Comprehensive Graph and Content Feature Based User Profiling

Peihao Tong[1], Junjie Yao[1(✉)], Liping Wang[1], and Shiyu Yang[2]

[1] School of Computer Science and Software Engineering,
East China Normal University, Shanghai, China
10122510148@student.ecnu.edu.cn, {junjie.yao,lipingwang}@sei.ecnu.edu.cn
[2] School of Computer Science and Engineering,
The University of New South Wales, Sydney, Australia
yangs@cse.unsw.edu.au

Abstract. Nowadays, users post a lot of their ordinary life records to online social sites. Rich social content covers discussion, interaction and communication activities etc. The social data provides insights into users' interest, preference and communication aspects. An interesting problem is how to profile users' occupation, i.e., professional categories. It has great values for users' recommendation and personalized delivery services. However, it is very challenging, compared to gender or age prediction, due to the multiple categories and complex scenarios.

This paper takes a new perspective to tackle the occupation prediction. We propose novel methods to transfer the commonly used social network feature and textual content feature into vector space representation. Specifically, we use the embedding method to transfer the social network feature into a low dimensional space. We then propose an integrated framework that combines the graph and content feature for the occupation classification problem. Empirical study on a large real social dataset verifies the effectiveness and usefulness of the proposed approach.

Keywords: User profiling · Graph embedding · Prediction model

1 Introduction

With the stunning growth of the social media fever, the amount of user contributed data is growing exponentially. Nowadays, ordinary users usually spend much more time surfing on the social media applications. Take Facebook, Twitter, and Weibo as examples, Facebook has more than 1.65 billion monthly active users every day; Twitter has around 500 millions registered users and generates several hundreds of million messages every day; Weibo from Sina has around 214 million registered users and 93 million active users every day[1]. Lots of users' life

[1] http://facebook.com
http://twitter.com
http://weibo.com
http://www.sootoo.com/content/654707.shtml.

M.A. Cheema et al. (Eds.): ADC 2016, LNCS 9877, pp. 31–42, 2016.
DOI: 10.1007/978-3-319-46922-5_3

record is collected and stored. It is becoming valuable to profile users' interests and preferences from social data [6,16]. The user profile mining has been a hot topic in recent years.

An important problem in user recommendation and content delivery is the occupation prediction [7]. Occupation describes users' professional area and work interest. However, occupation prediction is very challenging due to the lacking of enough data collection. The social media provides a great opportunity to dive into users' life record to extract occupation indicants. The occupation prediction from social data can be modeled as a classification problem of social features. Possible features include users' posted messages, his/her social networks and other kinds of activity features. The large social network data and the heterogeneous features inside social data domain hinder the effective processing [3,13,16]. Though there has already exist several attempts towards occupation inference from social data [4,7,13], the performance is still not comparable with other user attribute prediction.

This paper focuses on effective methods to process the social features for a better prediction performance. We propose a new framework that combines the representative content and graph features in order to analyze social data. For the content information, we propose a text classification approach. For the graph features, we resort to effective embedding way to process these features properly. Recently, many efficient techniques for analyzing the real-world network were proposed. One of the techniques is graph embedding [15]. Graph embedding is a representation learning process, aiming to transform the large-scale network to low dimensionality representation. It makes the network data much simpler and easier to handle. The most fantastic part is that, representation learning can combine different kinds of features in a unified space, enabling machine learning techniques more general and easier to tune. This paper exploits the graph embedding power and the corresponding unified occupation prediction approach.

This paper chooses a unique dataset from Sina Weibo, one of the Chinese largest social media platform. It has a special feature that some verified users' occupation information has already been properly annotated with editors. It alleviates our annotation cost and provides accurate ground truth for the prediction algorithm training. Weibo dataset has posts, social networks and users' other kind of information. In the preprocessing stage, we extract a network from the Weibo dataset via bi-follower relationship.

The proposed approach's framework is shown in Fig. 1. We use the graph embedding method to process the bi-follower network and deliver a low dimensional latent representation for every user. We concentrate on the messages' content of the users and merge the all the messages posted by each user as his/her content representation. On top of the new representation layer, we utilize the classification methods to process the graph feature and content feature respectively. We then merge the individual prediction results in a novel fusion way to finally deliver user's occupation estimation. We would illustrate the processing details of this unified occupation prediction in the later parts of this paper.

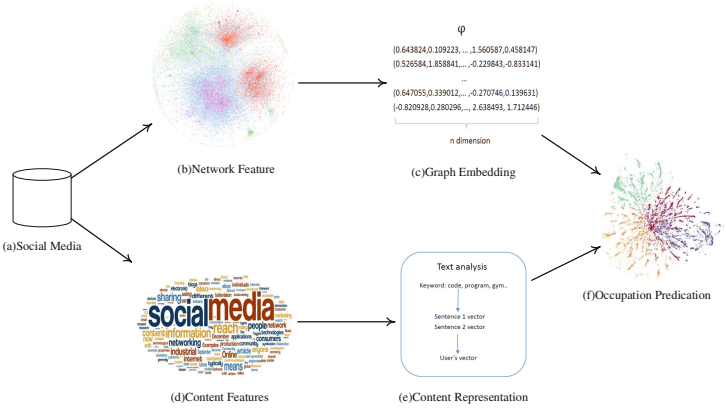


Fig. 1. The unified occupation prediction approach

Our main contributions of this paper are listed as follow:

1. First, we systematically analyze the feature representation of users and dive into the network structure to capture users' latent affiliations representation. This simplifies the feature extraction work and provides better foundation for the prediction task.
2. Second, We propose a multi-source supervised classification framework combined with both content-based and graph-based features. It is very versatile and able to cope with rich indicants of social domain.
3. At last, we conduct extensive experiments and the results clearly show the good performance of the proposed approach. The features are useful, especially the graph ones.

The rest of this paper is organized as follows. In Sect. 2 we discuss related work. Section 3 introduces the problem definition and approach framework. Section 4 illustrates the feature processing steps and the inference model. Section 5 presents the experiments on a real large dataset and finally we conclude this work in Sect. 6.

2 Related Work

The work in this paper is related to several areas. Here we briefly review the corresponding literature.

User Profiling: It is of great importance to a variety of user oriented applications. Lots of features and applications are investigated in recent years [6]. For example, [16] proposes a framework which utilizes the friendship to predict user's interests and friends. [4] takes the Twitter as its dataset and measures user's influence in different aspects. [7] combine the graph and text features to infer users' occupation and improve the performance. Other works also include

the mobility and social network prediction. However, they are either limited in the feature processing or the prediction accuracy. In this work, we focus on the novel low dimensional reduction and unified approach to guarantee the occupation prediction performance.

Graph Representation: Graph data is pervasive and ubiquitous across a lot of domains. Effective graph representation is vital for the processing. There are already many different processing methods, including not limited heterogeneous modeling [13], the joint propagation [16] and the embedding [2,11]. Traditional embedding algorithms usually suffer from the difficulties in processing large-scale network due to the complexity of these algorithms quadratic to the network's vertices [12]. SocioDim [14] is a graph embedding algorithm based on community detection and it generates the representation by the network's eigenvectors. In recent years, many novel graph embedding algorithms are proposed, performing well in handling large-scale network. [11] naturally combines graph modeling with language models. This paper focuses on effective graph representation methods for the social media prediction.

3 Preliminaries

In this section, we define the problem of occupation prediction and introduce the used datasets and the features.

3.1 Problem Definition

Definition 1. *We define graph $G = (V, E, O)$, where V is the set of the vertices in the network and the E is the set of vertices' correlation in the network and $E \in V \times V$. O is the set of the text information in vertex. In this work these text information usually implies vertex's occupation.*

In our work, we use the users' friendship network for analysis. For example, there is an edge between user a and user b if there is a bi-follower between a and b.

Definition 2. *Graph embedding aims to learn every vertex's low-dimensional representation which $V \in R^d$where d is the dimension of the representation and $d << |V|$. In other word, it tries to generate a function ϕ which $\phi(V)$ representing the low-dimensional vector representation where $\phi(V) \in R^d$.*

This paper discusses effective methods which process the network features and produce the latent low dimensional representation. Based on it, this paper then adopts the prediction technique to infer the user's occupation.

Definition 3. *Given a group set that contains I group labels, it can be represented $O = \{o_1, o_2, \ldots, o_I\}$. Every vertex in the network G has a label and the occupation inference is to infer the missing label of o_i. We model the occupation inference as a classification problem.*

3.2 Datasets

This paper uses one of China's largest social media platforms, i.e., Sina Weibo[2] as our dataset. In Sina Weibo, users can post, comment and re-share messages and interact with other users. This paper uses the users' verified occupation information, since it is more credible, acting as the ground truth for the prediction model training and testing.

After the pre-processing, we get the most common top groups. Their portions are shown in the following Table 1. We find that media accounted for the largest proportion, followed by entertainment, real estate and finance.

Table 1. Occupation distribution of verified users on Sina Weibo

Fashion	Government	Finance	Arts	Car	Real Estate	Construction
5.22 %	5.59 %	9.65 %	4.39 %	1.06 %	10.67 %	1.30 %
Education	Media	Service	Entertainment	Press	Sport	
5.18 %	25.82 %	4.05 %	11.32 %	3.54 %	7.51 %	

3.3 Features

Content Feature: Users in Sina Weibo can post the messages that called status. These messages are allowed within 140 words. A user can post any message that he/she likes. However, although users can post anything, users usually post the message that he/she is interested in.

We observe that users are usually interested in the things which are related to them, such as the things related to their occupations. For example, a programmer is more interested in IT filed than others and he is more likely to post a message that contains the words just like JAVA or other words in IT filed. So the status may imply the information that is related to the users' occupation. This paper takes the content of status as a source of occupation prediction features.

Network Feature: There are several networks in Sina Weibo. As we all know, users in Sina Weibo could comment and reweet others' messages. Relying on these relationships we could construct two types of networks. However, sometimes users will reweet or comment a message posted by a stranger, which leads the relationships being weak in practical. So we take the bi-follower relationship as the criterion to construct the network, since is more stable and general. In bi-follower network, the relationship is strong because users usually follow the people they are interested in or know well. The community feature in bi-follower network is more clear and it is easier for us to extract valuable information.

[2] http://weibo.com.

4 Prediction Approach

In this section, we discuss the used features and the unified approach for the occupation inference.

4.1 Content Features

This paper utilizes the text information of the user to infer the user's occupation. It concentrates on the content text and tries to transform the text information to latent representation. However, it's infeasible to embed the text directly because of the complexity of the content. So we resort to handle the content and pick a data structure for the content in order to process the content accurately and efficiently.

A user posts several messages. What we get is the representation of the content of the posted message, so we still need to do some process to transform it into the users' latent representation. In this paper, we choose to merge the representation by aggregating the messages' content representation.

There are many meaningless words in the text just like the preposition, conjunction and so on. Meanwhile, there are also some words that we do not concern about. We extract the key words via the term frequency and inverse document frequency criterion. In this way, similar content can be merged and then we use the feature selection metrics to reduce the representation length for each post.

4.2 Network Features

The objective of graph embedding is to learn a latent representation of a graph, which means transforming the large scale network to a low dimension vector representation. After embedding, the new format will be much easier for handling and we can adopt some machine learning techniques to process the data.

However, graph embedding has a risk that may lead to the loss of the structure's data. So it is important to select a suitable graph embedding algorithm to embed this data. DeepWalk [11] is a graph embedding algorithm which adopts to random walk to capture the structure of the network. Facts prove that random walk has good performance in capturing local structure and can adapt to the changed network easily. Besides, using random walk in the network, the frequency of the vertices appearing in the walks follows the power law distribution, while the word frequency in the natural language follows the same distribution. This discovery leads to adopting the natural language model to analyze graph data.

After generating the random walks, this approach continues to utilize the language model word2vec [9,10] to process the intermediate result. Language model aims to estimate the probability of the specific words' occurrence in the corpus. For Word2Vec, it is usually used to maximize the probability of the specific words sequence's occurrence in the corpus. Word2vec takes the words' corpus as the input and produces the distributed representation which called word embedding as output. There are two common models in word2vec. One is

CBOW and the other is called skip-gram. CBOW utilizes the context to predict a missing word. However, with the words sequence's length becoming longer, facts show that calculating the probability by CBOW model becomes difficult and even unfeasible. In contrast, skip-gram model removes the order constraint and uses a word to predict its context within window size rather than use the context to predict a missing word. This model is proved to be friendly to calculating.

Motivated by the content process, we regards a walk as a sentence and selects skip-gram because of the long length of the walk in the network. Its goal is to maximize the probability, targeting vertex's contexts make a co-occurence.

DeepWalk can be used to generate the latent representation of the vertices. We replace the vertex with the vector representation and use $\phi(V_i)$ represents the vector representation of V_i. In practice, it is a common way to adopt the log-likelihood function to handle the Eq. 1. We can get the new representation as following:

$$\max log(p(\{v_1, v_2, \cdots, v_{i-1}\}/v_i|\phi(v_i))) \tag{1}$$

Calculating the probability directly is still difficult and it is unfeasible in practical experiment. In order to derive the probability with guaranteed performance, we adopt a way called Hierarchical Softmax. It treats the occurrence frequency of the vertex in the walk as the weight of the leaf node in order to construct a Huffman tree for factorizing the conditional probability. After Huffman tree constructed, if we set the v_i as our target vertex and intend to compute the v_i's context v_j's occurrence probability $P(v_j|v_i)$, we should find the path from root to the leaf v_j in the Huffman tree.

In the path, non-leaf nodes should be given a vector representation θ as the undetermined parameter for the sake of the later calculation. And We set the left node as positive class and the right node as negative class. So According to the sigmoid function, we could get the probability of every level and thus we can get the final probability. Then by means of adjusting the vectors, we could maximize the probability and at the same time we get the vector representation.

4.3 Prediction Model

The framework we proposed combines the graph features and content features together. Therefore, there is a key problem that we have to solve is how to fuse the network feature and the text content feature.

There are two main fusion methods which are early fusion and late fusion. Early fusion fuses the features before processing this data while the late fusion concentrates on fusing the processed results of the different features' data. In our work, to explicitly compare each feature's contribution and guarantee the generality, we choose the late fusion method to fuse the graph feature and content feature.

In late fusion, there are still several strategies. The most common strategy is averaging. Averaging takes the results of the different learning machines and accumulates these results to calculate the average of the sum result. However, when a big gap exists among the learning machines' performances, the prediction

performance may be impacted and even be worse than some of the original learning machines. So in order to solve the problem, an improved version of averaging called weighted averaging is proposed which can be represented as $H(x) = \sum_{i=1}^{T} w_i h_i(x)$, where h is the individual learning machine and w is its weight. However, facts show that finding suitable weights for different learning machines is still difficult because of the noise of the data. There is another fusion strategy called stacking. Stacking is a method that needs an extra learning process to get the fusion result. For example, if the original learning machines take the training data x_i as input and produce the probability lists as the predict result $h_i(x_i)$, stacking combines the result as $H(x)$ which can be represented as Eq. 2

$$H(x) = h^{'}(h_1(x_1), h_2(x_2), \ldots, h_n(x_n)) \tag{2}$$

The fusion result could be regarded as the new input of the secondary learning machine and the prediction of this machine is the result of the fusion. Many studies proved that stacking is an efficient method in feature fusion. This paper compares averaging and stacking and chooses the stacking as our final fusion method. The result will be given in the experiment section.

5 Experiments

In this section, we present the empirical studies conducted on the real large dataset, i.e., Sina Weibo. We first introduce the baselines and the dataset description. We then discuss the prediction performance and reveal illustrate cases.

5.1 Dataset Setup

The network of the users in Sina Weibo is based on the relationship of bi-follower. There are 93,264 vertices and 1,688,681 edges used in this experiments. We choose the first 5 occupation groups as the evaluation set.

5.2 Baseline and Evaluation Metrics

To systematically compare the proposed methods' performance, we choose several baseline according to their features and methods:

Graph Partitioning: Graph Partitioning is an unsupervised partition method. It takes network G as input and decomposes the graph into different communities C. For every $c \in C$, we label the community with the most common group appears in c. In this experiment, we use the METIS library [1].

SocioDim: SocioDim algorithm is a representation learning algorithm based on community detection process [14]. SocioDim takes the network's communities in use and utilizes the network's eigenvectors to generate a latent representation in R^d where d is a small dimension. SocioDim usually takes advantage of the modular graph partitions of the network.

DeepWalk: We use Deepwalk approach [11] to get the low dimensional vector representation of users' social network information. It acts as the baseline to test the potential of the social network features.

Content Classification: Different from the methods above, this method takes the content of the messages as the input. After several processing just like extracting keywords, calculating weights and so on, we get the vector representation of the content for users and use the boost classification based on XGBoost package[3] to estimate users' occupation [8].

5.3 Effectiveness

We conduct the five-fold cross validation experiment, on top of scikit-learn package[4].

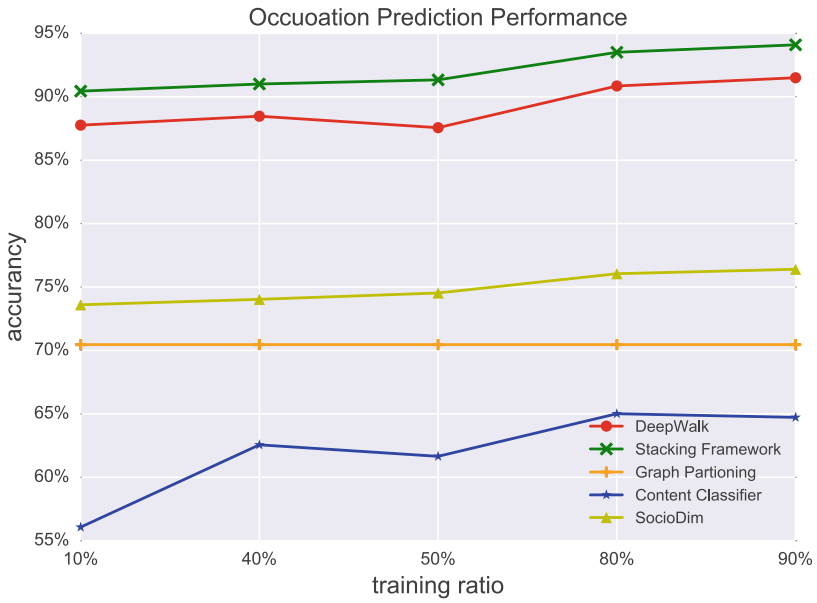


Fig. 2. Occupation prediction performance

Figure 2 shows the accuracy score of different methods. We can find that content classification method is not comparable with graph ones. Under five class prediction tasks, content feature is useful, though not competitive enough to provide satisfying performance. It is largely due to the short text collection. In the categories of graph based methods, the unsupervised partition method

[3] https://xgboost.readthedocs.io/en/latest/.
[4] http://scikit-learn.org/.

lags behind. The SocDim performance also can not compare with the new low dimensional space learning approaches. With the help of feature fusion, the proposed approach can gain a further increase on top of Deepwalk's already very high performance.

The result demonstrates that the unified framework with graph feature embedding and content feature fusion can improve the occupation prediction in social media domains. The classification based inference model is effective and able to cope with the social media challenges.

5.4 Training Ratio

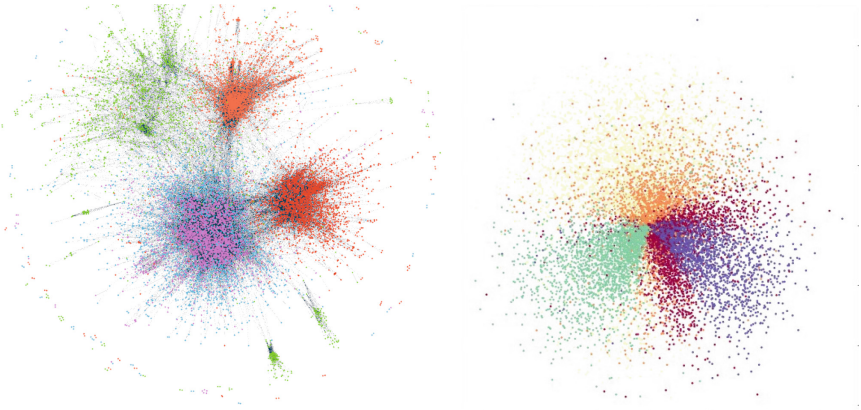
To test the sensitivity of the proposed approach, we extend the ratio scope and report the Micro/Macro F1 measures in Table 2. It is revealing that the accuracy is growing gradually with more training data. The network effect enables that a small portion of the training data can deliver satisfying results.

Table 2. Varying training ratio

Rate(%)	10	20	30	40	50	60	70	80	90
Micro-F1	87.65	90.44	89.80	91.00	91.33	91.72	92.25	93.50	94.10
Macro-F1	87.10	89.54	88.87	91.17	90.74	91.47	92.13	93.61	91.09

5.5 Showing Cases

To show the occupation prediction improvements, we choose to visualize the raw graph representation and the embedded prediction results in Fig. 3. The



(a) Raw Network (b) Embedded Network

Fig. 3. Top five occupation categories of users.

raw graph presentation is generated with the help of Gephi[5], and the embedded result is transformed using MDS algorithm [5]. We can clearly find the network coherence in the occupation categories are vivid, and the embedding process guarantees the occupation distinct and easy to interrupt.

6 Conclusions

In this paper, we discuss the problem of user profiling from social media data. We choose the occupation prediction to illustrate the features and models used in this new scenario. Based on a unique micro-blog dataset, we discuss the content and graph features for occupation prediction.

We propose to utilize the graph embedding and text classification for the prediction model. The fused approach also improve the individual based baselines. Extensive experiments on the large real dataset demonstrate the improvement and effectiveness of our new approach.

It is also remarkable that the comprehensive occupation prediction method can illustrate the users' professional interaction and posting activities. The future work of this paper can be focused on the occupation interpretation and friendship recommendation, with the help of efficient occupation pattern mining.

Acknowledgements. The research is supported by the National Natural Science Foundation of China under Grant No. 61502169, 61401155 and NSFC-Zhejiang Joint Fund for the Integration of Industrialization and Informatization Grant No. U1509219.

References

1. Abou-Rjeili, A., Karypis, G.: Multilevel algorithms for partitioning power-law graphs. In: 20th International Parallel and Distributed Processing Symposium, IPDPS 2006, p. 10-pp. IEEE (2006)
2. Bengio, Y., Courville, A., Vincent, P.: Representation learning: a review and new perspectives. IEEE Trans. Pattern Anal. Mach. Intell. **35**(8), 1798–1828 (2013)
3. Cao, S., Lu, W., Xu, Q.: GraRep: Learning graph representations with global structural information. In: Proceeding of CIKM, pp. 891–900 (2015)
4. Cha, M., Haddadi, H., Benevenuto, F., Gummadi, P.K.: Measuring user influence in twitter: The million follower fallacy. In: Proceeding of ICWSM, pp. 10–17 (2010)
5. Cox, T.F., Cox, M.A.: Multidimensional Scaling. CRC Press, Boca Raton (2000)
6. Farseev, A., Nie, L., Akbari, M., Chua, T.S.: Harvesting multiple sources for user profile learning: a big data study. In: Proceeding of ACM Multimedia, pp. 235–242 (2015)
7. Huang, Y., Yu, L., Wang, X., Cui, B.: A multi-source integration framework for user occupation inference in social media systems. World Wide Web **18**(5), 1247–1267 (2015)
8. Manning, C.D., Raghavan, P., Schütze, H., et al.: Introduction to Information Retrieval, vol. 1. Cambridge University Press, Cambridge (2008)

[5] https://gephi.org.

9. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781 (2013)
10. Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., Dean, J.: Distributed representations of words and phrases and their compositionality. In: NIPS, pp. 3111–3119 (2013)
11. Perozzi, B., Al-Rfou, R., Skiena, S.: DeepWalk: Online learning of social representations. In: Proceeding of SIGKDD, pp. 701–710 (2014)
12. Roweis, S.T., Saul, L.K.: Nonlinear dimensionality reduction by locally linear embedding. Science **290**(5500), 2323–2326 (2000)
13. Sun, Y., Norick, B., Han, J., Yan, X., Yu, P.S., Yu, X.: Integrating meta-path selection with user-guided object clustering in heterogeneous information networks. In: Proceeding of SIGKDD, pp. 1348–1356 (2012)
14. Tang, L., Liu, H.: Relational learning via latent social dimensions. In: Proceeding of SIGKDD, pp. 817–826 (2009)
15. Yan, S., Xu, D., Zhang, B., Zhang, H.J., Yang, Q., Lin, S.: Graph embedding and extensions: a general framework for dimensionality reduction. IEEE Trans. Pattern Anal. Mach. Intell. **29**(1), 40–51 (2007)
16. Yang, S.H., Long, B., Smola, A., Sadagopan, N., Zheng, Z., Zha, H.: Like like alike joint friendship and interest propagation in social networks. In: Proceeding of WWW, pp. 537–546 (2011)