Generating ordered list of Recommended Items: a Hybrid Recommender System of Microblog*

Yingzhen Li, Ye Zhang,
School of Mathematics & Computational Science, Sun Yat-Sen University

SUN YAT-SEN UNIVERSITY

*solution of Track 1 task, KDD Cup 2012

BACKGROUND

Observing the rise of twitter services' popularity in 2007, Tencent(www.qq.com), one of China's leading Internet service portal, launched microblog(China's Twitter) in 2010. Based on the large user group of its instant messaging service QQ(711.7 million[1]), Tencent Microblog has attracted large amount of registered users (425 million and 67 million daily active[2]) and became one of the dominant microblog platforms. Tencent invites celebrities and organizations to register and interact with users directly. Users can enjoy fun of Microblog directly on the website of Tencent Microblog or via the third-party port and related platforms. The service is embedded in Tencent's other leading platforms like QQ signature, Qzone(blog platform), Quan(SNS service) and Weixin(mobile messenger).

Logo of Tencen Microblog

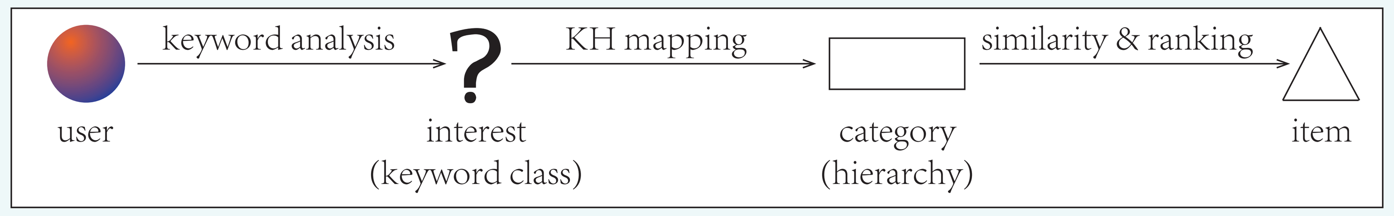
PROBLEM

While Tencent has the biggest microblog user groups, Sina Microblog took a commanding lead with 56.5% of China's microblogging market based on active users, and 86.6% based on browsing time over its competitors[3]. The existance of fake user group, widely used spammer strategy [4], and weird definition of active users considering those who write(including retweets and comments) or read microblog messages - no matter directly on the website or via third-party port or associated platforms - as acitve users contribute to the fake prosperity of Tencent Microblog which is far from the public perception. Another problem of Tencent Microblog is the low percentage of accepted item recommendation, less than 9% according to our sampling[5]. The item list doesn't update in time, and the recommendation often deviates from the preference of users.

Tencent Microblog(t.qq.com)

SOLUTION: HYBRID RECOMMENDER SYSTEM

Recommender systems can be categorized into contentbased algorithm[6], collaborative filtering[7], and influential ranking algorithm[8]. However each single algorithm has its unavoidable disabilities. Hence we design a hybrid approach considering user preference variance and similar interests among linked users, including keyword analysis, user taxonomy and generation of ordered item list.



The grade of recommending item i_k(belongs to category h_k) to user u_j(specified to its user class) is computed by:

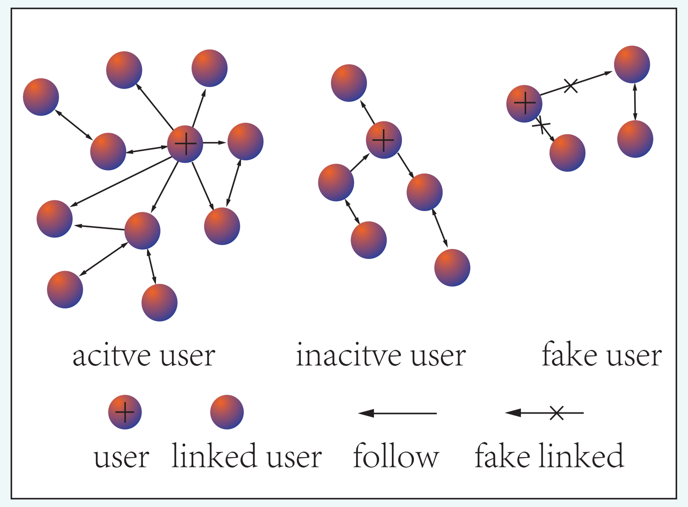
(active/inactive user) $\mathrm{grade}(u_j, i_k) = 2\,\mathrm{fond}(u_j, h_k)(\alpha\, \mathrm{hot}_k + (1-\alpha)\,\mathrm{sim}(u_j, i_k)) - 1$

(fake user) $\mathrm{grade}(u_j, i_k) = (1 + \mathrm{fond}(u_j, h_k))\,\mathrm{HOT}_k - 1$

where α is trained in training process and identical to u_j. Finally we find out the top-3 recommended items of the user.

Keyword Analysis

Noticing the existence of synonyms, we group the keywords into classes to extract user's(and item's) interests. But mining keyword classes directly in the huge user-keywordset is unrealistic, so we paralell the candidate generation by adopting and revising FDM[9]. Evidently the choice of (local/global)support and confidence affect the precision and complexity tremendously. We sampled 1000 users' keywords and found out that these users have their keyword weights average in 0.14, so we assign $\mathrm{supp\_local} = \mathrm{supp\_global} = 0.2$ and $\mathrm{conf\_local} = \mathrm{conf\_global} = 0.7$. Also we notice the ambiguity of keywords, 'apple' for instance, hence we insert these ambiguous keywords into different classes simultaneously.

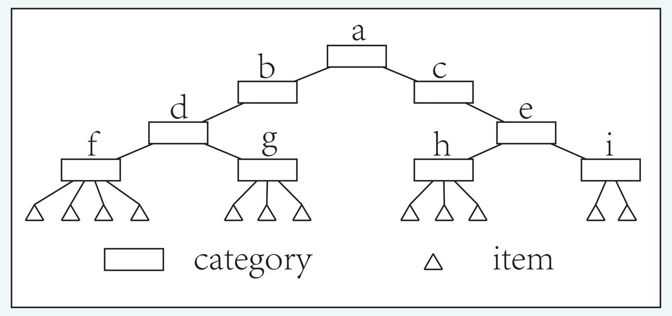


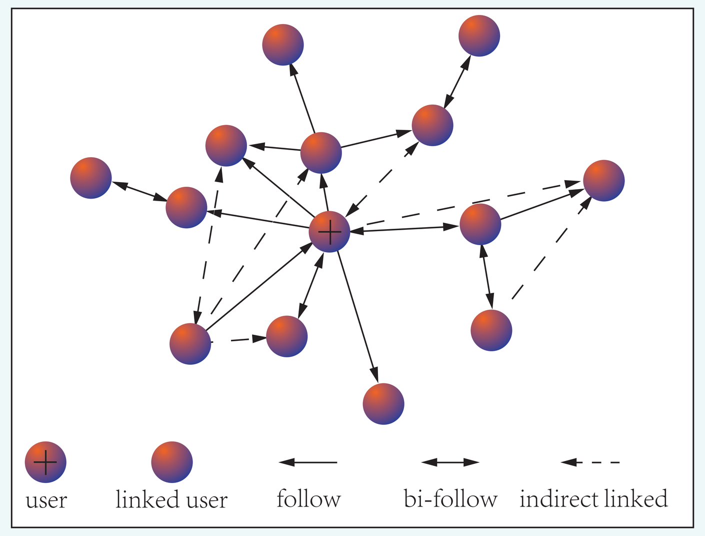
User Taxonomy

According to the number of tweets(due to the lost of login data) and interactions with others we group the users into 3 excluded groups - active, inactive and fake - to apply different types of strategies.In fact, lot of Tencent microblog users actually seldom login, and the messages generate from the third-party portor associated platforms are synchronized to their microblog, generating the fake illusion of their activeness. In addition we also classify the spammers as fake users. With statistics we conclude that only 33.2% of the users have written more then 100 tweets. We choose min_tweet = 100, min_interaction = 20, and separate the users into 3 groups.

Item Ranking(Computing hot_k and HOT_k)

An item is a specific user, which can be a famous person, an organization, or a group. Items are organized in different categories by Tencent according to their professional domains, which forms a hierarchy. Obviously the number of an item's followees reflect its popularity directly. We adopt that indicator and rank the items in categories (hot_k). Recommending high-ranked items in user's interested field promotes the possibility of acceptance effectively. We also recommend most popular items(indicator HOT_k) on the platform to those who show little of their preferences, especiallythe fake users.





Computing Similarity(u_j,i_k) and fond(u_j, h_k)

Recommending items with high similarity in preference is considerably effective to increase the percentage of acceptance. We extract the interests of users from their keyword classes. Noticing that few users actively write tweets thus not enough keywords, we design indirect collaborative filter to mine the potential interests of inactive users from their followees $\{u_n\}$and apply $\mathrm{fami}(u_j, u_n)$ to represent the familiarity between two users u_j and u_n to adjust the weights of potential interests. Then we define KH mapping which maps the interests to the hierarchy of items' professional domains, compute $\mathrm{fond}(u_j, h_k)$ which indicate user's preference of the category and obtain u_j's candidate items to compute their similarity $\mathrm{sim}(u_j, i_k)$.

EXPERIMENT & RESULTS

We sampled 5938 users for experiment. All the data is encrypted by Tencent. After training we noticed the variance of α among different user classes:

user class	user	followee	interaction	keyword	α
active	3919	46	87	10	0.33
inactive	1194	27	42	8	0.18
fake	825	18	2	5	/

Then we introduce the prediction evaluator[10] :

$$\mathrm{AP}(u_j) = \sum_{i=1}^{3} p(i)\,\Delta r(i)$$

where $p(i)$ is the precision of the i^{th} recommended item and $\Delta r(i)$ is the change in the recall from $i-1$ to i, and present the experiment's result and example prediction:

AP:

active	inactive	fake	total
0.41066	0.46879	0.33606	0.41198

u_j	user class	item	accepted item	AP (u_j)
2071402	active	1606902 1760350 1774452	1606902 1774452	0.83
942226	inactive	1606902 1606609 1774452	1606902	1.00
193889	fake	1760642 1774684 1774862	1774862	0.33

The result showed the high performance of our Hybrid Recommender System after training. To save time we can divide users into smaller classes and train the machine respect to each class.

REFERENCE

[1] Tencent. About tencent. http://www.tencent.com/enus/at/abouttencent.shtml, 2012.
[2] Tencent. Tencent announces 2012 first quarter results. http://www.tencent.com/enus/content/ir/news/2012/attachments/20120516.pdf, May 2012.
[3] Kyle. Sina commands 56% of china's microblog market. http://www.resonancechina.com/2011/03/30/sinacommands-56-of-chinas-microblog-market/, March, 2011.
[4] Baidu. Zombie fans on weibo. http://baike.baidu.com/view/4047998.htm, 2010.
[5] Y. Niu et al. The Tencent Dataset and KDD-Cup'12. KDD-Cup Workshop, 2012.
[6] M. McPherson et al. Birds of a feather: Homophily in social networks. Annual Review of Sociology, 2001.
[7] J. A. Konstan, et al. Grouplens: applying collaborative filtering to usenet news. Commun. ACM, 1997.
[8] Z. Wang, Y. Tan, and M. Zhang. Graph-based recommendation on social networks. In Web Conference (APWEB), 2010 12th International Asia-Pacific, 2010.
[9] D. Cheung, J. Han, V. Ng, A. Fu, and Y. Fu. A fastdistributed algorithm for mining association rules. In Parallel and Distributed Information Systems, 1996., Fourth International Conference on, dec 1996.
[10] M. Zhu. Recall, precision and average precision. Working Paper 2004-09, Department of Statistics & Actuarial Science, University of Waterloo, 2004.

ABOUT US

Yingzhen Li
Undergraduate student(Junior Year),
School of Mathematics & Computational Science,
Sun Yat-Sen University,
Guangzhou, China
Searching for PhD/Master programs in the U.S., 2013 Fall
E-mail: liyzhen2@mail2.sysu.edu.cn

Ye Zhang
Undergraduate student(Junior Year),
School of Mathematics & Computational Science,
Sun Yat-Sen University,
Guangzhou, China
Starting gruaduate study in SYSU, 2013 Fall.
E-mail: zhangye5@mail.sysu.edu.cn