Diversification Recommendation of Popular Articles in Micro-blog Scenario

Jianxing Zheng, Bofeng Zhang*, Guobing Zou, Xiaodong Yue
School of Computer Engineering & Science
Shanghai University
Shanghai, China
e-mail: bfzhang@shu.edu.cn

Abstract—**With the information overload in web services, micro-blog has been increasingly providing as a media for end-users to express their opinions. The notable feature of micro-blog articles is prone to be a burst of popularity during a short period. In addition, diverse interests make users bored in redundant items in most recommender systems. Therefore, providing users with diverse popular micro-blogs that suit their interesting topics is an important issue. In this paper, depending on forwarding number and comment number of micro-blogs, an effective model for popularity prediction is proposed to discover popular topics. Then, a MaxMin diversity algorithm based on content distance and popularity density is proposed to discover top k micro-blogs. Finally, we design a diverse personalized popularity attention (DPPA) recommendation approach for target user. We conduct extensive experiments on large scale micro-blog datasets. The experimental results show that our proposed approach can satisfy user's requirements with a higher recall than personal attention methods.**

Keywords—micro-blog; topic popularity prediction, diversity, personalized recommendation

I. Introduction

With the rapid development of web technology, the Internet has become a significant medium source through which thousands of people can share and obtain latest information. In particular, as a new emerging communication paradigm on the Internet, micro-blog spreads popular information from one user to millions of individuals, which makes people receive diverse information at anytime and anywhere. However, end-users have difficulty in discovering and reading their desirable information from tremendous number of popular micro-blogs. Therefore, how to effectively provide users with popular diverse micro-blogs that fit their personalized interests has received great attention. Currently, it has become an open important research issue. To our best knowledge, very little research is concerned about this issue.

In micro-blog scenario, a user reposts a micro-blog from other users to display one's interests and feelings, which can accelerate some topic popularly. Taking into account discovering the trend of blog topics, Liu et al. [1] had addressed the problem of how to predict the popularity trend of blog topics over time and provide predicted personalized popular topics to a target user. With heavy-tailed Power Law distribution, Xie et al. [2] proposed an interest-driven model in order to simulate basic user communication behaviors. Unfortunately, the popularity of topic in micro-blog scenario is submitted to a few micro-blogs but not all of micro-blog contents. People are immersed in forwarding particular micro-blogs to track a hot topic. Thus, it is important to analyze the popularity trend of topic in terms of a part of messages.

Additionally, people not only are immersed in popular topic but also concerned on diversified recommendation results that relevant to hot topic. However, there is a phenomenon that more and more people become bored in redundant items that provided by web service [7]. As users are likely to go through a small fraction of the available popular information, an important challenge is how to identify a small number of representative popular articles to present. The goal in such diversity settings is not just to select the most diverse articles for user's information needs; rather, highly popular articles that present all different angles of hot topic should be preferred. Moreover, diversification problems are typically NP-hard; greedy approximations of such solution have also been shown to significantly improve results quality. Hence, it is desirable to mine diverse micro-blogs of popular topic for target user in order to solve the problem of redundant interests.

In this paper, we propose a novel approach for micro-blog recommendation both considering the popularity and diversification of topic. First, we identify the popularity degree of topic by forwarding quantity and comment quantity of particular micro-blogs. Moreover, differing from most existing recommender systems where they mainly push user's interesting relevant documents, we exploit user's diverse articles in terms of the interesting popular topic. Lastly, diverse personalized popular micro-blogs are recommended to target user.

The remainder of this paper is organized as follows. Section II briefly introduces the related works. In section III, we present an overall framework of recommending diverse popular micro-blogs. The prediction of popular topic is illustrated in detail in Section IV. Section V presents the diversification discovery method of popular topic. Section VI shows several recommendation strategies. Section VII presents experimental results and some specific analysis. Finally, Section VIII makes conclusion and discusses our future work.

*Corresponding author. This work was partially supported by the National Natural Science Foundation of China (No. 61303096, 61103067), the Natural Science Foundation of Shanghai (13ZR1454600) and Innovation Program of Shanghai Municipal Education Commission (11ZZ85).

II. Related Work

A. Popularity Ttrend Discovering

Micro-blogs usually involve a large number of varying topics. Recently, several studies have been done to analyze the changes of blog features to discover the trend of topics. Gruhl [3] analyzed the blog text and investigated the information propagation of blog topics. Lee [4] applied a density-based online clustering method for mining micro-blog text streams to analyze temporal and geospatial features of real-world events.

Additionally, many researches focus on the analysis of a set of observation values by time order to build an adaptive model and predict the future trends. One of the most significant methods is the exponential smoothing method [1][5], which has a hypothesis that the time series is stable and regular. The exponential smoothing method is mainly used to assign more weight for recent observations in forecast than the older observations [6]. However, researchers mainly analyze the content of all documents to identify the dynamic trend of topic evolution, which could not fit for representative sample of interesting topics in micro-blog scenario. In our paper, we classify micro-blogs into topics via the fusion of classification method, and then use the lead central messages with high forwarding number and comment number to measure the popularity degree of topic.

B. Diversification Mining

Diversification has attracted a lot of attentions as a means of improving the performance of query results in recommender systems. J. C. et al. [7] designed the maximal marginal relevance method, which attempted to maximize relevance while minimizing the similarity of higher ranked documents. C. Z. et al. [8][9] developed and validated subtopic retrieval methods based on a risk minimization framework and introduce corresponding measures for subtopic recall and precision.

There have been various definitions of diversity, based on content, novelty and semantic coverage [10]. M. D. et al. [10] presented a Dis-C diverse subset of a query result containing objects such that each object in the result is represented by a similar object in the diverse subset and the objects in the diverse subset are dissimilar to each other. A. A. et al. [11] studied diversity-aware search by setting some captures and an efficient threshold algorithm to propose a low-overhead, intelligent data access prioritization scheme. D. V. et al. [12] developed a generalized framework for personalized diversification by researching the introduction of the user as an explicit random variable, which tried to maximize the probability that some query results is relevant to the user's diversity interest. C-N. Z et al. [13] designed a novel topic diversification method to balance and diversify personalized recommendation lists by improving user satisfaction. All these researches mainly focus on the diversification about the relevance of user interest. Apparently, in micro-blog scenario, a target user is highly interested in the popular topic, and we need to consider recommending diverse popular articles to improve the satisfaction of glimpsing lists for a user.

III. Overall Framework of Diverse Popular Micro-blog Recommendation

We propose a framework of diverse micro-blog recommendation with the aid of popular topic. It involves three major steps: popular topic prediction, diversification selection of topic, diverse personalized popular micro-blog recommendation. The framework is illustrated in the following Fig. 1.

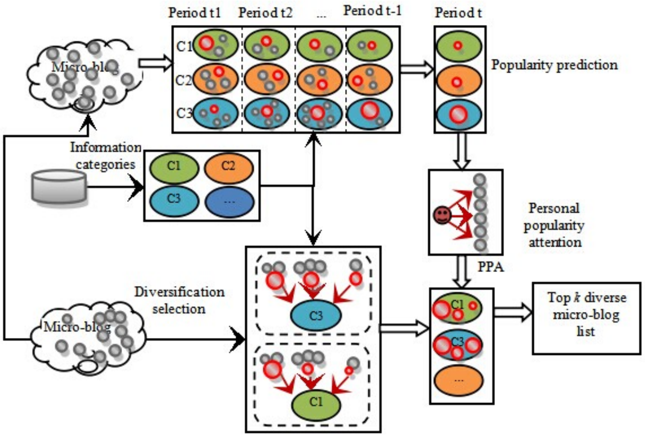

Fig. 1. Framework of diverse personalized popular micro-blog recommendation.

The first step of framework is partition of topic categories. The latest common micro-blogs are collected to be divided into predefined topic categories with machine learning classification algorithm. In popular topic prediction module, based on the forwarding number and comment number of representative micro-blogs in predefined topic, we design a popularity degree acquiring method for the given topic. In Fig. 1, the red micro-blog object in each topic category is used to calculate the popularity degree of topic in every time period. Furthermore, considering the popularity effect of adjacent time periods, the predicted popularity degree is calculated to verify the popularity trend of topic.

For the diversification selection module, taking into account coverage ability of each micro-blog, an algorithm for selecting diverse top k popular micro-blogs are presented. Here, we assume that a few micro-blogs with high popularity and dissimilar to each other could represent the diversification of popular topic.

To recommend diverse personalized popular micro-blogs, personal popularity attention (PPA) of topic is designed to rerank the appropriate desired interest for a target user. Based on personalized interesting popular topic, diversification selection algorithm is applied to select representative popular micro-blogs for recommendation. By considering personal interests, for topics C1 and C3, the red diverse popular micro-blogs are selected for target user.

IV. Topic Popularity Prediction

In this section, we propose an approach to calculate the actual popularity degree of topic. Then, based on the integration of actual popularity degree and predicted trend effect of topic, the prediction popularity degree can be modeled.

A. Partition of Topic Categories

In micro-blog scenario, micro-blog texts involve various kinds of topic so that they are necessary to be classified into appointed category [14]. In this paper, we use message features, i.e., a term-weight vector, to compute the similarity of arbitrary two micro-blogs, and classify the message into one of the eight topic categories with KNN algorithm.

Firstly, the short message is preprocessed by stopword removal and word stemming, then each micro-blog is represented as a term weight vector with a set of selected terms $m = \{(t_1, w_{1m}), (t_2, w_{2m}), \ldots, (t_p, w_{pm})\}$, defined in Eq.(1):

$$w_{tm} = tf_{tm} * idf_t\,, \qquad (1)$$

$$tf_{tm} = \frac{freq_{tm}}{\max_l(freq_{lm})}\,, \qquad (2)$$

$$idf_t = \log\frac{N_m}{n_t}\,, \qquad (3)$$

where $freq_{tm}$ is raw term frequency of term t in micro-blog m , and $\max_l(freq_{lm})$ is the frequency number of term l which has the maximum frequency in m . N_m represents the total number of micro-blogs and n_t is the number of micro-blogs containing term t as well as m .

The similarity between two micro-blogs by means of the cosine measure can be defined as:

$$sim(m_i, m_j) = \frac{\overrightarrow{m_i} \bullet \overrightarrow{m_j}}{\|\overrightarrow{m_i}\| \cdot \|\overrightarrow{m_j}\|} \qquad (4)$$

The time window size is set as seven days. In other words, all micro-blogs are slipped into several segments by time period; then articles in each segment are then partitioned into appointed topics. In each time period, KNN algorithm is applied to compute the distance for each micro-blog to all the samples. It helps to obtain k-nearest neighbor and then successively select the appropriate topic with the most number of neighbor sample micro-blogs. We divide each micro-blog to predefined topics such as Economy, IT, Health, Sports, Travel, Education, Career and Culture, which are from categorization corpus of Sogou Lab Data.

For topic category C , it can be formed by a group of micro-blogs as $\{m_1, m_2, \cdots, m_n\}$. Fig. 2 shows topic categories sample in different time periods.

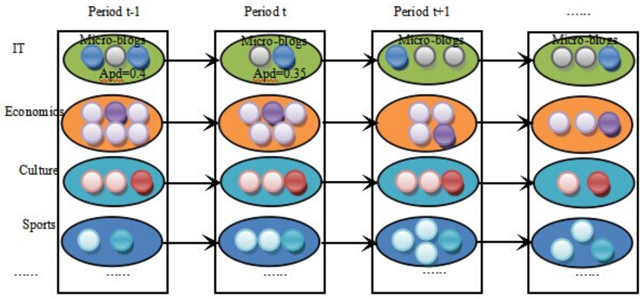

Fig. 2. Topic categories sample of micro-blogs in different time periods.

B. Calculation of Popularity Degree for Topic

1) Actual popularity degree

Considering the category C for period t , the average center potential of micro-blog m_k in C can be represented as $Acp_t(m_k)$ in Eq. (5):

$$Acp_t(m_k) = \frac{\sum_{m_i \in C,\, m_i \neq m_k} sim(m_i, m_k)}{|M(C)| - 1}\,, \qquad (5)$$

where $|M(C)|$ is the number of micro-blogs in topic category C . The larger $Acp_t(m_k)$ is, the greater the probability of representing other messages is. Hence, the average center of category C can be discovered according to Eq. (6):

$$Acp_t(C) = Acp_t(m_c)\,, m_c = \arg\max_{m_k \in M(C)} Acp_t(m_k) \; (6)$$

For each micro-blog m_k in category C , the actual popularity degree $Apd(m_k)$ can be acquired from the forwarding number and comment number of target message and its average center potential, as shown in Eq.(7):

$$Apd_t(m_k) = \log(1 + fc(m_k)) \times Acp_t(m_k)\,, \qquad (7)$$

where $fc(m_k) = \frac{|fn_t(m_k) + cn_t(m_k)|}{\max_{m_i \in M_t}\{fn_t(m_i) + cn_t(m_i)\}}$, and $fn(\cdot)$, $cn(\cdot)$ represents the number of forwarding and comments, respectively. M_t denotes the set of micro-blogs at time period t . Specially, the $Apd(m_k)$ marks zero when there is no forwarding and comment for m_k . Then, the popularity degree of category C is obtained by the maximum $Apd(m_k)$ in C , as defined in Eq. (8):

$$Apd_t(C) = Apd_t(m_p)\,, m_p = \arg\max_{m_k \in M(C)} Apd_t(m_k) \qquad (8)$$

2) Prediction popularity degree

According to the actual popularity degree of the most popular micro-blog, we make a prediction of popularity degree for a topic. As the double exponential smoothing method referred in [1], combining the absolute popularity degree $Apd_t(C)$ and prediction popularity degree $Ppd_t(C)$ in the preceding time period t, the prediction popularity degree of topic C at time period $t+1$ is calculated in Eq. (9):

$$Ppd_{t+1}(C) = \alpha \times Apd_t(C) + (1-\alpha) \times [Ppd_t(C) + b_t(C)] \quad (9)$$

where $Ppd_{t+1}(C)$ is the predicted popularity degree for topic category C at period t; $Apd_t(C)$ is the actual popularity degree for C at period t; $Ppd_t(C)$ is the predicted popularity degree for C at period t; $b_t(C)$ is the trend effect for C at period t. The parameter α is used to measure the relative importance of actual popularity degree and predicted trend effect in the preceding time period [1], which belongs to $[0,1]$.

Furthermore, we compute the predicted trend effect in the period t by combining the differences of predicted popularity degree at period t and $t-1$, and predicted trend effect in the preceding period $t-1$, as shown in Eq. (10):

$$b_t(C) = \delta \times [Ppd_t(C) - Ppd_{t-1}(C)] + (1-\delta) \times b_{t-1}(C). \quad (10)$$

Note that the parameter δ is used to balance differences of predicted popularity degree at period t and $t-1$, and predicted trend effect in the preceding period $t-1$.

V. DIVERSIFICATION OF TOPIC

In this section, based on popularity of topic category, the diversity problem of micro-blog selection is briefly defined, and top k diverse popular micro-blogs are discovered for target user.

A. The K-diversity Problem

For a given set of micro-blog objects in popular topic $U: |U| = n$ (e.g., search results, news or other messages), we want to select a representative micro-blog subset S of these objects such that each object from U is represented by a similar object in S and the micro-blog objects selected to be included in S are dissimilar to each other.

Definition 1: Let U be a set of micro-blogs, $d \geq 0$ a distance metric and f a function measuring the diversity of subset $S \subseteq U$. Let also k be a positive number. The k diversity problem is to select a subset S^* of U such that $S^* = \arg\max_{\substack{S \subseteq U \\ |S|=k}} f(S,d)$.

Definition 2: Given a set U of micro-blogs and distance metric d. $S \subseteq U$, the maximizing the minimum distance among S can be formally defined as $f_{MIN}(S,d) = \min_{\substack{m_i, m_j \in S \\ m_i \neq m_j}} d(m_i, m_j)$.

To simplify the presentation, for $\forall m_i, m_j \in U$, we assume that the distance metric of two micro-blog objects based on content defined as $d_c(m_i, m_j) = 1 - |sim(m_i, m_j)|$. Additionally, the distance metric about popularity degree between micro-blogs can be represented as $d_p(m_i, m_j) = \frac{|Apd(m_i) - Apd(m_j)|}{\max\{Apd(m_i), Apd(m_j)\}}$.

Taking into account the popularity trend of micro-blog, for $m_i \in U$, we use $N(m_i)$ to denote the set of popularity neighborhood of m_i : $N_r(m_i) = \{m_j \mid d_p(m_j, m_i) \leq r\}$. Furthermore, the average popularity density of micro-blog m_i in U is defined as $dens_r(m_i) = \frac{|N_r(m_i)|}{|U|}$.

The average popularity diversity reflects how many objects are nearly popular with the target object, and parameter radius r can adjust the size of neighborhood. Based on popularity diversity and distance metric of the two objects, the MaxMin greedy heuristic algorithm can select the most diverse popular top k micro-blogs from topic category.

B. Computing Diverse Subsets

We consider a series of micro-blog objects that can represent contents of all micro-blogs in topic C. The core idea of our diverse subsets discovery method is to process the input U as all micro-blog objects in C and to continuously select a diverse subset S, which allows that the objects in S can be far to each other and all objects in U are represented or replaced by at least one similar popular object in S. Algorithm 1 gives a specific k-diversity MaxMin method.

```
Algorithm 1: k-diversity MaxMin method.
Input: A set of micro-blog objects $U$ and radius $r$.
Output: An k-diversity subset $S$ of $U$.
1. $S \leftarrow \varnothing$;
2. select the first object $m_i$ with the largest
   $Apd(C) \times |dens_r(m_i)|$;
3. select the second object $m_i^2$, where $m_i^2$ satisfies
   $d_c(m_i^2, m_p) \times dens_r(m_i^1) =$
   $\max_{k=1}^{|U|}\{d_c(m_k, m_p) \times dens_r(m_k) \mid m_k \in U, m_p \in S\}$;
4. $S = S \cup \{m_i^2\}$;
5. **while** $|S| \leq k$ do
6.    select the object $m_i^*$, where $m_i^*$ satisfies
      $d_c(m_i^*, m_p) \times dens_r(m_i^*) =$
      $\max\{\min\{d_c(m_k, m_p) \times dens_r(m_k)\} : m_k \in U, m_p \in S\}$;
7.    $S = S \cup \{m_i^*\}$;
8. **end while**
9. return $S$.
```

Obviously, in steps 2 and 3, the first and second objects are most popular and dissimilar. In step 6, this algorithm shows us to select the object that covers the largest possible number of popular objects and keeps the farthest content distance with the objects in candidate set. In this case, the candidate object set includes the most popular micro-blogs, and mutually away from each other in aspect of contents. For a user, we can find one's personal interesting popular topic, and select diverse popular micro-blogs to make ideal recommendation.

VI. Recommendation Strategies

In this section, we rerank the appropriate topics by integrating the popularity with preference of a user and then perform top k-diversity popular micro-blog recommendation and provide some other recommendation strategies.

A. Personal Attention

In this subsection, we infer user's interest for topic category by their browsing behaviors. Let M_u be the set of micro-blogs that user u interested in, and $M_u(C)$ the set of micro-blogs divided in topic C. The number of micro-blogs involved C is used to estimate the preference score (PS) on C for target user u as follows:

$$PS_u(C) = \frac{|M_u(C)|}{|M_u|}. \qquad (11)$$

Then, by analyzing a user's preference score to a new micro-blog m_k, $PS_u(m_k)$, we can infer personal attention the user pays m_k, as shown in Eq. (12):

$$PA_u(m_k) = \frac{e^{|PS_u(m_k)|} - 1}{e - 1}, \qquad (12)$$

where $PS_u(m_k) = PS_u(C) \times \max_{m_j \in M_u(C)} \{sim(m_j, m_k)\}$ is the preference score of m_k in topic C. Moreover, top k micro-blogs most similar to the topic C with largest $PS_u(C)$ are selected to pushed target user.

B. Personal Popularity Attention

However, as detailed elaboration in above sections, the popularity degree of micro-blog reflects the collective trend effect in the next period. By considering personal interests, we need to recommend some popular topic as much as possible for target user so as to expand user's diverse interest. As is known, the harmonic mean method is omitting a lot of parameter tuning work comparing with arithmetic mean method. Therefore, we adopt the harmonic mean approach to combine the popularity with the preference score of topic for target user, as shown in Eq. (13):

$$PPS_u(C) = \frac{2 \times Ppd(C) \times PS_u(C)}{Ppd(C) + PS_u(C)}. \qquad (13)$$

According to the personalized popular topic, we can compute a user's preference popularity score to a new micro-blog m_k like Eq. (12), $PPA_u(m_k)$.

However, ranking $PPS_u(C)$ and selecting appropriate topic, k-diverse popular micro-blogs in topic can be supplied with the target user in terms of Algorithm 1.

VII. Experiiments and Discussions

In this section, we evaluate the effectiveness of our proposed method by comparing against several strategies.

A. Data Sets and Experimental Design

In our study, we adopt *Nlpir* dataset (http://www.nlpir.org/), which is from Sina Weibo website. The dataset was selected during a period from 4th of Nov. 2011 to 16th of Dec. 2011. For the dataset, 46 users and 1247 followee friends' common micro-blogs were selected elaborately for testing diverse popular interest recommendation. Statistics for dataset are shown in Table 1. In Table, common micro-blogs are separated into different time periods and classified into appointed eight topic categories in order to compute popularity degree of topics. In addition, users' micro-blogs are used to compare the performance of recommendation results.

TABLE I. Statistics Showing the Number of Users, Users' Micro-blogs and Common Micro-blogs

Nlpir dataset		
Users	***Users' micro-blogs***	***Common micro-blogs***
46	8326	41950

To evaluate the prediction accuracy of popularity, we adopt mean absolute error [1] as the evaluation metric. For a topic C_i, mean absolute deviation between prediction popularity degree and actual popularity degree is measured to verify the effect of popularity, as shown in Eq. (14):

$$MAE(C_i) = \frac{\sum_{t \in T} |Ppd_t(C_i) - Apd_t(C_i)|}{|T|} \qquad (14)$$

To verify the performance of diverse popularity recommendation in Algorithm 1, we use precision, recall and F1 measure metrics, defined as follows:

$$\Pr ecision = \frac{|I \cap R|}{|R|}, \qquad (15)$$

$$\text{Re}\,call = \frac{|I \cap R|}{|I|}, \qquad (16)$$

$$F1 = \frac{2 \times \Pr ecision \times \text{Re}\,call}{\Pr ecision + \text{Re}\,call}, \qquad (17)$$

where I is the set of micro-blogs in each time period and R is the set of recommended top k diverse micro-blogs. In our experiment, common messages are utilized to discover the popularity of topic. The proposed approach can provide users with personalized hot topic, which could improve the quality of popular micro-blog recommendation. We compare performance of this approach against PA method in Eq. (12) and PPA method in Eq. (13).

B. Results and Discussion

1) Popularity results

This section presents the experimental results of proposed popularity approach based on different weight settings of parameters. As interpreted in Eqs. (9) and (10), the values of α and δ are significant to determine the prediction popularity degree. On *Nlpir* dataset, we varied the value of δ with an increment of 0.1 under $\alpha = 0.5$ to examine the trend effect. Fig. 3 presents the average of MAE for different topic categories, including C1 (Economy), C2 (IT), C3 (Health), C4 (Sports), C5 (Travel), C6 (Education), C7 (Career) and C8 (Culture). As we can see, although there is no poignant change under different values of δ, we can still choose $\delta = 0.4$ to get the best prediction performance about popularity.

Based on above results, the best parameter setting of predicting popularity degree in micro-blog scenario is $\alpha = 0.5$, $\delta = 0.4$. By varying the value of α and $\delta = 0.4$, Fig. 4 shows different changes of MAE for eight topics. The prediction model achieves the lowest MAE under $\alpha = 0.4$ to 0.6 for most of categories, which means that the predicted popularity degree can be valid when we set half weight on the preceding actual popularity degree and preceding predicted popularity degree. Hence, such parameter settings are used in the rest of our experiments. Fig. 5 shows specific MAE values under $\alpha = 0.5$, $\delta = 0.4$. In Fig. 5, there are homogeneous errors nearly 0.06 in most of topics except for C2 (IT) and C5 (Travel). As we can see, built-in audiences of people interested in IT and Travel topics, which make them easier to keep stable popularity. In Figs. 6 and 7, we can observe that the popularity trend of rising or falling for different topics. Although the Apd of specific topic is different from Ppd, the basic trends of rising or falling for Apd and Ppd are consistent.

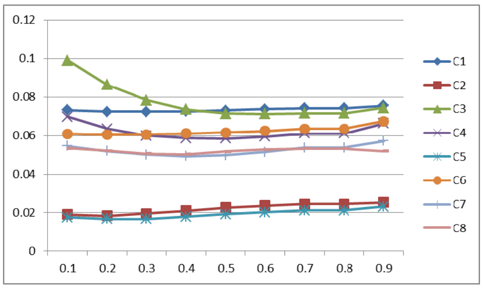

Fig. 3. Different MAEs under different δ ($\alpha =$ 0.5) for eight topics.

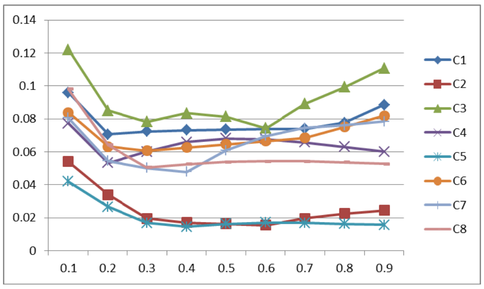

Fig. 4. Different MAEs under different α ($\delta =$ 0.4) for eight topics.

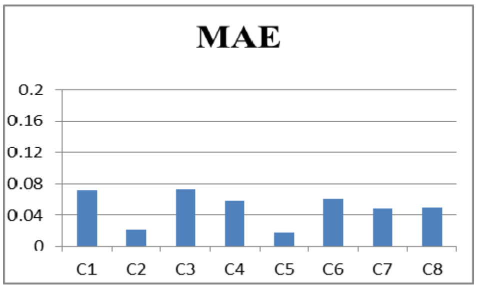

Fig. 5. Different MAEs for eight topics ($\alpha = 0.5$, $\delta = 0.4$).

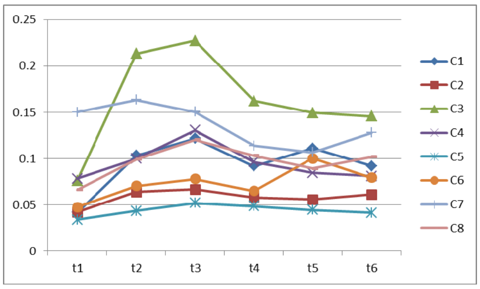

Fig. 6. Ppd of eight topics in different time periods.

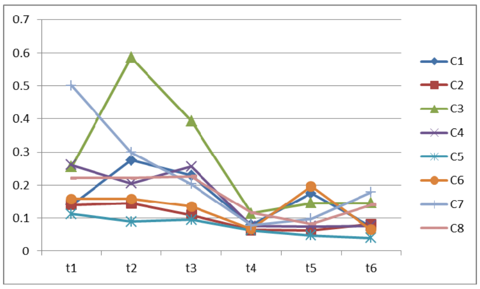

Fig. 7. Apd of eight topics in different time periods.

2) Recommendation results

Based on $\alpha = 0.5$, $\delta = 0.4$, for a personal popular topic in Eq. (13), we recommend each user in turn with different number of diverse personalized popular attention (DPPA) micro-blogs. For the diversity radius r, we choose $r = 0.1$ to set the size of neighborhood density.

Figs. 8, 9, 10 shows precision, recall, F1 curves under different recommendation list size k on *Nlpir* dataset, respectively. Here, the initial value of k is set 6. For the dataset, in Fig. 8, there is a relative benefit for the DPPA recommendation compared to the PPA and PA approaches. For example, for the case of recommendation list of size 18, we can see that the proposed DPPA method give a precision value of 0.45, which means that on average nearly 8 of the 18 recommended diverse popular micro-blogs are received by target user. In addition, the precision performance of DPPA strategy outperforms PPA approach when the recall performance displays better results, which is demonstrated by the result that on average 10 of the 18 recommended micro-blogs are received by target user in DPPA strategy. In Fig. 10, DPPA recommendation both considering the popularity and diversity shows better results than traditional personal approaches. This is likely because partial popular micro-blogs attended by target user at some time is limited, and diverse popular micro-blogs can cover wide interests of target user.

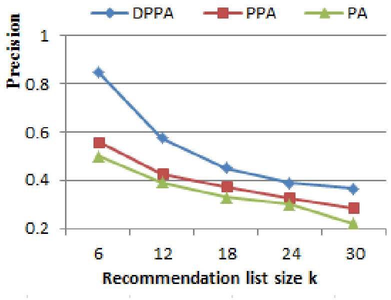

Fig. 8. Precision under different k based on DPPA, PPA and PA methods.

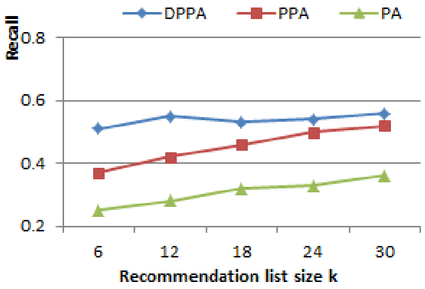

Fig. 9. Recall under different k based on DPPA, PPA and PA methods.

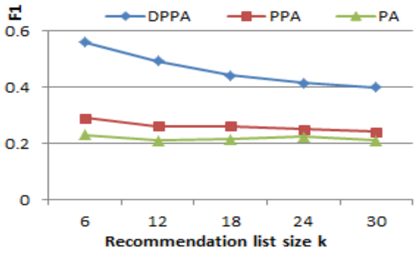

Fig. 10. F1 under different k based on DPPA, PPA and PA methods.

In the above figures, for DPPA and PPA methods, they both have higher performance than PA method without considering the popularity. Additionally, all approaches provide similar trend with the recommendation list size k increasing: the precision is decreasing and becoming smooth and recall is increasing. In micro-blog scenario, the number of meaningful popular micro-blogs in some topic is small, and they are prone to be focused during the propagation of hot events. These specific micro-blogs are discovered simply and the growth rate is lower than that of list size k , which leads to weaker recommendation performance gradually.

VIII. Conclusion

This paper proposes a novel recommendation method for popular micro-blogs by considering the diversity and personal attention. The method first models popularity degree of every topic by taking into account the forwarding number and comment number of particular micro-blog representing the center of topic. Then, diversification of popular topic is presented to identify representative popular micro-blogs that are dissimilar to each other. Finally, the diverse popularity approach is applied into personal interests to select top-k micro-blogs for recommendation. The extensive experimental results demonstrate that our proposed method is the most promising and effective one for popular topics, which can fit the diversity of interests in realistic micro-blog scenario.

In the future work, we will further polish user interests for each topic at multiple granularities so that it can be more effectively applied in our method for multi-layer diverse micro-blog recommendation. In addition, the selecting of diverse problem for representing the popular topic is an evolving problem, dynamic diversity method should be considered to effectively reduce the time consumption.

Acknowledgment

The authors are very grateful to the anonymous reviewers. Their many helpful and constructive comments and suggestions helped us significantly improve this work.

References

[1] D. R. Liu, P. Y. Tsai, P. H. Chiu, Personalized recommendation of popular blog articles for mobile applications, Information Sciences, 2011, vol. 181, no. 9, pp. 1552-1572.

[2] J. J. Xie, C. Z., M. Wu, Modeling microblogging communication based on human dynamics, in: Proceedings of the 8th International Conference on Fuzzy Systems and Knowledge Discovery, Beijing, China, 2011, pp. 2290-2294.

[3] D. Gruhl, R. Guha, R. Kumar, J. Novak, A. Tomkins, The predictive power of online chatter, in: Proceedings of the 11th ACM SIGKDD International Conference on Knowledge Discovery in Data Mining, New York, USA, 2005, pp. 78-87.

[4] C. H. Lee. Mining spatio-temporal information on microblogging streams using a density-based online clustering method, Expert Systems with Applications, 2012, vol. 39, no. 10, pp. 9623-9641.

[5] R. G. Brown, Smoothing, Forecasting and prediction of discrete time series, Courier Dover Publications, 2004.

[6] D. C. Montgomery, L. A. Johnson, J. S. Gardiner, Forecasting and time series analysis, second ed., McGraw-Hill, 1990.

[7] J. Carbonell and J. Goldstein. The use of MMR, diversity-based reranking for reordering documents and producing summaries. In 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, 1998, pp. 335-336.

[8] C. Zhai, W. W. Cohen, J. Lafferty. Beyond independent relevance: Methods and evaluation metrics for subtopic retrieval. In 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, 2003, pp. 10-17.

[9] C. Zhai, J. Lafferty, A risk minimization framework for information retrieval, Information Processing & Management, 2006, vol. 42, no. 1, pp. 31-55.

[10] M. Drosou, E. Pitoura, DisC diversity: result diversification based on dissimilarity and coverage. In Proceedings of the VLDB Endowment, 2012,vol. 6, no. 1, pp. 13-24.

[11] A. Angel, N. Koudas, Efficient diversity-aware search, In Proceedings of the 2011 ACM SIGMOD International Conference on Management of data, Athens, Greece, 2011, pp. 781-792.

[12] D. Vallet, P. Castells, Personalized diversification of search results, In Proceedings of the 35th international ACM SIGIR Conference on Research and Development in Information Retrieval, Portland, Oregon, USA, 2012, pp. 841-850.

[13] C.-N. Ziegler, S. M. McNee, J. A. Konstan, and G. Lausen. Improving recommendation lists through topic diversification. In Proceedings of the 14th international conference on World Wide Web, Chiba, Japan, 2005, pp.22-32.

[14] S. Shehata, F. Karray, M. S. Kamel, An efficient concept-based mining model for enhancing text clustering, IEEE Transactions on Knowledge Data Engineering, 2010, vol. 22, no. 10, pp. 1360-1371.