Measuring and Visualizing Interest Similarity between Microblog Users

Jiayu Tang, Zhiyuan Liu, and Maosong Sun

State Key Laboratory of Intelligent Technology and Systems
Tsinghua National Laboratory for Information Science and Technology
Department of Computer Science and Technology,
Tsinghua University, Beijing 100084, China
{tjy430,lzy.thu}@gmail.com, sms@tsinghua.edu.cn

Abstract. Microblog users share their life status and opinions via microposts, which usually reflect their interests. Measuring interest similarity between microblog users has thus received increasing attention from both academia and industry. In this paper, we design a novel framework for measuring and visualizing user interest similarity. The framework consists of four components: (1) Interest representation. We extract keywords from microposts to represent user interests. (2) Interest similarity computation. Based on the interest keywords, we design a ranking framework for measuring the interest similarity. (3) Interest similarity visualization. We propose a integrated word cloud scenario to provide a novel visual representation of user interest similarity. (4) Annotation data collection. We design an interactive game for microblog users to collect user annotations, which are used as training dataset for our similarity measuring method. We carry out experiments on Sina Weibo, the largest microblogging service in China, and get encouraging results.

Keywords: interest similarity, information visualization, microblogging, keyword extraction.

1 Introduction

Microblogging is a new form of blogging in the Web 2.0 era. Typical microblogging services include Twitter[1] and Sina Weibo[2]. Microblogging services allow users to post small elements of content such as short text messages, images, videos, etc., the so-called *microposts*. Microposts are made by succinctly broadcasting information within a certain length, e.g., 140 English or Chinese characters. Users usually share information, update daily activities, and seek knowledge on microblogging services. Microposts can thus reflect user interests to a certain extent, and we can identify a user's major interests by extracting representative words and phrases in the microposts [1]. We call these words and phrases as *keywords*.

[1] http://twitter.com/
[2] http://weibo.com/

J. Wang et al. (Eds.): WAIM 2013, LNCS 7923, pp. 478–489, 2013.

As a typical social media, the relations between users in microblogging have attracted many attentions from both research and commercial communities. On a microblogging service, a user may follow and pay close attention to anyone who s/he is interested in. The reasons behind the following behaviors are complicated. For example, a user may follow another one just because they have social connections in the real-world, or because they share similar interests so that they can keep an eye and communicate with each other. Currently, most researches focus on studying the following behaviors through the following-network structure. In fact, however, since the microposts imply the interests of users, we can model user relations via their similarities of interests. Therefore, as an alternative to the perspective of network structure analysis, we use keywords as the representation of user interests, and propose a novel framework to measure the interest similarity between users.

In this paper, a novel framework is proposed for measuring and visualizing interest similarity between microblog users. We extract keywords from microposts to represent user interests. With interest keywords, a support vector machine for ranking (SVM-rank) model is learned, which measures the interest similarity effectively. Then, we extend Wordle [2], a widely-used text visualization, into a integrated word cloud scenario to make viewers comprehend the interest similarity between microblog users clearly and intuitively. Besides, we design an interactive game for microblog users to collect user annotations for SVM-rank model training.

The rest of this paper is organized as follows. We first briefly discuss related work in Section 2, followed by the details of our framework in Section 3. Then, we evaluated and verified our framework on a real-world microblogging service to show the effectiveness of our framework in Section 4. Finally, we conclude with a discussion and future directions in Section 5.

2 Related Work

2.1 Microblogging Analysis

Over the last couple of years, microblogging has been investigated from several perspectives. Java et al. [3] studied the topological and geographical properties of Twitter's social network to understand the community structure in microblogging. Kwak et al. [4], Wu et al. [5], and Bakshy et al. [6] investigated the diffusion of information on Twitter. Zhao and Rosson [7] qualitatively investigated the motivation of the users who use Twitter. Krishnamurthy et al. [8] analyzed the characteristics of Twitter users in such aspects as classes, behaviors, and social networks.

There are also some researches investigating user interests in microblogging. Using natural language processing tools, Piao and Whittle [9] extracted named entities and core terms to identify individual Twitter users' interests. Using TFIDF and TextRank, Wu et al. [10] also extracted keywords from the Twitter microposts to label user's interests and concerns. Yamaguchi et al. [11] extracted tags from the names of "Twitter List" (i.e., user groups) to discover appropriate

topics for list members and identify their common interests. Michelson and Macskassy [12] leveraged Wikipedia as a knowledge base to categorize the entities in the Twitter microposts and built a topic profile for each Twitter user. Banerjee et al. [1] analyzed real time user interests in different cities by mining content-inductive and usage-inductive keywords, each of which represents different class of user interests.

As shown in the above-mentioned works, keywords have been proved to be an appropriate and encouraging way to identify user interests. In this paper, we also aim to extract keywords to estimate microblog users' interests for further measurement.

2.2 Word Cloud

A word cloud, also known as a tag cloud, is a popular way of representing text data. It's typically used in the Web 2.0 era to visualize the frequency distribution of key terms which depict the website content. A word cloud is originally organized in horizontal lines, and popularized in the photo sharing site Flickr[3] to show the user-generated tags of photos in 2004. A cloud encodes word importance or frequency information via font size. There have been several tools for generating word clouds from text provided by the Web users. Wordle[4] is one of the most outstanding and appealing tools. It is widely used in the social community and mainstream media [2]. The word cloud created by Wordle is significantly different from regular ones because of its striking graphic statements: the words are arranged tightly and the display space is thus efficiently used; words can be placed in different orientation.

While Wordle is a state-of-the-art tool for generating a simple word cloud, it cannot show any relationship among words; that is, text content about two or more subjects cannot be recognized by the word cloud created by Wordle. Paulovich et al. [13] and Cui et al. [14] used a sequence of word clouds to show different document collections. Nevertheless, to the best of our knowledge, it is still difficult for current word cloud schemes to demonstrate the commonness and difference between two collections in an intuitive way.

The impact of the visualization on the user experience has also been studied [15]. A comparative study of several word cloud layouts suggested that appropriate layouts should be carefully selected according to the expected user goals [16].

3 Framework

In this section, we describe the framework for measuring and visualizing the similarity between microblog users (shown in Figure 1). We first show how to extract keywords from microposts to represent user interests (3.1). Further, we present a ranking approach to measuring the interest similarity between users

[3] `http://www.flickr.com/`
[4] `http://www.wordle.net/`

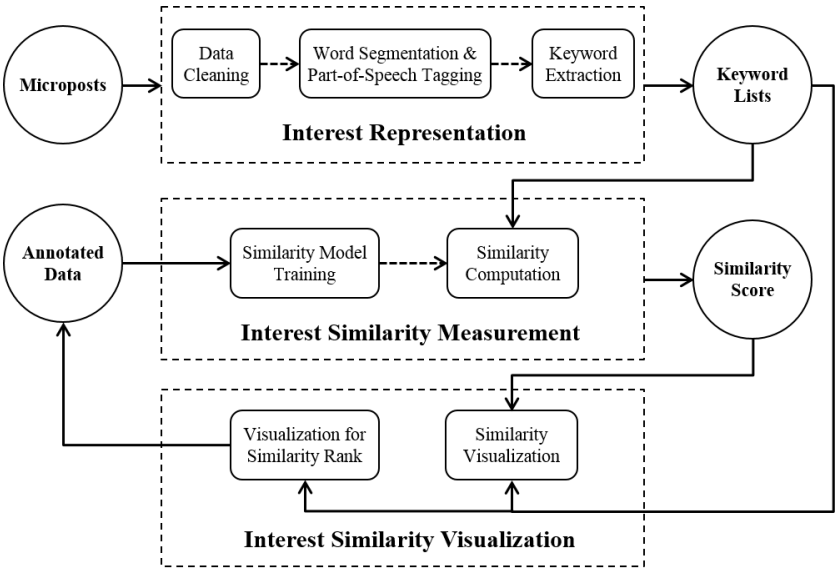

Fig. 1. Framework for measuring and visualizing the interest similarity between microblog users

(3.2). Then, we propose a integrated word cloud visualization to provide a novel and clear representation of user interest similarity (3.3).

3.1 Interest Representation

We identify a microblog user's interests by extracting keywords from his/her microposts. In this paper, we take Sina Weibo, the largest microblogging service in China, as our research service, and get users' microposts from its APIs[5]. On Sina Weibo, an overwhelming majority of the microposts are in Chinese. Hence, we perform Chinese word segmentation and part-of-speech (POS) tagging before keyword extraction.

Data Cleaning. Before word segmentation and POS tagging, we clean microposts and only keep plain text data in preprocessing. In China, a significantly large percentage of microposts are retweets [17]. A retweet usually contains two parts: the original micropost and a comment by the retweeting user. A user may retweet a micropost simply because it's popular, and the micropost usually contains more information about the user who post it originally than that about the retweeter. Therefore, we only retain the comment by the retweeter in a retweet micropost.

A user may use emoticons in microposts to help express his/her sentiment. As emoticons cannot directly help keyword analysis and further similarity

[5] `http://open.weibo.com/wiki/`

computation, we remove them from the microposts. Additionally, we also remove the user names mentioned in posts, URLs and other none-texts from microposts.

Word Segmentation and POS Tagging. With the clean data, we perform word segmentation and POS tagging using a practical system THULAC [18]. Other POS taggers such as Stanford Log-linear Part-Of-Speech Tagger [19] or Apache OpenNLP POS Tagger[6] can also be easily embedded in our framework to support other languages. After filtering the stop words, only notional words with specific concepts are selected as candidate words for keyword extraction: common nouns, person names, place names, institute names, idioms and other proper nouns.

Keyword Extraction. With filtered notional words, we perform keyword extraction derived from an efficient and effective framework proposed by Liu et al. [20], in which a translation-based method and a frequency-based method are combined. The translation-based method can summarize appropriate keywords in spite of the noise caused by varied subjects and the frequency-based method can find new words used in microposts. Their experiments on Sina Weibo have shown that this framework is effective and efficient for identifying user interests.

To measure interest similarity, we need to extract keywords from two microblog users' microposts. After keyword extraction, we get two keyword lists, each of which indicates a user's interests. In the keyword list, each keyword is assigned a weight.

3.2 Interest Similarity Measuring

Although the task of computing interest similarity is important for both academia and industry, there is actually no gold standard for directly evaluating the computation of interest similarity between microblog users. It is intuitive to use the cosine similarity of two keyword lists for similarity computation. The cosine scores in isolation, however, are usually too small due to the sparsity of keyword matrix, which cannot well illustrate the similarity between microblog users. In this paper, we propose to take multiple features together for measuring user similarity. For this method, the challenge is how to collect annotation data and design an approach to supervised learning.

We first introduce the method for similarity computation, and then introduce the collection method of annotation data.

Similarity Computation. With two keyword lists, we first use the vector space model to represent keyword lists and compute the cosine similarity between them. We select the cosine score, the number of common keywords and ratios of the number of common keywords to the numbers of two keyword lists as the features for measuring interest similarity between microblog users, using our similarity model learned from a Ranking SVM algorithm [21] (experiments

[6] `http://opennlp.apache.org/`

Fig. 2. Interface of the game application for collecting annotation data. A user press one Select button to choose whether user A or user B is more similar to himself/herself.

for the model selection are shown in Section 4.1). The value v derived from the similarity model is mapped into a specific score s by the following sigmoid function:

$$s = \frac{100}{1 + e^{-v}} \tag{1}$$

The final similarity score s is called *interest exponent* with range from 0 to 100.

Annotation Data Collection. To train the SVM-rank model, we design an interactive game for microblog users to collect annotation data. Compared to tell the preference between two users according to interest similarity, it is difficult for an annotator to give an absolute score of his/her similarity with a user. Hence, we ask a microblog user to provide his/her interest preference between his/her two friends. Moreover, if we only show interest keyword lists during the annotation, an annotator is difficult tell exactly how similar two users are. Therefore, in the interactive game, we use our novel visualization method to demonstrate the interest similarities (see Section 3.3 for detailed introduction) to help annotation. We finally implement a Web application to gather annotation data.

Figure 2 shows a snapshot of the game application. With a user ID on Sina Weibo, the application can access user data to generate a pair of word clouds as well as the information about the keyword numbers at each time. A user can click the Select buttons to annotate whether the left two microblog users are more similar or the right two ones, without knowing in advance who they are respectively. Then, the application will show the real identities of user A and user B. Moreover, the user can share the visualization with his/her annotation as a micropost on Sina Weibo, which can attract more users to use our game application.

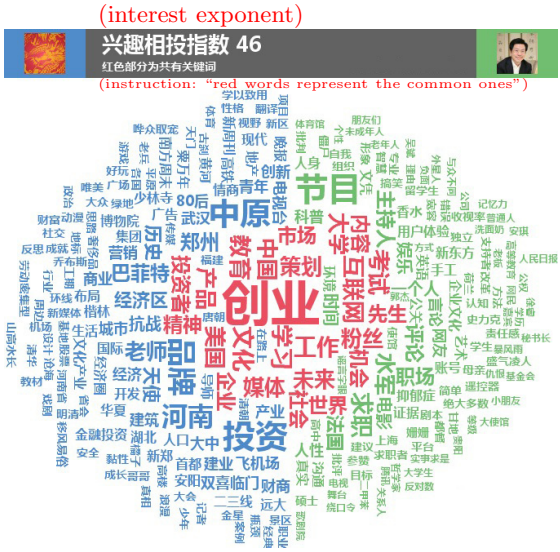

Fig. 3. Interest similarity visualization of two users on Sina Weibo

3.3 Interest Similarity Visualization

With interest keyword lists and their similarity score, we can create similarity visualization for viewers. By showing two microblog users' keywords and their common ones at the same time on canvas as well as the interest exponent, an overall view of interest similarity between two users is provided. Figure 3 shows an example of our similarity visualization, in which the common keywords of two microblog users, such as "创业 (startup)," "互联网 (internet)," and "教育 (education)," are shown in the center of the canvas. Moreover, the avatar of each user is placed on the corresponding side at the top of the canvas to make our visualization more self-explanation for viewers. In this section, we describe the visual representation and the rationale of our visualization.

Word Colors. Different colors are used to represent three kinds of keywords to make the visualization intuitionistic and aesthetic. Assuming that viewers are more interested in the common keywords of two microblog users, these ones are colored in striking red and are placed in the center of the canvas. The distinctive keywords of each user are separately colored in visually distinguishable blue and green.

Font Size. Following the most accepted visual design, font size is used to indicate word weight. Big words catch viewers' attention more easily than small ones [22], and the font size of the common keywords is thus bigger than that of the distinctive ones owned by each user. To balance the distinction and legibility, the variation of the font size of the different keywords has to be critically chosen. The font size $csize_i$ for a common keyword i is calculated as follows:

$$tsize_i = C_{max} - (C_{max} - C_{min})\sqrt{\frac{w_{max} - w_{H_i}}{w_{max} - w_{min}}} \tag{2}$$

$$csize_i = (\alpha v + \beta)tsize_i \tag{3}$$

where w_{max} is set to the maximum weight of the common keywords, w_{min} is set to the minimum weight of the common keywords, and w_{H_i} is the harmonic mean of the weights of i in two users' keyword lists. C_{max}, C_{min}, α, and β are the constant factors. v is the similarity value got from the similarity model, and the size of the common keyword is thus adjusted according to the similarity between microblog users. The more similar two microblog users are, bigger the size of the common keywords is.

The font size $dsize_j$ for a distinctive keyword j owned by only one microblog user u is calculated as follows:

$$dsize_j = D_{min} + (D_{max} - D_{min})(\frac{w_j}{w'_{max} - w'_{min}})^{\mu} \tag{4}$$

where w'_{max} is set to the maximum weight of the keywords of u, w'_{min} is set to the minimum weight of the keywords of u, and w_j is the weight of j. D_{max}, D_{min}, and μ are the constant factors.

Word Layout. Since Wordle's layouts are proven to be very compelling [2] and circular layout with decreasing weight is suitable for finding major concerns [16], we mimic Wordle's design rationale: the words with high weight are placed centrally on canvas, and the small ones fill the rest spaces to provide a holistic view. The layout of keywords proceeds on the basis of the pseudo code proposed in [2]. The common keywords have high priority when determining where to be placed. The distinctive keywords of each user are placed on one side of the dividing line (shown in Figure 4), which visually separates two users' keywords.

The number of keywords may vary widely with respect of the amount of users' microposts. To form an aesthetic view, the dividing line of two users' distinctive keywords thus has to be dynamic rather than always in the middle of the canvas.

Fig. 4. A case when the numbers of two users' keywords are quite different

Given two microblog users A and B, user A's keywords have to be placed on the left side of the dividing line, the boundary of which is $xpos_l$, and user B's have to be placed on the right side of the dividing line, the boundary of which is $xpos_r$. When placing user A's keywords, the position of $xpos_l$ is calculated as follows:

$$xpos_l = \frac{(\frac{n_A}{n_B})^\gamma}{1 + (\frac{n_A}{n_B})^\gamma} \tag{5}$$

where n_A is the number of user A's keywords, and n_B is the number of B's. γ is the constant factor and we experimentally set it to a value of 0.25. When all of user A's distinctive keywords are placed on the canvas and user B's are to be placed, the position of $xpos_r$ is calculated as follows:

$$xpos_r = min(xpos_l, xpos_{A_{max}}) \tag{6}$$

where $xpos_{A_{max}}$ is set to the rightmost position of user A's distinctive keywords (shown in Figure 4). Equation 6 thus creates a harmonious visual effects when the numbers and weights of the keywords of two users are quite different.

4 Experiments

4.1 Similarity Model Training

As aforementioned, given two interest keyword lists, we use the following features for training the similarity model:

- The cosine score of two keyword lists.
- The number of common keywords.
- The ratio of the number of common keywords to the number of keywords in the shorter list.
- The ratio of the number of common keywords to the number of keywords in the longer list.

With these features and target annotation by the Sina Weibo users of our game application, we are able to train the similarity model and evaluate its performance. In the experiments, our task is formalized as follows. Given a microblog user A and his/her two friends B and C, we get two pairs of users, e.g., (A, B) and (A, C). Between the two pairs of the users, our model should determine which pair is more similar with each other than another pair. We collect 500 groups of annotations. In each group, a user selects two friends and annotates which one is more similar to him/her.

We can regard the problem as a classification problem. That is, given a user and his/her two friends, the pair that is more similar with each other is annotated as the positive instance ($y = 1$) while another is negative ($y = 0$). To prevent over-fitting in training, we apply 10-fold cross-validation. The most simple and efficient algorithm for classification is linear model, while the state-of-the-art classification algorithm is support vector machines (SVM). In this paper, we use

Table 1. Accuracy on training set

Method	Parameters	Accuracy
LIBSVM	b=1, the rest is default	93.2%
LIBLINEAR	s=6, the rest is default	92.0%
Ranking SVM	default	94.0%

LIBSVM [23] and LIBLINEAR [24] as the toolkit of SVM and linear model, both of which are the most widely used tools in natural language processing and machine learning. Table 1 shows the evaluation results.

The task can also be regarded as a ranking problem. That is, for a user as the input query, we rank his/her friends with more similar ones ranked higher. The problem can be addressed by learning-to-rank algorithms. Learning-to-rank algorithms can be divided into three approaches, including point-wise, pair-wise and list-wise. Our ranking task is naturally a pair-wise ranking problem. Therefore, we select the stat-of-the-art pair-wise algorithm, Ranking SVM [25], to solve our problem. Cross validation accuracy of Ranking SVM shows that the learned rule to the similarity model is effective (see Table 1).

The evaluation results show that the ranking assumption is more effective than the classification assumption for computing the similarity of user interests. This is not surprising because that our task is more like a ranking problem than a classification problem. For example, when a user tries to compare his/her friends in our game application, s/he is more concerned about the order. In the classification setting, however, it rigidly sets the preferred as 1 while the other as 0. This does not conform to reality comprehensively, because the distances between two friends will not always be 1.

4.2 Performance of the Framework on Sina Weibo

We apply our framework for measuring and visualizing the interest similarity on Sina Weibo. From 20th March, 2012, to 31th December, 2012, users on Sina Weibo have used our online system to visualize the interest similarities between themselves and their friends for more than 140,000 times. This phenomenon indicates our framework is effective and attractive.

It is usually difficult to quantify how well people welcome a new visualization technique; however, most users describe our visualization using “interesting”, “intuitionistic”, and “beautiful” in their microposts and on the message board of our system. A great deal of positive feedback indicates our visualization of interest similarity is satisfactory.

5 Conclusion

In this paper, we propose a novel framework for measuring and visualizing interest similarity between microblog users. By applying Ranking SVM method

on interest keywords extracted from microposts, we measure microblog users' interest similarity effectively, and the integrated word cloud visualization makes viewers comprehend the interest similarity clearly and intuitively. Besides, the interactive and attractive game we designed for collecting user annotations can help to train SVM-rank model constantly for better performance. Since applied on Sina Weibo, the largest microblogging service in China, our framework has attracted more than 140,000 times of usage in 9 months and has got plenty of positive feedback, which shows our framework is effective and encouraging.

We will consider the following work as the future research plan: (1) Our framework for measuring interest similarity does not hold the similarity transitivity. For example, when $sim(A, B) > sim(A, C)$ and $sim(A, C) > sim(B, C)$, our method does not guarantee that $sim(A, B) > sim(B, C)$, where $sim(X, Y)$ is the similarity score of user X and user Y derived from the similarity model. We will improve the framework for measuring interest similarity to hold the similarity transitivity. (2) It is obvious that the interests of most users will change over time. We will incorporate time factors into our framework. (3) We will learn to recommend relevant and useful information, such as users with similar interests and articles on relevant topics, according to the results of interest similarity measurement.

Acknowledgements. This work is supported by the Key Project in the National Science and Technology Pillar Program under Grant No. 2009BAH41B04 and the Singapore National Research Foundation under its International Research Centre @ Singapore Funding Initiative administered by the IDM Programme Office. The authors would like to thank Shiqi Shen for his help.

References

1. Banerjee, N., Chakraborty, D., Dasgupta, K., Mittal, S., Joshi, A., Nagar, S., Rai, A., Madan, S.: User interests in social media sites: an exploration with micro-blogs. In: CIKM 2009, pp. 1823–1826. ACM, New York (2009)
2. Viegas, F.B., Wattenberg, M., Feinberg, J.: Participatory Visualization with Wordle. IEEE Transactions on Visualization and Computer Graphics 15, 1137–1144 (2009)
3. Java, A., Song, X., Finin, T., Tseng, B.: Why we twitter: understanding microblogging usage and communities. In: WebKDD/SNA-KDD 2007, pp. 56–65. ACM, New York (2007)
4. Kwak, H., Lee, C., Park, H., Moon, S.: What is Twitter, a social network or a news media? In: WWW 2010, pp. 591–600. ACM, New York (2010)
5. Wu, S., Hofman, J.M., Mason, W.A., Watts, D.J.: Who says what to whom on twitter. In: WWW 2011, pp. 705–714. ACM, New York (2011)
6. Bakshy, E., Hofman, J.M., Mason, W.A., Watts, D.J.: Everyone's an influencer: quantifying influence on twitter. In: WSDM 2011, pp. 65–74. ACM, New York (2011)
7. Zhao, D., Rosson, M.B.: How and why people Twitter: the role that micro-blogging plays in informal communication at work. In: GROUP 2009, pp. 243–252. ACM, New York (2009)

8. Krishnamurthy, B., Gill, P., Arlitt, M.: A few chirps about twitter. In: 1st Workshop on Online Social Networks, pp. 19–24. ACM, New York (2008)
9. Piao, S., Whittle, J.: A Feasibility Study on Extracting Twitter Users' Interests Using NLP Tools for Serendipitous Connections. In: PASSAT/SocialCom 2011, pp. 910–915. IEEE CS Press, New Jersey (2011)
10. Wu, W., Zhang, B., Ostendorf, M.: Automatic generation of personalized annotation tags for Twitter users. In: HLT 2010, pp. 689–692. ACL, Stroudsburg (2010)
11. Yamaguchi, Y., Amagasa, T., Kitagawa, H.: Tag-based User Topic Discovery Using Twitter Lists. In: ASONAM 2011, pp. 13–20. IEEE CS Press, New Jersey (2011)
12. Michelson, M., Macskassy, S.A.: Discovering users' topics of interest on twitter: a first look. In: AND 2010, pp. 73–80. ACM, New York (2010)
13. Paulovich, F.V., Toledo, F.M.B., Telles, G.P., Minghim, R., Nonato, L.G.: Semantic Wordification of Document Collections. Computer Graphics Forum 31, 1145–1153 (2012)
14. Cui, W., Wu, Y., Liu, S., Wei, F., Zhou, M.X., Qu, H.: Context-Preserving, Dynamic Word Cloud Visualization. IEEE Computer Graphics and Applications 30, 42–53 (2010)
15. Rivadeneira, A.W., Gruen, D.M., Muller, M.J., Millen, D.R.: Getting our head in the clouds: toward evaluation studies of tagclouds. In: CHI 2007, pp. 995–998. ACM, New York (2007)
16. Lohmann, S., Ziegler, J., Tetzlaff, L.: Comparison of Tag Cloud Layouts: Task-Related Performance and Visual Exploration. In: Gross, T., Gulliksen, J., Kotzé, P., Oestreicher, L., Palanque, P., Prates, R.O., Winckler, M. (eds.) INTERACT 2009 Part I. LNCS, vol. 5726, pp. 392–404. Springer, Heidelberg (2009)
17. Yu, L., Asur, S., Huberman, B.A.: What Trends in Chinese Social Media. arXiv:1107.3522v1 (2011)
18. A stacked model based on word lattice for Chinese word segmentation and part-of-speech tagging, `http://nlp.csai.tsinghua.edu.cn/thulac`
19. Toutanova, K., Klein, D., Manning, C.D., Singer, Y.: Feature-rich part-of-speech tagging with a cyclic dependency network. In: NAACL 2003, pp. 173–180. ACL, Stroudsburg (2003)
20. Liu, Z., Chen, X., Sun, M.: Mining the interests of Chinese microbloggers via keyword extraction. Frontiers of Computer Science in China 6, 76–87 (2012)
21. Joachims, T.: Optimizing search engines using clickthrough data. In: KDD 2002, pp. 133–142. ACM, New York (2002)
22. Halvey, M.J., Keane, M.T.: An assessment of tag presentation techniques. In: WWW 2007, pp. 1313–1314. ACM, New York (2007)
23. Chang, C., Lin, C.: LIBSVM: A library for support vector machines. ACM Transactions on Intelligent Systems and Technology 2, 1–27 (2011)
24. Fan, R., Chang, K., Hsieh, C., Wang, X., Lin, C.: LIBLINEAR: A Library for Large Linear Classification. Journal of Machine Learning Research 9, 1871–1874 (2008)
25. Joachims, T.: Training linear SVMs in linear time. In: KDD 2006, pp. 217–226. ACM, New York (2006)