Summarizing Microblogging Users with Existing Well-defined Hashtags

Shuangyong Song, Yao Meng, Zhongguang Zheng
Internet Application Laboratory, Fujitsu R&D Center Co., Ltd.
Ocean International Center, No.56, Dong Si Huan Zhong Rd,
Chaoyang District, Beijing 100025, China
{shuangyong.song, mengyao, zhengzhg}@cn.fujitsu.com

Abstract

The rapid increasing popularity of microblogging has made it an important information seeking platform, and the typical way for a user to get useful information is by following those whose tweet content can draw interest. However, it is not practical to read all the tweets in order to decide whether to follow a user or not. Therefore, a brief and effective user description method is required, which is the focus of this paper. We design a model to automatically detect the most representative hashtags of a microblogging user, with which s/he can be easily known by others. The experimental results on Sina-Weibo, one of the most popular micro-blogging sites in China, show that our model can achieve a better performance than several baseline methods.

Keywords

Microblogging, well-defined hashtags, user summarization, user description, user interest detection, user interest ranking.

1. Introduction

Social media have shown increasing significance in our daily life, catering to different users for various purposes. In particular, microblogging serves as one of the most popular social media platforms where users can write up their status, utter opinions on some trending topics, and follow others to be acquainted with news or information of their interest (Java et al. 2007). Typically, a microblogging user can be updated with real time information published by those in follow, and a decision whether to follow a user is usually based on this user's tweet content in that people usually locate friends who post information relevant to their interests (Han and Lee, 2014). However, reading a dozen of tweets published recently is insufficient for a full understanding of a user, while a thorough reading is too time-consuming. Therefore, a convenient method that enables a concise and comprehensive description of each microblog user, is required for others to quickly grab a user's main interests.

Wu et al. (2010) proposed a model to generate personalized annotation tags for twitter users, which can create a simple user description effectively; however, tags such as single words or phrases usually have low readability and are prone to ambiguities without context. Sentence ranking techniques in text summarization area can also be used to get user description results (Wolf and Gibson, 2004), but this kind of techniques are unlikely to achieve the expected results in a bunch of tweets rather than a single long text in consideration of that top ranking tweets may have high semantic repeatability and just contain a user's one or two main interests. To solve these problems, in this paper we propose a model to automatically detect the most the representative hashtags of a user that can fully reflect his/her interests.

A particular and important feature of microblogging is the hashtag, a short-hand convention adopted by microblogging users to manually assign their posts to a wider corpus of messages on the same topic. They are denoted with a short string between two # symbols, often a name or abbreviation (Carter et al. 2011), such as '#the big bang theory#', '#Michael Jackson#' and '#computer hardware engineering#'. Hashtags are used to be brief topical markers, and they are usually adopted by users that contribute similar content or express a related idea. In essence, hashtags can highlight the topic of tweets, and make the tweets be easily searched and understood by other users (Bao et al. 2013). Therefore, many users are keen to add hashtags for their tweets.

In this paper, we detect the most related hashtags for each user which can be easily understood by others. In view of the arbitrariness of hashtags which may lead to mess and uneven quality (She and Chen, 2014), we utilize two effective factors to filter out hashtags of low quality and get 'well-defined' hashtags of high quality. Based on our previous research, hashtags' user acceptance degree and development tendency are two important factors for evaluate the quality of them, with which we can get some 'well-defined' hashtags (Song et al. 2015). A summarization of a user with several well-defined hashtags can be a concise and comprehensive description of him/her.

Generally, a user may change his/her interests over time, which is observable by reading content of his/her tweets (Song et al. 2012). Although some existing news categories are available for classifying abounding tweets of different users (Han and Lee, 2014), they are not reliable to analyze a single user's interest. For example, a songwriter may publish tweets about songwriting, singing and reviews on other singers, and it is inappropriate to classify all those content into the 'song' category. Therefore, our model first detects a user's interests on a proper granularity by clustering his/her tweets, and then ranks those interests based on tweets' quantity and novelty, finally detects topic-related well-defined hashtags for each interest to describe the user.

The remaining of this paper is organized as follows: In section 2, we provide a brief review of the related work. The introduction of our method is proposed in section 3. In section 4, some analysis of our experimental results is given. Finally we make a conclusion and discuss our plans for future work in section 5.

2. Related Work

Microblogging content analysis has attracted a number of researchers' interest. Nevertheless, to our knowledge, no previously research has addressed the issue of tagging a user's personal interests with most related well-defined hashtags to his/her microblogging content. Our work is related to the research on detecting the accurate interest of microblogging users. In this section, we discuss those related work.

Han and Lee (2014) propose an approach that estimates user interest from social media. They employ heterogeneous media to map microblogging user's tweets into categories, with a hybrid method that integrates a topic model with *TF-ICF* for extracting both explicitly presented and implicitly latent features. Wu et al. (2010) designed a system to automatically extract keywords from tweets to tag microblogging users' interests and concerns. Although this system can effectively summarize users' interests, automatically extracted keywords have low readability, such as 'tech', 'video' and 'stars'. Song et al. (2014) improve the aforementioned interest detection methods by considering keyphrases instead of keywords, which renders a better readability but some noisy phrases cannot be avoided. Bao et al. (2013) suggest a temporal and social probabilistic matrix factorization model to predict users' potential interests in micro-blogging by describing a user's interest with the hashtags used in his/her tweets. Unfortunately only 20% of the tweets are with hashtags based on our incomplete statistics, so this method may miss a majority of the useful information. Lim and Datta (2012) point to user following links as a key factor to reflect microblogging users' interest, and by calculating similarity between users to detect twitter communities with common interests. Obviously this method is not able to improve the detection of user interest from tweets. In this paper, our goal is to detect the most relevant hashtags to a user's tweets as an exact description of his/her interest which should have higher readability than automatically extracted keywords or keyphrases.

3. Proposed Model

3.1. System Architecture of the Proposed Model

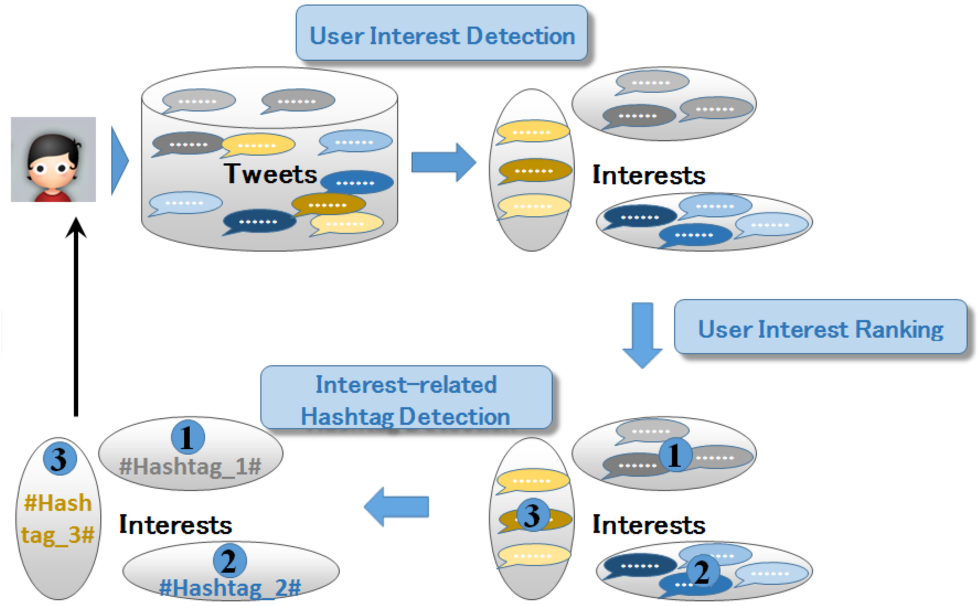

Figure 1: System architecture of the proposed model.

In this subsection, we present the architectural design of our proposed microblogging user summarization model. The overview of the system architecture is shown in Figure 1, which consists of three functional modules, namely, user interest detection, user interest ranking and interest-related hashtag detection.

In user interest detection module, we design a clustering method with self-adaptive threshold for the interest detection task in order to solve the ‘multi-granularity problem’ of user interest as aforementioned. In user interest ranking module, quantity and novelty of tweets are accounted to evaluate the degree of a user’s specific interests, and then all the detected interests are ranked based on their degree values. In interest-related hashtag detection module, we consider two factors to evaluate the recommendation value of a hashtag to a detected user interest: one is the quality of a hashtag and the other is the semantic similarity between the hashtag and the user interest. The mechanism of each functional module in our proposed model will be explored in details in the following subsections.

3.2. User Interest Detection

Generally, a user may have a variety of interests, and different users may have varying interest granularity. For example, a user may have interest in sports, movies and games, and we need to extract tweets from each of those domains to describe his/her interest, while another user may just be interested in sports, so we need to extract tweets about different types of sports to describe him/her, such as sports stars, sports lotteries and sports gossip. Therefore, we propose a self-adapting clustering method, which can detect a user's interest based on his/her different interest granularity.

We first apply *ICTCLAS* system (Zhang et al. 2003) to perform Chinese word segmentation on the tweet corpus, then utilize *JGibbLDA* version topic model (Phan et al. 2008) to realize the dimension reduction in word vectors of tweets because the 'tweet - word' matrix is sparse and difficult to be well analyzed (Bao et al. 2013). *JGibbLDA* is a Java implementation of *Latent Dirichlet Allocation* (Blei et al. 2003) using Gibbs sampling for parameter estimation and inference. In our paper, we empirically set the number of topic as 50, and the maximum number of iterations is 100. Then we design a clustering method with self-adaptive threshold for the interest detection task. The clustering threshold δ for a user in our method is defined in formula (1) as 'average *Euclidean distance* value of all his/her tweet couples', which is based on his/her own unique 'interest distribution granularity'.

$$\delta = w * \frac{\sum_{i=1}^{x-1} \sum_{j=i+1}^{x} Ed(V(m_i), V(m_j))}{x * (x-1)/2} \tag{1}$$

In formula (1), assuming that a user has published x tweets, x*(x-1)/2 represents the total number of 'tweet couple'. w is a weight parameter, which we empirically set as 0.9 (Song and Meng, 2015). $V(m_i)$ and $V(m_j)$ are the topic vectors of tweets m_i and m_j, and $Ed(V(m_i),V(m_j))$ means *Euclidean distance* between $V(m_i)$ and $V(m_j)$. After calculating δ, we start the clustering with a random tweet: assuming that it is a cluster and if the *Euclidean distance* between a new tweet and this cluster is smaller than δ, we put the new tweet into this cluster, otherwise we assign the new tweet to a new cluster. Consequently, all the tweets will be assigned to a steady cluster in the end, and we use $S(c_k)$ to denote each set of tweets in a cluster c_k. The process of user interest detection with our model is described in Algorithm 1.

Algorithm 1. The process of detecting users' interests

```
Input:
dataset: $\{m_1, \ldots, m_x\}$;
dimension reduction model;
weight parameter: $w$;
random seed: 1.
Output:
the detected user interests, all the $S(c_*)$
Clustering:
get Clustering threshold $\delta$ with $w$.
initialize $Cluster\_number = 0$.
For $i = 1, \ldots, x$
  Step 1: vectorize $m_i$ as $V(m_i)$ with dimension reduction model.
    If $Cluster\_number = 0$
      set $Cluster\_number = 1$
      set $m_i \in c_1$
      e.g. add $m_i$ into $S(c_1)$
    End If
    Else
      boolean $if\_create\_new\_cluster$ = true;
      For $k = 1, \ldots, Cluster\_number$
        If $\exists\ (Ed(V(m_i),V(m_j)) \leq \delta)$ with $\forall\ (m_j \in c_k)$
          set $m_i \in c_k$
          e.g. add $m_i$ into $S(c_k)$
          set $if\_create\_new\_cluster$ = false;
          End For
        End If
      End For
      If $if\_create\_new\_cluster$ = true
        set $Cluster\_number = Cluster\_number + 1$
        set $m_i \in c_{Cluster\_number}$
        e.g. add $m_i$ into $S(c_{Cluster\_number})$
      End If
    End Else
End For
```

3.3. User Interest Ranking

In generally, two factors have been recommended to reflect a microblogging user's degree of interest on a topic. The first one is the number of his/her published tweets on a particular topic as reported previously (Macdonald and Ounis, 2008; Song et al. 2014). The other factor is the novelty of those topic-related tweets because a user's recent tweets are more contributive to reflect his/her interest in comparison to earlier tweets (Bao et al. 2013; Zheng et al. 2015). Therefore, we take into account both the number of tweets and their timestamps to measure a user's interest degree in a cluster. Formula (2) shows how we calculate the interest degree with those two factors:

$$I(c_k) = \sum_{i=1}^{T_k} \exp(-\frac{t_p - t_i}{\gamma}) \qquad (2)$$

where $I(c_k)$ means the given user's interest degree in the k^{th} cluster (interest), and T_k is the total number of tweets in cluster c_i. t_p is the present time while t_i is the published time of the i^{th} tweet. γ is the kernel parameter, deciding the speed of decaying. In this paper, we experimentally choose the most effective value for γ in the following section.

3.4. Interest-related Hashtag Detection

After clustering tweets of similar content together, we detect a key hashtag for each cluster to represent this user's interest. Assuming that there are a total of H hashtags in our hashtag pool, we denote each hashtag as h_n, where $1 \leq n \leq H$. We evaluate the 'representative degree' of a hashtag h_n to a user interest (cluster) c_k with three factors: the first one is the semantic similarity between h_n and c_k, which we denote as $Sim(h_n, c_k)$, and the other two factors are the user acceptance degree and development tendency of h_n, which have been reported in our previous work (Song et al. 2015) for evaluating a hashtag's quality. In this paper, we denote the user acceptance degree of h_n as $A(h_n)$, and denote the development tendency of h_n as $D(h_n)$.

Topic model has been widely used in microblogging hashtag recommendation (Ding et al. 2012), and we also use topic vectors as representations of tweets in our work. Furthermore, we utilize Cosine Similarity between semantic vectors of h_n and c_k to measure $Sim(h_n, c_k)$, while the semantic vector of h_n is obtained from the integration of the topic vectors of tweets which contain h_n, and the semantic vector of c_k is obtained from the integration of the topic vectors of tweets which are clustered into c_k. Then we denote the frequency of h_n as $N(h_n)$, and get the $A(h_n)$ as the logarithm of $N(h_n)$, e. g. $\log[N(h_n)]$. Finally, we get the $D(h_n)$ via the following steps: we utilize polynomial spline estimator (Huang et al. 2004) to estimate the time curve of h_n with its historical data, and then calculate the slope of the continuous curve of h_n on the given day, which we denote as $G(h_n)$, and use the *Sigmoid*

Function (Tarde, 1903) to normalize the slope value into a real number on the interval (0,1), which is the value of $D(h_n)$. After calculation of the three factors, $Sim(h_n, c_k)$, $A(h_n)$ and $D(h_n)$, the formula of the representative degree of h_n to c_k is given as below:

$$\begin{aligned} R(h_n,c_k) &= Sim(h_n,c_k)*(Q(h_n))^{\lambda} \\ &= Sim(h_n,c_k)*[A(h_n)*D(h_n)]^{\lambda} \\ &= CosSim[V(h_n),V(c_k)]*\left\{\log[N(h_n)]*\left(\frac{1}{1+e^{-G(h_n)}}\right)\right\}^{\lambda} \end{aligned} \tag{3}$$

In formula (3), $R(h_n, c_k)$ means the representative degree of h_n to c_k, and $Q(h_n)$ means h_n's quality. The parameter λ is used to adjust the significance of the $Sim(h_n, c_k)$ and $Q(h_n)$, where $0 < \lambda < \infty$. Then the hashtag with the highest value of $R(h_n, c_k)$ will be chosen as the key hashtag of c_k. Finally, we rank those key hashtags according to the ranking of the $I(c_k)$, and the top Z key hashtags will be taken as the final user description result.

The notations used throughout this paper are summarized in Table 1 for an easy reading and understanding this paper.

δ	Clustering threshold
x	Published tweets number of a given user
w	weight parameter for δ
m_i	The i^{th} tweet of given user, where $1 \le i \le x$
$V(m_i)$	The topic vectors of tweet m_i
$Ed(V(m_i),V(m_j))$	*Euclidean distance* between $V(m_i)$ and $V(m_j)$
c_k	The k^{th} cluster (interest) of the given user
$I(c_k)$	The given user's interest degree on the k^{th} cluster (interest)
$S(c_k)$	The set of tweets in c_k
T_k	The total number of tweets in the cluster c_k
t_p	The present time
t_i	The published time of the i^{th} tweet
γ	The kernel parameter, deciding the speed of decaying.
H	Total number of hashtags in our hashtag pool
h_n	The n^{th} hashtag in our hashtag pool, where $1 \le n \le H$
$Sim(h_n, c_k)$	The semantic similarity between h_n and c_k
$Q(h_n)$	The h_n's quality
λ	The parameter to adjust significance of $Sim(h_n, c_k)$ and $Q(h_n)$
$A(h_n)$	The user acceptance degree of h_n
$D(h_n)$	The development tendency of h_n
$G(h_n)$	The slope of the continuous curve of h_n on the given day
$N(h_n)$	Total number of usage of h_n
$R(h_n, c_k)$	The representative degree of h_n to c_k
Z	Number of top ranking hashtags which we choose as final result

Table 1: Summary of notations used in this paper.

4. Experiments

4.1. Experimental Data and Parameter Setting

Our dataset is obtained from *Sina Weibo*, a well-known Chinese microblogging service. It contains 290,638 tweets collected from *Sina Weibo* 'public timeline' API over the time period from 08:56, Jun 09, 2013 to 16:40, Nov 23, 2013. For evaluation experiment, we choose users who have published more than 100 tweets, from the friend list of the first author of this paper, in thatour model aims at a concise user description for those who has published a huge amount of tweets. The chosen users accounted for around 86.5% of the users in this friend list (77 users published more than 100 tweets and 12 users not), which partly indicates that the application range of our model is very wide in microblogging. For those users, the average number of tweets is around 1,012, and the average time-range is a little more than 3 years.

The parameter λ is used to adjust the significance of the two factors in formula (3). A value of λ equal to 1 indicates that $Q(h_n)$ and $Sim(h_n,c_k)$ have same impact on $R(h_n,c_k)$ of a hashtag, while a value of λ smaller than 1 indicates that $Q(h_n)$ has a smaller impact on the $R(h_n,c_k)$ than $Sim(h_n,c_k)$ does, and vice versa. Since $R(h_n,c_k)$ just reflects the precision of a recommended hashtag, we use the average precision value (V_{AP}) of hashtag with the highest $R(h_n,c_k)$ for each user interest as the reference for setting λ. The description of the V_{AP} is given in formula (4).

$$V_{AP} = \frac{1}{N(u)} \sum_{i}^{N(u)} \frac{1}{N(c,u_i)} \sum_{j=1}^{N(c,u_i)} B(c_j,u_i) \tag{4}$$

In formula (4), $N(u)$ means the total number of chosen users in our experiments, and $N(c,u_i)$ means the number of detected interests of user u_i. $B(c_j,u_i)$ means a binary distributions of the 'Interest-related Hashtag Detection' result for each user interest, while if the detected interest-related hashtag is the same as the manual tagged one, $B(c_j,u_i)$ is set to be 1, otherwise $B(c_j,u_i)$ is set to be 0.

Figure 2 shows the V_{AP} results when varying λ to seven different values from 0.1 to 10. As shown, the curve peaks at $\lambda = 0.5$, which indicates that $Sim(h_n,c_k)$ has larger impact on $R(h_n,c_k)$ than $Q(h_n)$ does. Therefore, we set $\lambda = 0.5$ in the following experiments.

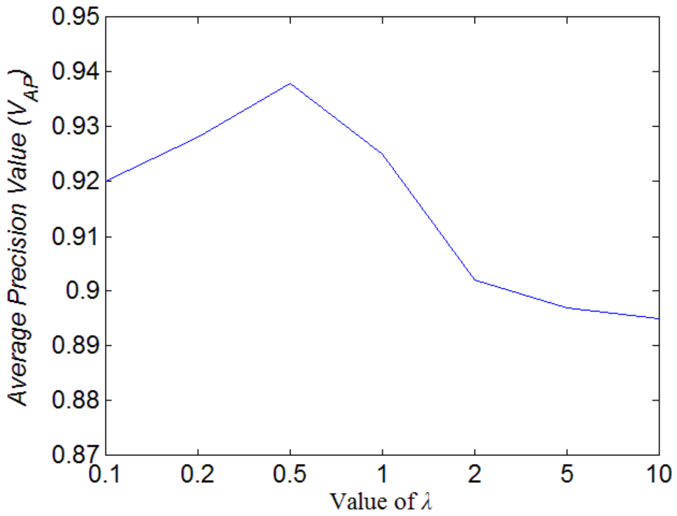

Figure 2: 'Average precision value (V_{AP})' results by different values of λ.

4.2. Evaluation Metrics and Baselines

To measure the effects of our model, we use *Precision*, *Recall* and *F-value* as the evaluation metrics. Meanwhile, five other user description methods are conducted in the same way as baselines:

Hashtag Frequency: ranking the hashtags which have been used in a user's tweets, with their frequency, and then choosing the top ranking *Z* key hashtags as the description of this user's interest. This method can be taken as an initial hashtag-based microblogging user summarization method;

Classification: using online news categories knowledge to classify users' tweets (Han and Lee, 2014) and then choosing one key hashtag from each of the *Z* most important categories as the user summarization result;

Annotation Tag: generating top *Z* personalized annotation words to label Twitter user's interests and concerns with a proposed *TextRank* method (Wu et al. 2010);

Sentence Ranking: using a *PageRank* based sentence ranking method, which is proved as the most effective sentence ranking method (Wolf and Gibson, 2004), to rank all tweets of a user and choose the top *Z* sentences as the user summarization result;

Latest Tweets: the latest published tweets are automatically shown in each user's homepage, and we take this description of user as a baseline: using latest *Z* tweets to represent a user, which is the same as the display mode of microblogging platforms (Song et al. 2012).

4.3. Gold Standard Generation

To evaluate the model, we have the 77 chosen users score the five model results with an integral score between 1 and 5, considering the *Precision* and *Recall* respectively. Then for each of the 77 users, we invite one of his/her microblogging friend who is also his real-life friend to score the results by considering the same factors. We get the value of *intra-class correlation coefficients* (*ICC*) (McGraw and Wong, 1996) between 'self-scoring' and 'friend scoring' as 0.908, which indicates the validity. Finally we use one fifth of the average score of 'self-scoring' and 'friend scoring' as the final *Precision* and *Recall*, furthermore get the final *F-value*.

4.4. Experimental Results and Discussions

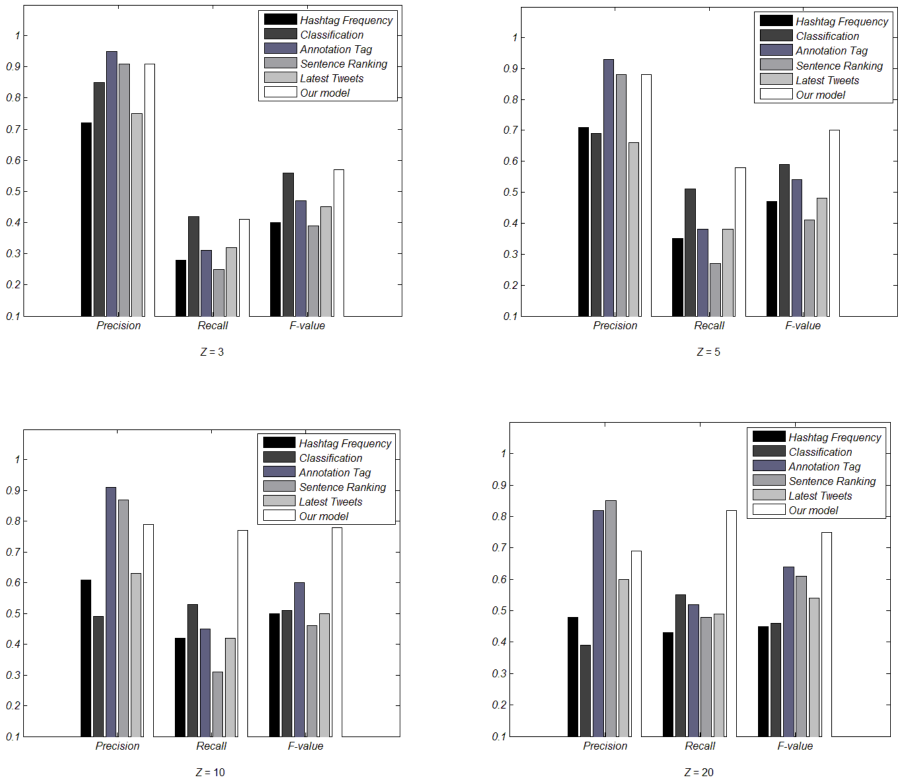

Figure 3: Result comparison of the five different models with *Precision*, *Recall* and *F*-value when Z = 3, 5, 10 and 20 respectively.

Figure 3 compares the evaluation results of our model and the five baseline methods, when Z is set as 3, 5, 10 and 20 respectively. *Annotation Tag* method made the highest *Precision* when *Z*=3, which is due to top ranking tags enable higher refinement than sentences. The *Precision* of our model is rarely lower than *Annotation Tag* method. This is because defined hashtags may have content disparity from the real tweets content despite of a higher readability. However, the *Recall* is not as good as the *Precision* for the *Annotation Tag* method because top ranking words may be in a same domain. For example, ‘dinner’, ‘coupon’, and ‘restaurant’ are detected as the top ranking tags of a user, while his other interests on movies and gossip are ignored. The *Sentence Ranking* method shows similar limitations.

Our model achieves the highest *F-value* when *Z* is 10, which suggests that each user may have as many as 10 different main interest points. Besides, our model exhibits the highest *Recall* on almost every *Z* as it is capable to improve the disparity of tweets in the final result. *Classification* method gives a higher *Recall* than other three methods when *Z* is 3 or 5, but this advantage reduces when *Z* is 10 or 20, in that it does not account for the granularity of different interests within the same domain. As for *Hashtag Frequency* method, although hashtags are also used as user description, similar to our model, the drawback is to summarize users who rarely use hashtags in his/her published tweets.

Table 2 gives some examples of user summarization result with their representative hashtags. Taking *user1* as an example, she is a female *IT* researcher. Although she published only a few hashtag-embedded tweets, we can summarize her with well-defined hashtags as per our proposed model. We can see that her main points of interest are *IT* news and enjoyment, which provides a brief and effective interest summary for potential followers. Overall, 4 of these 5 users have representative hashtags #Daily positive energy# or #Positive energy#, indicating that positive social events and news are popular among the microblogging service.

5. Conclusion and Future Work

In conclusion, we propose a novel model to facilitate a simple description of microblogging users, which is a convenient way for them to be easily known by others. The proposed model first detects and ranks user interest, and then extracts key hashtags from each interest. In our experiments, this method is shown to be effective to enhance the performance of user description in microblogging. To the best of our knowledge, no other work on user interest based representative hashtag extraction of microblogging users has been done to date.

In our future work, we plan to design a model to estimate the dynamic interest similarity between users, for recommending friends not only in microblogging but also in other types of social media, such as social networking sites and Vertical BBS. In addition, we will try to use word embedding technique to improve representation of hashtags.

Furthermore, the event-oriented user recommendation and hashtag-oriented expert finding are also part of our future work.

Users	Top 10 Representative Hashtags
user1	#IT news# (#IT 新闻#); #Cawayi# (#萌#); #Overseas purchasing# (#海外代购#); #Daily positive energy# (#每日正能量#); #Fun time# (#开心一刻#); #Talking to Beauty# (#与美对话#); #New film recommendation# (#电影推荐#); #Virgo# (#处女座#); #Music Bar# (#音乐吧#); #Phoenix Entertainment live# (#凤凰娱乐现场#).
user2	#Food stories# (#美食工场#); #Midnight canteen# (#深夜食堂#); #Love life love shopping# (#爱生活爱购物#); #Xinhua share News# (#新华分享#); #Thanks for your love# (#谢谢你的爱#); #Wallace Chung# (#钟汉良#); #Tour group# (#驴行团#); #Delicious food DIY# (#美食 DIY#); #Daily positive energy# (#每日正能量#); #Pregnancy# (#怀孕#).
user3	#Microblogging hotspots# (#微热点#); #Positive energy# (#正能量#); #I love American TV series# (#我爱看美剧#); #Enjoy colorful style# (#享受多彩 style#); #Dad where are we going# (#爸爸去哪儿#); #Beauty & Cosmetics# (#美容化妆#); #Walking photograph# (#随手拍#); #New balance# (#新百伦#); #Fashion# (#时尚#); #Thanksgiving Day# (#感恩节#).
user4	#Graduation Season# (#毕业季#); #Christmas Eve# (#圣诞夜#); #Last 8 days of 2013# (#2013 年最后 8 天#); #Old photos# (#老照片#); #Fast & Furious# (#速度与激情#); #Microblogging public welfare# (#微公益#); #Daily positive energy# (#每日正能量#); #Gucci# (#Gucci#); #Girls Generation# (#少女时代#); #Music Bar# (#音乐吧#).
user5	#Autohome microblogging# (#汽车之家官方微博#); #Self-guided tour in Korea# (#韩国自由行#); #No Make Up contest# (#素颜大赛#); #Phoenix Entertainment live# (#凤凰娱乐现场#); #Bangkok shopping# (#曼谷购物#); #Cawayi# (#萌#); #New media# (#新媒体#); #Spy photos of new cars# (#新车谍照#); #Good morning# (#早安#); #Let's immomo# (#陌陌吧#).

Table 2: Examples of user summarization result with representative hashtags (translated into English).

6. References

Bao, H., Li, Q., Liao S. S., Song, S., Gao, H. 2013. A New Temporal and Social PMF-based Method to Predict Users' Interests in Microblogging. *Decision Support Systems*, 55(3): 698-709.

Blei, D. M., Ng, A. Y., Jordan, M. I. 2003. Latent Dirichlet Allocation. *Journal of Machine Learning Research* 3: 993-1022.

Carter, S., Tsagkias, M., Weerkamp, W. 2011. Twitter hashtags: Joint Translation and

Clustering. In *Proceedings of the 3rd International Conference on Web Science*, pp. 1-3.

Ding, Z., Zhang, Q., Huang, X. 2012. Automatic Hashtag Recommendation for Microblogs using Topic-Specific Translation Model. In *Proceedings of the 2012 International Conference on Computational Linguistics*, pp. 265-274.

Han, J., Lee, H. 2014. Characterizing user interest using heterogeneous media. In *Proceedings of the 23rd International World Wide Web Conference*, pp. 289-290.

Huang, J. Z., Wu, C. O., Zhou, L. 2004. Polynomial spline estimation and inference for varying coefficient models with longitudinal data, *Statistica Sinica*, 14: 763-788.

Java, A., Song, X., Finin, T., Tseng, B. 2007. Why we twitter: understanding microblogging usage and communities. In *Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 Workshop on Web Mining and Social Network Analysis*, pp. 56-65.

Lim, K. H., Datta, A. 2012. Following the follower: detecting communities with common interests on twitter. In *Proceedings of the 23rd ACM Conference on Hypertext and Social Media*, pp. 317-318.

Macdonald, C., Ounis, I. 2008. Voting Techniques for Expert Search. *J. Knowledge and Information Systems*, Volume 16, Number 3, September 2008, pp. 259-280.

McGraw, K. O., Wong, S. P. 1996. Forming inferences about some intra-class correlation coefficients. *Psychological Methods*, 1: 30-46.

Phan, X-H., Nguyen, L-M., Horiguchi, S. 2008. Learning to classify short and sparse text & web with hidden topics from large-scale data collections. In *Proceedings of the 17th International World Wide Web Conference*, pp.91-100.

She, J., Chen, L. 2014. TOMOHA: TOpic MOdel-based HAshtag Recommendation on Twitter. In *Proceedings of the 23st International World Wide Web Conference*, pp. 371-372.

Song, S., Li, Q., Bao, H. 2012. Detecting dynamic association among twitter topics. In *Proceedings of the 21st International World Wide Web Conference*, pp. 605-606.

Song, S., Meng, Y., Sun, J. 2014. Detecting Keyphrases in Micro-blogging with Graph Modeling of Information Diffusion. In *Proceedings of the 13th Pacific Rim International Conference on Artificial Intelligence*, 26-38.

Song, S., Meng, Y. 2015. Detecting representative tweets of micro-blogging users, In *Proceedings of the Eighth International C* Conference on Computer Science & Software Engineering*, pages 110-112.

Song, S., Meng, Y., Zheng, Z. 2015. Recommending Hashtags to Forthcoming Tweets in Microblogging. In *Proceedings of the 2015 IEEE International Conference on Systems, Man, and Cybernetics*.

Tarde, G. 1903. The laws of imitation. *New York: Henry, Holt and Co.*

Wolf, F., Gibson, E. 2004. Paragraph, word, and coherence-based approaches to sentence ranking: a comparison of algorithm and human performance. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics*, pp. 383-390.

Wu, W., Zhang, B., Ostendorf, M. 2010. Automatic generation of personalized annotation tags for twitter users. In *Proceedings of the Conference of the North American Chapter of the Association of Computational Linguistics*, pp. 689-692.

Zhang, H., Yu, H., Xiong, D., Liu, Q. 2003. HHMM-based Chinese lexical analyzer ICTCLAS. In *Proceedings of the second SIGHAN workshop on Chinese language processing*, pp. 184-187.

Zheng, N., Song, S., Bao, H. 2015. A Temporal-Topic Model for Friend Recommendations in Chinese Microblogging Systems. *IEEE Trans. on Systems, Man & Cybernetics: Systems*, Volume 45.