Clustering Users in Twitter Based on Interests

Abstract

Twitter has recently emerged as a popular social microblogging service. There are over 100 million users in Twitter nowadays, little is known yet about Twitter at user level. In this paper, we investigate the problem of clustering users in Twitter based on their interests. Solving this problem is very important in many different fields, such as user recommendation, personalized services, viral marketing, etc. To address this problem, we first compute user similarity leveraging both textual contents and social structure, according with Twitter's role, not only a news media but also a social network. These features include tweet text, URLs, hashtags, following relationship and retweeting relationship, all of them are closely correlated with user's interests. Then we use user similarity as a measure to cluster users. To assess effectiveness of our method, we propose the clustering metrics "average number of mutual following links per user in per cluster" in Twitter. Experimental results show that our method can successfully cluster users in Twitter, and performs much better than random selection. From a side view, our experiment also shows that users in our dataset of Twitter can be approximately categorized into 400 clusters.

Keywords: Twitter; Clustering; User; User similarity;

1. Introduction

Twitter has recently emerged as a popular social microblogging service where users share and discuss about everything, including news, jokes, what they are going through, and even their mood. Nowadays there are over 100 million users in Twitter, and almost 200 million messages are posted every day. Although Twitter is very popular, little is known yet about it at user level. So in this paper, we investigate the problem of clustering users based on their interests in Twitter.

The benefits of solving this problem are multifold. First, clustering users can be used in user recommendation as well as tweet recommendation; second, some secondary services such as real-time search engines and trend analysis, will probably evolve different personalized services [1]; third, clustering users is also crucial in viral marketing, for example, online marketers probably can accurately post different advertisements to different groups of users. Despite that solving this problem is very important in many different fields, this paper is the first special study about clustering users in Twitter to the best of our knowledge.

In this paper, we first compute the similarity between users based on their interests, and then use it as a measure to cluster users in Twitter. With the purpose of computing user similarity, we leverage different features which reflect one's interests, including both textual contents and social structure. The textual content in Twitter mainly encompasses three features, tweet text, URLs and hashtags embedded in tweets. As we know, tweet text contains rich information about author of the tweet, such as what the user is talking about, what the user is going through and so on, so tweet text is potentially useful in determining interests of an individual user. The second useful textual feature is URLs embedded in tweets. Boyd et al. [2] have noted that 22% of tweets contain a URL, and due to the limitation of 140 characters, more and more users prefer to share information by the use of URLs. Another textual feature is hashtags, which are frequently used to create and follow topics, and users that shared common hashtags intuitively have similar interests. The social structures on Twitter mainly refer to two features, following relationship and retweeting relationship, which have bean proved to be closely correlated with user's interests [3, 4]. Weng et al. [3] have noted that users follow each other based on shared interests, and following relationship is a strong indicator of similar interests among users. Welch et al. [4] noted that retweeting relationship is more closely correlated with topical similarity of user interests, because a user would not retweet something easily, except that he/she likes or approves it very much.

The remainder of this paper is organized as follows. We review prior work in Twitter in Section 2. In Section 3, we provide an overview of collected data and demonstrate the particular characteristics the data takes on at user level. In Section 4, we elaborate our method to cluster users based on textual content and social structure in detail. The experiment results of our method are described in Section 6. We get the conclusion and describe some future work in Section 7.

2. Related work

Twitter plays an important role in many different fields, such as politics, marketing, emergencies and even our daily life [5, 6, 7, 8]. Tumasjan et al. [5] found that Twitter was indeed used extensively for political deliberation and messages in Twitter validly mirrored the offline political sentiments. Bollen et al. [6] showed that collective mood states derived from Twitter was correlated to the value of the Dow Jones Industrial Average over time. Sakaki et al. [7] successfully approximated the epicenter of earthquakes in Japan by treating Twitter users as a geographically-distributed sensor network. Zhao et al. [8] pointed out that people in Twitter not only updated their daily life activities with friends but also shared information with interested observers.

Given the popularity and importance of Twitter, plenty of researchers have studied the characteristics of Twitter [2, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19]. Java et al. [9] presented a study focused on examining the topological and geographical properties of twitter’s social network. Kwak et al. [10] conducted a large-scale study to analyze the topological characteristics of Twitter and its power as a new medium of information sharing. Grier et al. [11] and Lee et al. [12] investigated the problem of spam detection in Twitter. Suh et al. [13] studied how and why certain information spreads more widely than others, Petrovic et al. [14] further studied the problem whether a tweet will spread widely and Hong et al. [15] predicted popular messages in Twitter. Boyd et al. [2] analyzed the mechanism retweet in detail and Nagarajan et al. [16] gave a qualitative examination of retweet practices. Ramage et al. [17] characterized tweets using hashtags, emotions as well as social structure, Galuba et al. [18] further characterized the propagation of URLs and predicted the information cascades in Twitter. Cao et al. [19] detected near duplicate messages in Twitter. We note that, although Twitter is currently a hot subject of research, few researchers have investigated characteristics of Twitter at user level.

Additionally, most of these study about users in Twitter is limited in the problem of finding “influencials” in Twitter [3, 4, 20], with other tasks overlooked. Cha et al. [20] measured user influence by the use of indegree and retweets. Weng et al. [3] improved topic-sensitive pagerank, leveraging following relationships between Twitter users as well as topical similarity distilled from tweets of users. Welch et al. [4] improved what Weng et al. have done, using retweeting relationships instead. In this paper, we propose the problem of clustering users in Twitter, with the purpose of understanding Twitter better at user level.

3. Data Description

In order to cluster users in Twitter, we use Twitter’s Developer API to collect user data. We keep an eye on 116846 unique users, discarding the non-English or inactive users, and finally get 45772 English users who have posted at least 100 tweets and have more than 20 friends.

Boyd et al. [2] have shown that only 5% of tweets contain a hashtag, and 22% of tweets include a URL. Suh et al. [13] conducted a large-scale analysis on Twitter and noted that 11.2% of tweets were retweets. However, they both investigated the data at tweet level. We give a close look on active users and find some important phenomenon at user level as follows:

1. The proportion of users, half of whose tweets contain at least a URL, is 47.5%, and only 0.2% of users have never used URLs in their tweets.

2. The proportion of users, 25% of whose tweets encompass a hashtag at least, is 40.9%, and only 8.3% of users have never used hashtags in their tweets.

3. 32.1% of users use the mechanism retweet in their 30% of tweets or more, and 13.2% of users have never retweeted others.

These findings show that there is a very widely use of URLs, hashtags and retweets at user level, and prove that it is necessary to take these features into account when computing user similarity.

4. Clustering users

In this section, we elaborate our method in detail. In section 4.1 and 4.2, we compute user similarity by the use of both textual contents and social structure respectively, including tweet text, URLs, hashtags, following relationship and retweeting relationship. Section 4.3 aggregates these feature similarities together and gets the final user similarity. Section 4.4 uses the final user similarity as well as classic clustering algorithms to cluster users.

4.1. Textual contents

Textual contents in Twitter mainly refer to tweet text, URLs, and hashtags. All of them are closely correlated with user's interests.

4.1.1. Text Similarity

The goal of this section is automatically identifying the topics that users are interested in based on the tweets they published and computing tweet text similarity based on these topics. We should note that the tweets here don't encompass any URLs or hashtags because they will be handled particularly later. To avoid the problem of small size of a single tweet and the sparseness of data, we aggregate the tweets published by an individual user into a big document. Thus, each document essentially corresponds to a user, and finding topics that users are interested in just means finding latent topics in these documents. With this purpose, Latent Dirichlet Allocation (LDA) model [21] is applied, which is an unsupervised machine learning method to identify latent topics from large document collections.

As a generative probabilistic model, LDA generates each document as follows: First, pick a topic from its distribution over topics for each document; second, sample a word from the chosen topic's distribution; finally, repeat the two above processes for all the words in the document. Thus, each document is represented as a probability distribution over some topics, while each topic is represented as a probability distribution over a bag of words. The graphical representation of these generative processes is shown in Figure 1. In this figure, the boxes are "plates" representing replicates. The outer plate represents documents, while the inner plate represents the repeated choice of topics and words within a document. The variable in the bottom of box represents the number of repetitions. θ and

ϕ represent the distribution over topics and the distribution over words for each topic respectively. α and β are hyper-parameters related with θ and ϕ. A topic *z* is sampled from the multinomial distribution θ, and a word ω from the multinomial distribution ϕ associated with topic *z* is sampled consequently.

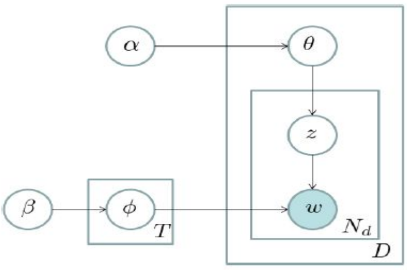

Fig.1 Graphical model of LDA

There are a lot of methods to infer parameters in models. In this study, Gibbs sampling is applied for model parameter estimation. Experientially, the number of latent topics is set to 200, α and β are set to 0.25 and 0.1. Using GibbsLDA++ [22], we get the results as follows:

1. *UT*, a $U \times T$ matrix, where *U* is the number of users and *T* is the number of topics. UT_{ij} captures the probability that user u_i has the latent topic t_j.

2. *WT*, a $W \times T$ matrix, where *W* is the number of total unique words occurring in the documents, and *T* is the number of topics. WT_{ij} means the probability that the word ω_i has been assigned to topic $t_{j.}$

3. *Z*, a $1 \times N$ vector, where N is the total number of words in the tweets. Z_i is the topic assigned for word ω_{i}.

When coming to compute user similarity in Twitter, matrix *UT* is of particular importance among the three matrixes mentioned above. We define text similarity between users as follows:
Definition 1. Text similarity between two users u_i and u_j can be calculated as:

$$sim_{text}(i,\ j) = 1 \Big/ \sqrt{D_{js}(i,\ j)} \tag{1}$$

$D_{js}(i, j)$ is the Jensen-Shannon Divergence between the two users' probability distributions $UT_{i.}$ and $UT_{j.}$, and is defined as:

$$D_{js}(i,j) = \frac{1}{2}(D_{kl}(UT_{i.} \| M) + D_{kl}(UT_{j.} \| M)) \tag{2}$$

Where $M = \frac{1}{2}(UT_{i.} + UT_{j.})$ *, and D_{kl} is the Kullback-Leibler Divergence which defines the divergence from distribution Q to P as:* $D_{kl}(P \| Q) = \sum_i P(i) \log \frac{P(i)}{Q(i)}$.

4.1.2. URL Similarity

The second feature we use to compute user similarity is URLs embedded in tweets. A minor difficulty here is the common use of URL shortening services such as bit.ly, TinyURL, etc. This prevents making use of keywords or other interesting artifacts the URL may contain directly, and makes additional processing of the data necessary. Fortunately, with help of *LongUrl*[1], we can get the expanded URL as well as some metadata such as titles and descriptions for each shortened URL. And then we aggregate all these titles of a user's URLs into a document, corresponding to the user. Finally, similar with section 4.1.1, we compute URL similarity sim_{url} using these documents.

4.1.3. Hashtag Similarity

Hashtag is a convention to create or follow a thread of discussion, which is similar with tags in website *del.icio.us*[2]. Between two users, we measure hashtag similarity based on the number of their common hashtags and the importance of these hashtags. Additionally, given that even the same hashtag may have different weights in these two users, we define hashtag similarity between users as follows:

Definition 2. Hashtag similarity between users u_i and u_j can be calculated as:

$$sim_{hashtag}(i,j)=\sum_{k=1}^{n}\left(1-\left|\frac{N_{ik}}{|H_i|}-\frac{N_{jk}}{|H_j|}\right|\right)\frac{N_{ik}+N_{jk}}{|H_i|+|H_j|} \quad (3)$$

Where $|H_i|$ is the total number of hashtags published by u_i, n is the number of hashtags that appear both in u_i and u_j, N_{ik} represent the number of the kth hashtag in user u_i . In Eq (3), $\left|\frac{N_{ik}}{|H_i|}-\frac{N_{jk}}{|H_j|}\right|$ represent the difference of the kth hashtag's weights in u_i and u_j, and $\frac{N_{ik}+N_{jk}}{|H_i|+|H_j|}$ represents the the kth hashtag's whole weight in the two users.

Generally, the bigger n is (the two users have more common hashtags), the smaller $\left|\frac{N_{ik}}{|H_i|}-\frac{N_{jk}}{|H_j|}\right|$ is (the importance of the hashtag in two users are more similar), and the bigger $\frac{N_{ik}+N_{jk}}{|H_i|+|H_j|}$ is (the more important the hashtag is between these two users), the bigger hashtag similarity $sim_{hashtag}(i,j)$ is.

4.2. Social Structure

When mentioning the social structure in Twitter, we mainly refer to following and retweeting relationships here. Weng et al. [3] have noted that users in Twitter don't follow people randomly or reciprocally. Namely, a twitterer follows a friend because she/he is interested in the topics the friend publishes in tweets, and the friend followed back because she/he finds they share similar topic interest. This phenomenon is called "homopyily", which demonstrates that following relationship is closely correlated with users' interests. Welch et al. [4] further noted that retweeting relationship can reflect users'interests better. In this section, we mainly investigate following and retweeting relationship between users, with the purpose of computing sim_{follow} and $sim_{retweet}$.

4.2.1. Following Similarity

The goal of this section is to compute the following similarity between users using following structure. Intuitively, if two users have many common friends and followers, they are quite similar. So

[1] http://longurl.org/

[2] http://www.delicious.com/

we take these two factors into account, and define following similarity between users as follows:

Definition 3. Following similarity between users u_i and u_j can be calculated as:

$$sim_{follow}(i,j) = \frac{c_{friend}}{\sqrt{|Friend_i|}\sqrt{|Friend_j|}} + \frac{c_{follower}}{\sqrt{|Follower_i|}\sqrt{|Follower_j|}} \quad (4)$$

$|Friend_i|$ is total number of the users that u_i follows, $|Follower_i|$ is total number of the users that follow u_i. c_{friend} represents number of the two users' common friends, while $c_{follower}$ represent number of the two users' common followers. F is a n × n matrix, where n is total number of users in data.

4.2.2. Retweeting Similarity

As mentioned before, if two users retweet the same person frequently, the two users may have similar interests. Additionally, whether the two users retweet each other is a stronger indicator of similar interests. With these two factors into consideration, we define retweeting similarity between users as follows:

Definition 4. Retweeting similarity between users u_i and u_j can be calculated as:

$$sim_{retweet}(i,j) = \frac{c_{retweet}}{\sqrt{|R_i|}\sqrt{|R_j|}} + \frac{n_{ij}+n_{ji}}{|R_i|+|R_j|} \quad (5)$$

$|R_i|$ is the number of users whom u_i retweet, $c_{retweet}$ is the number of users that u_i and u_j both retweet. n_{ij} represent the number of times that u_i retweet u_j while n_{ji} represent the number of times that u_j retweet u_i.

4.3. Aggregation

The definitions presented in Section 4.1 and 4.2 compute user similarity using different features respectively. In this section, we aggregate these feature similarities together to get final user similarity. We define final user similarity between users as follows:

Definition 5. The final similarity between users u_i and u_j can be calculated as:

$$sim(i,j) = \gamma_t sim_{text}(i,j) + \gamma_u sim_{url}(i,j) + \gamma_h sim_{hashtag}(i,j) + \gamma_f sim_{follow}(i,j) + \gamma_r sim_{retweet}(i,j) \quad (6)$$

$\gamma_t, \gamma_u, \gamma_h, \gamma_f, \gamma_r$ are parameters between 0 and 1 to control the weights of different feature similarities, and $\gamma_t+\gamma_u+\gamma_h+\gamma_f+\gamma_r=1$. The bigger sim(i, j) is, the more similar the two users are.

4.4. Clustering algorithms

We use the final user similarity computed in last section as well as classical clustering algorithms to cluster users in this section. There are two well known cluster algorithms: hierarchical clustering and *k*-means. Hierarchical algorithm is too slow to handle large-scale dataset such as users in Twitter, whereas *k*-means is not only effective but also very fast. So in this paper, we apply *k*-means to cluster users in Twitter. A minor difficulty in applying *k*-means is the selection of parameter *k*, which decides the number of clusters and influences the final clustering effectiveness. We will talk about how to select appropriate *k* later particularly.

5. Experimental results

5.1. Evaluation metrics

As mentioned above, because of homopyily phenomenon in Twitter, users following each other probably share common interests. Given that the “mutual following relationship” is a strong indicator of user’s similarity, intuitively, to assess the effectiveness of our approach, we propose the evaluation metrics “the average number of mutual following links per user in per cluster (FPUPC)”. A bigger FPUPC value means that the users in the same cluster are more similar and the users in different clusters are less similar, namely clustering results are better.

5.2. Selecting aggregation parameters

As mentioned in section 4.3, we use $\gamma_t, \gamma_u, \gamma_h, \gamma_f, \gamma_r$ to adjust the influence of different features in the final user similarity. To select the parameters for different features, we first apply each feature similarity in the clustering task respectively, and then select appropriate parameter for each feature based on the cluster results. If a feature performs better than the others, the aggregation parameter of this feature will be bigger than others.

Tab.1 FPUPC values for each single feature in clustering task

Features	text	URL	hashtag	following	retweeting
FPUPC values	0.056	0.04	0.009	0.053	0.061

Table 1 shows experimental results using different feature similarity respectively in the form of FPPUPC values. We note that retweeting relationship have a greatest impact on clustering, whereas the feature hashtag plays a least important role. The aggregation parameters for features $\gamma_{feature}$ are defined as follows:

$$\gamma_{feature} = FPUPC_{feature} / \sum_{all\ features} FPUPC_{feature} \tag{7}$$

Finally, $\gamma_t, \gamma_u, \gamma_h, \gamma_f, \gamma_r$ are assigned as 0.25, 0.18, 0.04, 0.24 and 0.29 respectively.

5.3. Selection of k

The parameter *k* in *k*-means means the expected number of clusters and the selection of *k* is crucial to final clustering results. Figure 2 shows the comparison between our method and random clustering with different *k*. From Figure 2, we can note that our method performs much better than random clustering. We also find that our method gives a best performance when *k* is selected around 400. In this case, the FPUPC value reaches 0.32, whereas the value for random clustering is 0.004, which is two orders of magnitudes smaller.

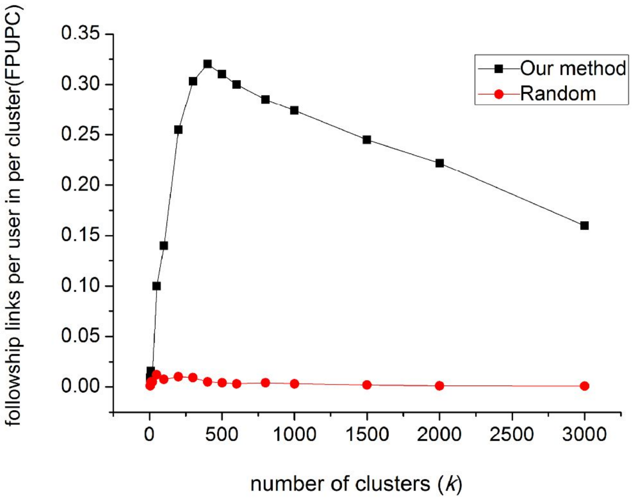

Fig.2 Clustering results of our method and Random Clustering

5.4. A visual case of clustering results

We list three users randomly selected from a single cluster in table 2. From their descriptions, we can visually see that our method can cluster users quite well.

Tab.2 Screen names and descriptions of three random users in a same cluster

Users	Description
@Mr_J_Shannon	#Freelance #Web #Designer from #Chicago looking to build my network and client base.
@victormirandamx	Professional of IT Security, IaaS Architect and Product Designer, Senior System Engineer, Lead Auditor ISO 27001, CISM, CNNA, CCSA, CCSE, PMP & Web Designer
@thecodebakery	Web technology news to your Twitter feed.

6. Conclusion

A fundamental task of understanding Twitter at user level is clustering users, which is the focus of this paper. We first compute user similarity by the use of different features which reflect one's interests, including both textual contents and social structure. Then we use user similarity as a measure to cluster users. To assess effectiveness of our method, we propose clustering metrics "average number of following links per user in per cluster". Experimental results show that our method can successfully cluster users in Twitter, and get a much better performance than random selection.

Nevertheless, as an early attempt to cluster users in Twitter, our work still has space for improvement. We envision two directions towards which our work can evolve. First, we intend to investigate the feasibility of applying other clustering algorithms to cluster users; second, with the popularity of some recent functions in Twitter such as "list", we plan to take these features into account when computing user similarity.

References

[1] Zaifan Jang, Bin Wang. User Behvior Based Personal Information Retrieval. IN *CCIR'10*, 2010

[2] D. Boyd, S. Golder, and G. Lotan. Tweet, tweet, retweet : Conversational aspects of retweeting on Twitter. In *43rd Hawaii International Conf. on System Sciences*, page 412, 2008.

[3] J. Weng, E-P. Lim, J. Jiang, and Q. He. Twitterrank: Finding topic-sensitive influential twitterers. In *WSDM'10*, 2010

[4] J. Welch, D. He, U. Schonfeld, and J. Cho. Topical Semantics of Twitter Links. In *WSDM'11*, 2011

[5] A. Tumasjan, T. O. Sprenger, P. G. Sandner, and I. M. Welpe. Predicting Elections with Twitter: What 140 characters Reveal about Political Sentiment. In *Proc. Of the Fourth International AAAI Conf. on Weblogs and Social Media. ICWSM'10*, 2010

[6] J. Bollen, H. Mao, and A. Pepe. Determining the public mood state by analysis of microblogging posts. In *Procceedings Of the Alife XII Conf.* MIT Press, 2010

[7] T. Sakaki, M. Okazaki, Y. Matsuo. Earthquake Shakes Twitter Users : Real-time Event Detection by Social Sensors. In *WWW'10*, 2010

[8] D. Zhao and M. B. Rosson. How and why people twitter: the role that micro-blogging plays in informal communication at work. In *Proceedings of the ACM 2009 international conference on Supporting group work.* ACM, 2009

[9] A. Java, T. Finin, X. Song, B. Tseng. Why We Twitter: Understanding Microblogging Usage and Communities. In *WBDKDD'07*, 2007

[10] H. Kwak, C. Lee,H. Park, S. Moon. What is Twitter, a Social Network or a News Media? In *WWW'10*, 2010

[11] C. Grier, K. Thomas, V. Paxson and M. Zhang. @spam: The Underground on 140 Characters or Less. In *CCCS'10*, 2010

[12] K. Lee, J. Caverlee and S. Webb. Uncovering Social Spammers: Social Honeypots + Machine Learning. In *SIGIR'10*, 2010

[13] B. Suh, L. Hong, P. Pirolli and H. Chi. Want to be Retweeted? Large Scale Analytics on Factors Impacting Retweet in Twitter Network. In *IEEE Second International Conference on Social Computing (SocialCom)*, pages 177-184. IEEE. 2010

[14] S. Petrovic, M. Osborne and V. Lavrenko. RT to Win! Predicting Message Propagation in Twitter. In *AAAI2011*, 2011

[15] L. Hong, O. Dan and B. D. Davison. Predicting Popular Messages in Twitter. In *WWW2011*, 2011

[16] M. Nagarajan, H. Purohit and A. Sheth. A Qualitative Examination of Topical Tweet and Retweet Practices. In *Proceedings of Fourth International AAAI Conference on Weblogs and Social Media (ICWSM'2010)*, pages 295–298, 2010

[17] D. Ramage, S. Dumais and D. Liebling. Characterizing Microblogs with Topic Models. In *AAAI'10*, 2010

[18] W. Galuba, K. Aberer, D. Chakraborty, Z. Despotovic and W. Kellerer. Outtweeting the Twitterers-Predicting Information Cascades in Microblogs.

[19] P. Cao, J. Li, M. Tong, Y. Liu and Xueqi Cheng. Detecting Near Duplicate Messages in Twitter. In *Journal of chinese information processing*, 2011

[20] M. Cha, H. Haddadi, F. Benevenuto and K. P. Gummadi. Measuring User Influence in Twitter : The Million Follower Fallacy. In *AAAI'10*, 2010

[21] D. M. Blei, A. Y. Ng and M. I. Jordan. Latent dirichlet allocation. In *Journal of Machine Learning Research*, 3:993-1022, 2003

[22] X. H. Phan, L. M. Nguyen and S. Horiguchi. Learning to classify short and sparse text & web with hidden topics from large-scale data collections. In *Procceedings of WWW '08*, pages 91-100, 2008