# Inferring Social Network User's Interest Based on Convolutional Neural Network

Yanan Cao[1], Shi Wang[2](✉), Xiaoxue Li[1], Cong Cao[1], Yanbing Liu[1], and Jianlong Tan[1]

[1] Institute of Information Engineering,
Chinese Academy of Sciences, Beijing, China
{caoyanan, lixiaoxue, caocong,
liuyanbing, tanjianlong}@iie.ac.cn

[2] Institute of Computing Technology, Chinese Academy of Sciences,
Beijing, China
wangshi@ict.ac.cn

**Abstract.** Learning microblog users' interest has important significance for constructing more precise user profile, and can be useful for some commercial applications such as personalized advertisement, or potential customer analysis. Existing works generally utilize text mining or label propagation methods to solve this problem, which leverage either the user's publicly available comments or the user's social links, but not both. As we will show, these learning methods achieve limited precision rates. To address this challenge, we consider the interest inference task as a multi-value classification problem, and solve it using a convolutional neural network architecture. We innovatively present an ego social-attribute network model which integrates the target users' attributes, social links and their comments, and represent the ego SA network as the input fed to CNN. As a result, we assign each microblog user one or more interest labels (such as "loving sports"), which is different from previous approaches using non-uniform interest keywords (such as "basketball", "tennis", etc.). Experimental results on SMP CUP and Zhihu dataset showed that the precision rate of user interest inference reached 77.9% at best.

**Keywords:** Social-attribute network · Convolutional neural network · User interest inference

## 1 Introduction

As an important social media service, microblog is a wonderful platform where people share their thoughts, status and even their personal information. As the continuously increasing of microblog users, the analysis of users' attributes, relations and behaviors has received more and more attention both in academic and industry. Specifically, microblog users' interests can reflect users' preference and also have a close relationship with users' other attributes such as gender, age and occupation. Therefore, modeling users' interests has important significance for getting more precise user profile, and can be useful for commercial applications such as personalized advertisement, or potential customer analysis. At present, the miss rate of users' registration

D. Liu et al. (Eds.): ICONIP 2017, Part V, LNCS 10638, pp. 657–666, 2017.
https://doi.org/10.1007/978-3-319-70139-4_67

interest tags is higher than 70% [1], which means that most users' implicit interests should be learned. As important data sources, microblog contents and social links involve personal preferences which directly reflect user interests.

To mine user interests from microblog contents, existing studies have proposed two major methods including TextRank [2, 3, 5] and Topic Model [6, 7]. Although these methods are considered the state-of-the-art unsupervised keyword extraction and clustering methods, they face two challenges. On one hand, extracted interest keyword may provide an ambiguous representation of the topic. On the other hand, the topic model can obtain potential topics in texts, but explicit topic semantic labels are not given.

To make use of social links, label-propagation-based works [16–19] propagate missing attribute values from label nodes to unlabel nodes. The foundation of label-propagation-based work is homophily, which means that two linked users share similar attributes. Main label-propgation algorithms include MV, GSSL and CP are used in different kinds of attribute value inference, such as school, location, interest, etc. The average precision reached 60% to 70% and computing cost is typically high.

To deal with these above problems, we propose a novel method to infer user interest based on both users' attributes, social links and comments. Here, we consider the interest inference task as a multi-value classification problem, and solve it using the convolutional neural network architecture. We innovatively present an ego social-attribute network model which integrates the target users' attributes, social links and their comments, and represent the ego SA network as the input fed to CNN. Our CNN architecture contains two layers of convolution, which capture latent relations between the target user and his neighbor nodes, and the output is the probability distribution over interest labels. During the pre-processing stage, we mine and cluster frequent interest phrases, which have clearer semantic information than keywords, from users' comments based on the method proposed in [20]. For each user, we select the top N topics his interest phrases belong to as important attributes.

We evaluate our method both on the SMP CUP 2016 dataset and Zhihu dataset, which contains about 20 thousand Sina microblog users and 30 million contents in Chinese. Experimental results show that CNN architecture can achieve better results than traditional classifier models on precision, and performs well on time complexity. The precision rate of user interest inference reached 77.9% at best.

The main contributions of this paper are as follows:

- We present an ego social-attribute network model which integrates the target users' attributes, social links and their comments.
- We design a convolutional neural network architecture under ego SA network to infer users' implicit interests, which specially performs well on multi-valued interest inference.

## 2 Related Works

Existing interest inference works can be roughly classified into two categories, text mining based method and label propagation based one.

**Text Mining Based Methods:** Researchers tried to use TextRank to build a word-based graph and to use PageRank [8] to get top n candidate keywords as users' interest keywords, which gained 31.2% precision and recall rate of 43.1% [5]. Some researchers describe users' interests by using a set of tuples of content directives (categories to which user interests belong) and action indicators (actions related to interest categories), which can effectively exploit the real-time interests of microblog users [10]. Others consider the time distribution of microblog contents and use the time series to classify users' contents [3]. The precision rate of the classification was increased to 67%. These methods make use of the statistical properties or semantic information of words in text. They have made some effects in mining the interest information of microblog users, but they can't make use of statistical features in documents and between documents, and can't solve the ambiguity problem of interest words either.

The topic model performs better in this respect. Zhang used LDA to extend the text feature space, and then used words' frequency to extract the hot topic, which makes the hot topic rank higher [14]. Ramage also use the aggregated information to train the LDA model, and the experimental results show that the model is more suitable for the modeling of "author-feature topic" [15]. Weng proposed Twitter-LDA to filter the non-hot topic words and compared them with the distribution of hot topics in traditional media [6]. They find that most of the topics in microblog contents are about the daily life of users, which more reflect users' personal interests. These studies showed that the topic model can efficiently mine interests from the sparse and short text such as microblog contents by using the distribution of words and topics in the text and the distribution of topics and documents. However, in the existing works, the semantic information of topics and the categories of user interests are not clearly identified.

**Label Propagation Based Methods:** To make use of social links, label-propgation-based works [16–18] propagate missing interest values from labeled nodes to unlabeled nodes. Li present a hidden factor in social connections-relationship type and propose a co-profile users' attributes and relationship types base on this development [16]. Through iteratively profiles attributes by propagation via certain types of connections, and profiles types of connections based on attributes and the network structure, their algorithm profiles various attributes accurately. Dong design different strategies for computing the relational weights between users' attributes and social links and used a graph-based semi-supervised learning (GSSL) algorithm to infer attributes [17]. Dougnon proposed a new lazy algorithm PGPI to infer user profiles by using rich information (such as group memberships) without training [18]. These methods are used in different kinds of attribute value inference, such as school, location, interest, etc. The average precision reached 60% to 70%. Interest inference using social structure and attributes could also be solved by a social recommender system in [19]. However, such approaches have higher computational complexity than above methods.

To address above shortages of existing research, we consider the interest inference task as a multi-value classification problem, and solve it using a convolutional neural network architecture. And we present an ego social-attribute network model which integrates the target users' attributes, social links and their comments.

## 3 Problem Definition

In this section, we will formally introduce the definition of user interests' inference problem. We start by describing our social-attribute network model, which integrates social structure and user attributes in a unified framework. Here, user attributes not only involve gender, status, country, etc., but also contain the top N topical phrases extracted from user comments, which may reflect the user interests in a certain extent.

**Definition 1 (Social-Attribute Network).** We denote a Social-Attribute network (SA network for short) as $G = (V, E, t)$, where $V$ is the set of nodes, $E$ is the set of links, and $t$ a function that maps a node to its node type, i.e., $t_u$ is the node type of $u$. Nodes corresponding to users, attributes and interests are respectively called *social nodes*, *attribute nodes* and *interest nodes*, which are represented as $S$, $A$ and $I$. Links between social nodes are called *social links*, links between social nodes and attribute nodes are called *attribute links*, and links between social nodes and interest nodes are called *interest links*. Additionally, for a given node $u$ in the SA network, we denote by $\Gamma_{u,s}$, $\Gamma_{u,A}$, $\Gamma_{u,I}$, respectively the sets of all social neighbors, attribute neighbors and interests of $u$, and the neighbor nodes of $u$ is $\Gamma_u = \Gamma_{u,s} \cup \Gamma_{u,A} \cup \Gamma_{u,I}$.

**Definition 2 (ego SA network).** The ego social-attribute network of a target user $u$ is represented as a graph $EG_u = (V', E', t)$, where $V' = \bigcup\limits_{v \in u \cup \mathit{\Gamma}_{u,s}} (\Gamma_{v,A} \cup \Gamma_{v,I}) \cup u \cup \Gamma_{u,s}$, $E' = \{(u', v') | u', v' \in V')\}$, and $t$ a function that maps a node to its node type. Later in this article, we call the target user in an SA ego network an ego user.

Figure 1 illustrates an example SA network, in which the social node set is $S$={$v_1$, $v_2$,$v_3$,$v_4$,$v_5$}, attribute node set is $A$={*status*, *gender*, *age*, *country*} and interest set is $I$= {*sports*, *music*, *drawing*}. For a given ego user $v_1$, its social neighbor set is $\Gamma_{u,s}$={$v_2$, $v_3$}, its attribute neighbor set is $\Gamma_{u,A}$={*student*, *China*}, and its interest node set is $\Gamma_{u,I} = \phi$. Next, we aim to predict $\Gamma_{u,I}$ in the ego network $EGv_1$, which contains social nodes {$v_1$,$v_2$,$v_3$}, their attribute nodes, interest nodes and links between these.

More formally, the problem of interest inference of a target user $u$ in an ego network $G$ is to predict the interest link of $u$ based on his attributes and neighbor nodes' information in the ego network $EG_u$. We represent these information as an $n$*$k$ matrix and feed it to a CNN model to implement the multi-value classification.

## 4 Proposed Architecture

The model architecture, shown is Fig. 2, is a slight variant of the CNN architecture of [21]. The input is an $n * k$ matrix which represents an ego SA network. The second layer contains several convolution operations, and we take multiple convolution kernels of different sizes to extract features of input data.

We use a notation $f(\cdot)$ to represent a neural network. Any feed-forward neural network with $P$ layers can be seen as a composition of functions $f$, corresponding to each layer $p$.

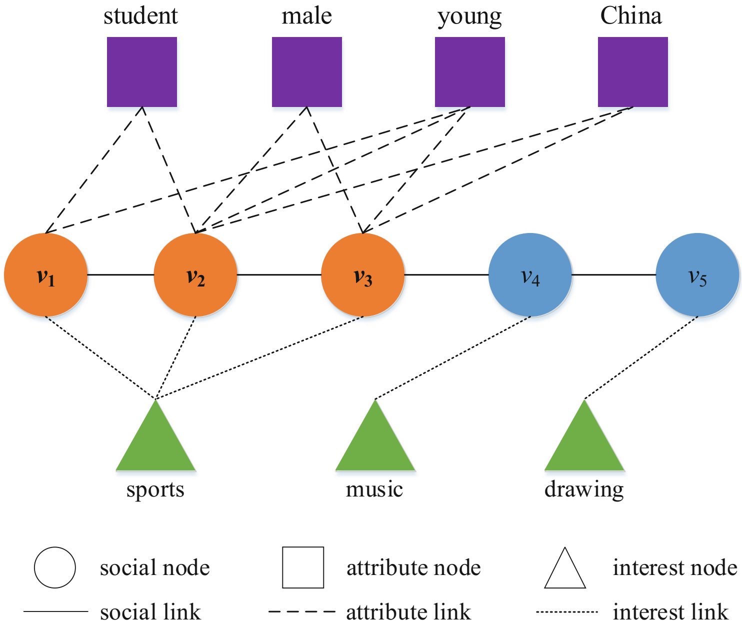

**Fig. 1.** An example of a social-attribute network and an ego SA network of $v_1$. All nodes and edges belong to a full social-attribute network and an ego-network of $v_1$ is represented as orange nodes and their attributes (Color figure online)

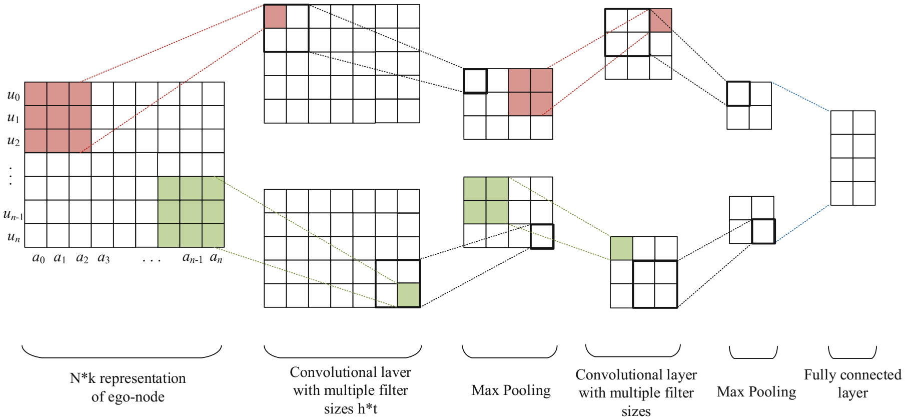
$u_0$
$u_1$
$u_2$
⋮
$u_{n-1}$
$u_n$
$a_0$ $a_1$ $a_2$ $a_3$ . . . $a_{n-1}$ $a_n$
N*k representation of ego-node
Convolutional laver with multiple filter sizes h*t
Max Pooling
Convolutional layer with multiple filter sizes
Max Pooling
Fully connected layer

**Fig. 2.** CNN architecture for social network user's interest inference

$$f_\theta(\cdot) = f_\theta^P(f_\theta^{P-1}(\cdot \cdot \cdot f_\theta^1(\cdot) \cdot \cdot \cdot)) \tag{1}$$

And initially, $f_\theta^0(\cdot) = F_i$ where $i \in (1, n)$. In the following, we could introduce our architecture layer by layer.

### Input Feature Map

Let $u_1 \in R^k$ be the $k$-dimensional attribute vector corresponding to the $i$-th social node in the ego-network and $u_0$ is the ego-node specially. An ego-network of size $n$ (padded or pruned where necessary) is represented as

$$u_{1:n} = u_1 \oplus u_2 \oplus \ldots \oplus u_n \tag{2}$$

where $\oplus$ is the concatenation operator. In general, $u_{i:i+j}$ refer to the concatenation of social nodes $u_i$, $u_{i+1}$, …, $u_{i+j.}$Each attribute value of $u_i$ is first passed through the lookup table layer, producing a numeric vector $ML_i$ of the same size as $L_{i\_}$. The feature can be viewed as the initial input of the standard convolution neural network. More formally, the initial input feature map fed to the convolution layer can be written as

$$f_\theta^P(\cdot) = ML_i = LTF(F_i) \tag{3}$$

### Convolution Layer

A convolution operation involves a filter $w$, which is applied to a window with size $h * r$ to produce a new feature. For example, a feature is generated from a window of social nodes $u_{i:i+h\text{-}1}$ and a window of attribute nodes $a_{j:j+r\text{-}1}$by

$$c_i = f(w \cdot u_{i:i+h-1} a_{j:j+r-1} + b) \tag{4}$$

where $b$ is a bias term and $f$ is a non-linear function. This filter is applied to each possible window of nodes in the ego-network to produce a feature map.

In our architecture, we use two convolution layers. The input vector can be fed to the standard neural network layer which performs affine transformations over their inputs

$$f_\theta^P(\cdot) = ReLU(wf_\theta^{P-1}(\cdot) + b) \tag{5}$$

Here ReLU is the active function. As for standard affine layers, convolution layers often stacked to extract higher level features.

### Max Pooling Layer

Local feature vectors extracted by the convolutional layers have to be combined to obtain a global feature vector, with a fixed size independent of the $L_{i\_}$, in order to apply subsequent standard affine layers. Then, we apply a max pooling over the feature map and take the maximum value as the feature corresponding to this particular filter. The idea is to capture the most important feature with the highest value for each feature map. This pooling scheme naturally deals with variable matrix size. Formally, given a matrix $f_\theta^{P-1}(\cdot)$ output by a convolution layer $p-1$, the max pooling layer output a vector $f_\theta^{P-1}(\cdot)$

$$[f_\theta^p]_i = \max_i [f_\theta^{p-1}]_{i,t} \tag{6}$$

where $t$ is the number of layer $p-1$ output. The fixed size global feature vector can be then fed to the standard affine network layers.

**Fully Connected Softmax Layer**
We have described the process by which one feature is extracted from one filter. The model uses multiple filters to obtain multiple features. These features from the penultimate layer and are passed to a fully connected softmax layer whose output is the probability distribution over interest labels.

# 5 Experiments and Results

In this section, we start with the introduction of the datasets and experimental setting. We then describe comparative methods and the evaluation on precision.

## 5.1 Datasets

There is no public benchmark in social network user interest inference problem. So, we use a dataset provided from SMP CUP 2016 (a microblog user profile contest held by Sina) which contains more than 2000 users and 230,000 microblog contents. This dataset is divided from a real Sina microblog dataset in Chinese, which contains about 46,000 users, more than 30,000,000 microblog contents. SMP CUP dataset provides plenty of user contents but inadequate user attributes. So, we constructed another dataset containing 20000 users, 50 contents and 13 dimensions of attributes for each user, in which the user attribute, user link and user comments are crawled from the online social network website Zhihu. The interest label for each user is extracted semi-automatically and verified annually.

**User Attribute Selection:** We select 8 attributes (gender, age, status, major, university, i.e.) which are commonly used information in social media platforms. In our dataset, there are some missing attribute value and noisy information. For example, some user's status is 'loving money'. In this case, we use specific tag to represent missing attribute values and noisy data.

**Topic Mining:** We preprocess users' contents by deleting specific symbols and removing duplication. Chinese stop words are also removed for phrase mining and topic modeling steps. Then, we extract users' interest candidate phrases using an effective topical phrase mining method [19]. In order to identify the topics users interested in, we semi-automatically construct a hierarchical topic knowledge base and utilize it to identify users' topical phrases. For each user, his top 5 topical phrases are used as user attributes in the input matrix.

After these preprocessing procedures, above datasets are divided to training dataset and test dataset according to the proportion of 2:1.

## 5.2 Experimental Setting

We compared the performance of the proposed algorithm with four state-of-the-art algorithms: Linear Regression, Naïve Bayes classifiers [9], Graph Semi-Supervised Learning and Majority Voting. These four algorithms predict the value of target user's gender, status and major respectively.

Linear Regression (LR): we construct a linear function by using our training dataset, and predict the missing values using this function.

Naïve Bayes (NB) Classifiers: NB infer user profiles strictly based on correlation between attributes values which is as well as our UPS model.

Graph Semi-Supervised Learning (GSSL) [11] and Majority Voting (MV) [6] infer user's profiles by using the social structures which is the same as our CNN model. For algorithms which need specific parameters, we use empirical value to achieve the state-of-art results.

Both the Linear Regression and Naïve Bayes classifiers perform well on binary classification. Because inferring user's potential interest is a problem of multi-classification, we also use a general neural network (NN) algorithm as a comparative method. Note that, these three methods don't use user links in the interest inference procedure, which is different from GSSL and MV.

Contrast experiments were performed on a computer with a fourth generation 64-bit Core i5 processor running Ubuntu 14.5 and 16 GB of RAM.

## 5.3 Experimental Results

We evaluate the accuracy of CNN and the comparative models on user interest inference. Table 1 shows the evaluation results on SMP CUP dataset and Zhihu dataset. We can see that, on both datasets, CNN significantly outperforms other methods, which achieved 52.6% and 77.9% precision respectively. This result demonstrates that CNN has great advantages in dealing with multiple classification problems. Besides, the effectiveness of NN and GSSL (which make good use of multiple dimensions of user attributes and user links respectively) is also acceptable. MV is the worst model in user interest inference, because its strategy is too simple to deal with this problem. Compared with NN, CNN improved the accuracy significantly for its model architecture and the use of social links in the neural network, which reflects that CNN could learn the hidden relation between the ego node and its social neighbors.

**Table 1.** Accuracy of different algorithms on two datasets

| Dataset\ Accuracy | CNN | NN | NB | LR | GSSL | MV |
|---|---|---|---|---|---|---|
| SMP CUP dataset | 52.6% | 48.2% | 44.3% | 48.4% | 49.0% | 40.2% |
| Zhihu dataset | 77.9% | 72.3% | 67.4% | 70.6% | 71.30% | 52.50% |

We also evaluate the impact of different training dataset size on above classification models, including CNN, NN, NB, LR. The evaluation result is demonstrated in Fig. 3. We can see that the effectiveness of these classifiers was generally improved as the training data increases. When the training set contains less than 5,000 users, the effectiveness of CNN and NN is almost equivalent. Because under this situation, users' links are sparse and unbalanced, which conducts a negative impact on the effectiveness of CNN.

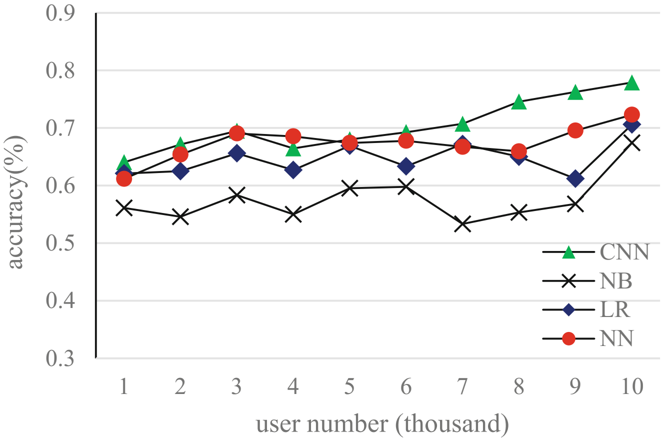

**Fig. 3.** The accuracy of four algorithms on different dataset size (Zhihu)

## 6 Conclusion and the Future Work

In contrast to other classifier models, CNN performs better on predicting multi-valued interest for online social network users. Our experimental results showed that it is possible to learn users' potential interests based on user links and user profiles. Our CNN architecture is just a single channel model. In the future work, we would like to use a multichannel architecture which may make better use of attributes and topical phrases. Furthermore, we will evaluate the impact of users' topical phrases on our algorithms.

**Acknowledgement.** This work was supported by the National Natural Science Foundation of China grants (NOs. 61403369, 61602466), the National Key Research and Development program of China (Nos. 2016YFB0801304, 2016YFB0800303).

## References

1. Ding, Y.X., Xiao, X., Wu, M.J.: Predicting users' profiles in social network based on semi-supervised learning. J. Commun. **35**(8), 15–22 (2014)
2. Vu, T., Perez, V.: Interest mining from user tweets. In: Proceedings of ACM International Conference on Information and Knowledge Management, San Francisco, CA, USA (2013)
3. Yang, T., Lee, D.W., Yan, S.: Steeler nation, 12th man, and boo birds: classifying Twitter user interests using time series. In: Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks and Mining, pp. 684–691 (2013)
4. He, L., Jia, Y., Han, W.H., Ding, Z.H.Y.: Mining user interest in microblogs with a user-topic model. China Commun. **11**, 131–144 (2014)

5. Mihalcea, R., Tarau, P.: Textrank: bringing order into texts. In: Proceedings of 2004 Conference on Empirical Methods in Natural Language Processing, pp. 404–411 (2004)
6. Zhao, W.X., Jiang, J., Weng, J., He, J., Lim, E.-P., Yan, H., Li, X.: Comparing Twitter and traditional media using topic models. In: Clough, P., Foley, C., Gurrin, C., Jones, G.J.F., Kraaij, W., Lee, H., Mudoch, V. (eds.) ECIR 2011. LNCS, vol. 6611, pp. 338–349. Springer, Heidelberg (2011). doi:10.1007/978-3-642-20161-5_34
7. Zhang, C.Y., Sun, J.L., Ding, Y.Q.: Topic mining for microblog based on MB-LDA model. J. Comput. Res. Dev. **48**(10), 1795–1802 (2011)
8. Page, L., Brin, S., Motwani, R., Winograd, T.: The pagerank citation ranking: bringing order to the web. Technical report, Stanford Digital Library Technologies Project (1998)
9. Lindamood, J., Heatherly, R., Kantarcioglu, M., et al.: Inferring private information using social network data. In: Proceedings of the 18th International Conference on World Wide Web, pp. 1145–1146. ACM (2009)
10. Banerjee, N., Chakraborty, D., Dasgupta, K., et al.: User interests in social media sites: an exploration with micro-blogs. In: Proceedings of the 18th ACM Conference on Information and Knowledge Management, pp. 1823–1826 (2009)
11. Hu, X., Sun, N., Zhang, C., Chua, T.S., et al.: Exploiting internal and external semantics for the clustering of short texts using world knowledge. In: Proceedings of the 18th ACM Conference on Information and Knowledge Management, pp. 919–928 (2009)
12. Abel, F., Gao, Q., Houben, G.-J., Tao, K.: Semantic enrichment of Twitter posts for user profile construction on the social web. In: Antoniou, G., Grobelnik, M., Simperl, E., Parsia, B., Plexousakis, D., De Leenheer, P., Pan, J. (eds.) ESWC 2011. LNCS, vol. 6644, pp. 375–389. Springer, Heidelberg (2011). doi:10.1007/978-3-642-21064-8_26
13. Musat, C.C., Velcin, J., Trausan-Matu, S., Rizoiu, M.A., et al.: Improving topic evaluation using conceptual knowledge. In: Proceedings of the Twenty-Second International Joint Conference on Artifical Intelligence-Volume, vol. 3, pp. 1866–1871 (2011)
14. Zhang, S., Luo, J., Liu, Y., Yao, D., et al.: Hotspots detection on microblog. In: 2012 Fourth International Conference on Multimedia Information Networking and Security(MINES), pp. 922–925. IEEE (2012)
15. Ramage, D., Hall, D., Nallapati, R., et al.: Labeled LDA: a supervised topic model for creditattribution in multi-labeled corpora. In: Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing, pp. 248–256 (2009)
16. Li, R., Wang, C., Chang, K.C.C.: User profiling in an ego network: co-profiling attributes and relationships. In: Proceedings of the 23rd International Conference on World Wide Web, pp. 819–830 (2014)
17. Dong, Y., Tang, J., Wu, S., Tian, J., et al.: Link prediction and recommendation across heterogeneous social networks. In: 2012 IEEE 12th International Conference on Data Mining (ICDM), pp. 181–190. IEEE (2012)
18. Dougnon, R.Y., Fournier-Viger, P., Nkambou, R.: Inferring user profiles in online social networks using a partial social graph. In: Barbosa, D., Milios, E. (eds.) CANADIAN AI 2015. LNCS, vol. 9091, pp. 84–99. Springer, Cham (2015). doi:10.1007/978-3-319-18356-5_8
19. Ye, M., Liu, X., Lee, W.C.: Exploring social influence for recommendation a probabilistic generative model approach. In: SIGIR (2012)
20. El-Kishky, A., Song, Y., Wang, C., Voss, C.R., Han, J.W.: Scalable topical phrase mining from text corpora. PVLDB **8**(3), 305–316 (2015). Also, In: Proceedings of 2015 International Conference on Very Large Data Bases (VLDB 2015), Kohala Coast, Hawaii, September 2015
21. Krizhevsky, A., Sutskever, I., Hinton, G.E.: ImageNet classification with deep convolutional neural networks. In: Proceedings of NIPS (2012)