on the keyword's popularity vendors thus have an incentive to invent new keywords or early adopt already existing keywords.

- The data comprise 2,824,853 links between 1,523 online retailers and 915,271 keywords that they attached to their products. Time span of the data is from 12 November 2009 to 21 June 2014 (40,360 hours in total).

- shop-keyword_data_timestamp.zip (In unix timestamp)
- shop-keyword-time.zip (In readable date/time)

Data from Alibaba

- Another data (shop-category.zip) in the platform have already been shared with CNR. It contains online shops, the categories of products it provides to buyers, and the number of deals and income of the shop.

- The Shop-Category data contains:

1. “shop id”,
2. “the first categories id”,
3. “the second categories id”,
4. “the number of deals of the category in the year 2012”,
5. "the number of product sold in the year 2012 (sometimes one deal include several products)",
6. "the income in the categories [unit: yuan] (the price of product multiplied by the number of the product sold. Here, it is the price before discount so this income maybe less than real income)"

Data from other sources

- Tencent data (tencent-user-group.zip) is the user-group records in the Chinese biggest instant messaging software service called *QQ*. This software is similar to MSN, but allow users to set up discussion groups. Each user can join multiple discussion groups. The data contains 500,2340 users, 103,681 groups, 11,165,558 records. There is no time information.

- Microblog data (microblogs.zip) is the post records in the Chinese biggest microblogging website *weibo.com.* It is very similar to Twitter. We crawled the posts data from users in the website between 2012 Feb 4 and 2012 Feb 7. Each post is saved in files with user ID, post ID and the created time. In total, there are 477,138 users, 4 million posts.

Data from other sources

- Douban data (douban-user-item-social.zip) is the user rating records and social network in *douban.com* which is an online community allowing registered users to record information and create content related to film, books, music, and recent events and activities in Chinese cities.

- There are two files: the first one describes the social network between users, the second is the rating records on movies (5-star rating). In total, there are 24,324 users, 9,699 items, 3,811,792 rating records and 74,544 friendship records. The data has no time information.

Connections to Growthcom

- D4.2 Upload of data in the open platform.

Upload the data collected in the first 12 months to the open access platform. The main objective of WP4 is to setup a universal data portal as a whole, including storage, security, processing, which could further support research in other WPs, and the ecology problem in Alibaba as well.

- MS3 Product and company data collected
- MS5 Integration of initial collected data by WP4

Future work

- More data from Alibaba

  Taobao.com, *Tmall.com* (possibly also from *alibaba.com*, *etao.com*),

  Data for interconnected networks,

  Data with time information.

- Some research on the data

  Empirical analysis on the network structure,

  Information spreading patterns,

  Designing a better recommendation algorithm,

  Developing a reputation system.

Publication list of Hangzhou team

- 1. Yu-Xiao Zhu, Xiao-Guang Zhang, Gui-Quan Sun, Ming Tang, Tao Zhou and Zi-Ke Zhang. Influence of reciprocal links in social networks. ***PLoS ONE*** 9(7): e103007.
- 2. Da-Cheng Nie, Zi-Ke Zhang, Jun-Lin Zhou, Yan Fu, Kui Zhang. Chongjing Sun and Yan Fu. Information Filtering on Coupled Social Networks. ***PLoS ONE*** 9(7)(2014): e101675.
- 3. Zi-Ke Zhang, Chu-Xu Zhang, Xiao-Pu Han, Chuang Liu. Emergence of Blind Areas in Information Spreading. ***PloS ONE*** 9(3)(2014) e95785.
- 4. Ning Xi, Zi-Ke Zhang,Yi-Cheng Zhang, Zehui Ge, Li She, Kui Zhang. Cultural Evolution: The Case of Babies' First Names. ***Physica A*** 406 (2014) 139–144
- 5. Ye Sun, Chuang Liu, Chu-Xu Zhang and Zi-Ke Zhang. Epidemic Spreading on Weighted Complex Networks. ***Physics Letters A*** 378 (2014) 635–640.
- 6. Zi-Ke Zhang, Lu Yu , Kuan Fang, Zhi-Qiang You, Chuang Liu, Hao Liu, Xiao-Yong Yan. Website-Oriented Recommendation Based on Heat Spreading and Tag-Aware Collaborative Filtering. ***Physica A*** 399 (2014) 82-88.
- 7. Zi-Ke Zhang, Ye Sun, Chu-Xu Zhang ,Kuan Fang, Xiang Xu, Chuang Liu, Xueqi Wang, Kui Zhang. Diagnosing and Predicting the Earth's Health via Ecological Network Analysis. ***Discrete Dynamics in Nature and Society*** (2013) 741318

Thank you!