A Microblog Recommendation Algorithm Based on Multi-tag Correlation

Huifang Ma[1,2(✉)], Meihuizi Jia[1], Meng Xie[1], and Xianghong Lin[1]

[1] College of Computer Science and Engineering, Northwest Normal University,
Lanzhou Gansu 730070, China
mahuifang@yeah.net
[2] Key Laboratory of Intelligent Information Processing of Chinese Academy of Sciences,
Institute of Computing Technology, Chinese Academy of Sciences, Beijing 10085, China

Abstract. In this paper, we present a microblog recommendation algorithm based on multi-tag correlation. Firstly, a tag retrieval strategy is designed to add tags for unlabeled users, the initial user-tag matrix is then constructed and user-tag weights are set. In order to represent user interests accurately, we fully investigate the associations between the tags. Both inner and outer correlation between tags are defined to conquer the problem of sparsity of user-tag matrix. The user interests can then be decided and microblogs can be recommended to users. Experimental results show that the algorithm is effective for microblog recommendation.

Keywords: Microblog recommendation · Tag retrieval · User-Tag weight · Tag correlation

1 Introduction

As a typical representative application of web 2.0, microblog has attracted a great number of users and rapidly developed in recent years. It is necessary to develop new algorithms to provide the personalized service for these users. And high quality information should be accurately pushed based on the user's interest[3].

In this paper, we present a microblog recommendation algorithm based on multi-tag correlation. First, a novel user tag retrieval strategy is developed to select the tags for unlabeled users and a user-tag matrix is created to represent the initial weight of users' tags. Second, we construct a correlation matrix of multi-tag by investigating inner and outer correlation between tags. Third, the original user-tag matrix is updated by correlation matrix of multi-tag to obtain the final weight.

The basic outline of this paper is as follows: Section 2 presents user method. The experiments and results are demonstrated in Section 3. Lastly, we conclude our paper in Section 4.

2 Our Approach

2.1 User Tag Retrieval and User-Tag Matrix Construction

If the tagging service is provided by microblog system, the built-in tags can be directly used. Otherwise, a tag retrieval method is adopted to acquire the personal tags from the microblog posted by that user.

S. Zhang et al. (Eds.): KSEM 2015, LNAI 9403, pp. 483–488, 2015.
DOI: 10.1007/978-3-319-25159-2_43

Table 1. Notions in tag retrieval

notion	definition
$U=\{u_1,u_2,\dots,u_i,\dots u_N\}$	The microblog user dataset
N	The number of users
$D_i=\{d_{i1},d_{i2},\dots,d_{iM_i}\}$	The microblog collection for u_i
M_i, i=(1,2,…,N)	The number of microblog for u_i
$D=\bigcup_{i=1}^{N} D_i$	The microblog dataset of all users
$T_i=\{t_{i1},t_{i2},\dots,t_{im_i}\}$	The terms set for microblog dataset D_i
m_i, $m_i \ll M_i$	The number of terms
$L_i=\{l_{i1},l_{i2},\dots,l_{in_i}\}$	The tag set for u_i
n_i	The number of tags for u_i
$L=\bigcup_{i=1}^{N} L_i$	The collection for all tags
n	The total number of tags in tag collection L

The most critical step for tag retrieval is to select most representative words from users previously posted microblog as query words. Ideally, the query words should be (i) well represent the main content of the microblog, and (ii) be topically indicative. The clarity score[1] is used to measure the topical-specificity of a certain term. Let the *j-th* term l_{ij} be a candidate word for selection, we use l_{ij} as a single-term query to retrieve its top-g_i most relevant microblogs, denoted by Q_{lij}. We use equation (1) to calculate the clarity score of term l_{ij}.

$$Clarity\left(l_{ij}\right)=\sum_{l_{ij}\in T_i} P\left(l_{ij}\mid Q_{l_{ij}}\right)\log\frac{P\left(l_{ij}\mid Q_{l_{ij}}\right)}{P\left(l_{ij}\mid D_i\right)} \tag{1}$$

Then the score of the *j-th* word is computed as the equation below:

$$s_j = tf_j \times clarity\left(l_{ij}\right) \tag{2}$$

n_i words with the highest weight are chosen as user u_i*'s* tags, and each tag is assigned a normalized weight:

$$normalized\left(s_j\right)=\frac{s_j}{\sum_{x=1}^{n_i} s_x} \tag{3}$$

The tag weight vector $V_i=\left(w_{i1},w_{i2},\dots w_{in}\right)$ is created for user u_i to represent the initial weights of tags[4]. If the tags are obtained from the above tag retrieval scheme, we use Eq. (3) to generate the initial weights for these tags. Otherwise, if the tags are provided by the tagging service, each tag is treated as of equal importance. Assuming that u_i has Z_i tags, the initial user's tag weight w_{ij} are defined as follows:

$$w_{ij}=\begin{cases}1/Z_i, & \text{if tag } l_j \text{assigned to } u_i,\\ normalized\left(s_j\right), & \text{tag retrieval,}\\ 0, & \text{otherwise,}\end{cases} \tag{4}$$

Based on users' weight vectors, we create a $N*n$ matrix M_{ul}, which is defined as:

$$M_{ul}=\begin{bmatrix}\overrightarrow{V_1}\\ \overrightarrow{V_2}\\ \cdots\\ \overrightarrow{V_N}\end{bmatrix}=\begin{bmatrix}w_{11} & w_{12} & \cdots & w_{1n}\\ w_{21} & w_{22} & \cdots & w_{2n}\\ \vdots & \vdots & \vdots & \vdots\\ w_{N1} & w_{N2} & \cdots & w_{Nn}\end{bmatrix} \tag{5}$$

Where N denotes the total number of users, n is the number of tags and w_{ij} is the weight of the *j-th* tag for the *i-th* user in matrix.

2.2 Multi-tag Correlation and the Recommendation Algorithm

Sun et al[2] have considered associations between terms for microblog hot topic detection, inspired by this, we investigate correlations between tags to update the original user-tag matrix.

If both tags are marked by one particular user, an **inner correlation** between two tags is established. Based on Jaccard similarity, the inner correlation between tags l_j and l_k can be quantified as:

$$LIR\left(l_j,l_k\right)=\frac{1}{|H|}*\sum_{y\in H}\frac{w_{ij}w_{ik}}{w_{ij}+w_{ik}-w_{ij}w_{ik}} \tag{6}$$

$|H|$ represents the number of elements in $H=\{y|(w_{ij}\neq 0)\&(w_{ik}\neq 0)\}$, We normalize $LIR(l_j,l_k)$ to [0,1], and the normalized inner correlation of tags l_i and l_k are defined as:

$$N-LIR\left(l_j,l_k\right)=\begin{cases}1, & j=k\\ \dfrac{LIR\left(l_j,l_k\right)}{\sum_{j=1,j\neq k}^{n}LIR\left(l_j,l_k\right)}, & j\neq k\end{cases} \tag{7}$$

If two users u_1 and u_2 are simultaneously marked by the same tag, an **outer correlation** between two tags is established. Where tag l_q is a linked tag which links l_j and l_k. The outer correlation between two tags l_j and l_k linked by term l_q can be formalized as:

$$LOR\left(l_j,l_k\mid l_q\right)=\min\left(N-LIR\left(l_j,l_q\right),N-LIR\left(l_k,l_q\right)\right) \tag{8}$$

We then define the outer correlation between two tags l_j and l_k with all the linked terms and normalize the values to [0,1] as:

$$N-LOR\left(l_j,l_k\right)=\begin{cases}0, & j=k\\ \dfrac{\sum\limits_{\forall t_q\in E} LOR\left(l_j,l_k\,|\,l_q\right)}{|E|}, & j\neq k\end{cases} \tag{9}$$

Where $|E|$ denotes the number of link tags in $E=\{l_q|(N\text{-}LIR(l_j,l_q)>0)\&(N\text{-}LIR(l_k,l_q)>0)\}$.

Given a pair of tags l_j and l_k, the tag correlation between them can be defined as:

$$LR\left(l_j,l_k\right)=\begin{cases}1 & j=k\\ \alpha*N-LOR\left(l_j,l_k\right)+\left(1-\alpha\right)*N-LIR\left(l_j,l_k\right) & otherwise\end{cases} \tag{10}$$

Where α ($\alpha\in\left[0,1\right]$) determines the importance of inner correlation and outer correlation between multi-tags. Then we create a $n\times n$ multi-tag relationship matrix M_{lr}, and $LR(l_j,l_k)$ are elements in this matrix, which not only includes inner correlation between tags but also consider outer correlations between multi-tags. The final user-tag matrix M_{re} can be defined:

$$M_{re}=M_{ul}\times M_{lr} \tag{11}$$

Given a microblog d_p, the ranking function $f(u_i,d_p)$ for a user $u_{i,}$ is defined as[4]:

$$f\left(u_i,d_p\right)=\boldsymbol{E}_P\cdot\left(\boldsymbol{V}_i^{'}\right)^{\mathrm{T}} \tag{12}$$

Which denotes the similarity between microblog d_p and user u_i. Each microblog d_p is represented as $\boldsymbol{E}_p=\left(w_{p1},w_{p2},\ldots,w_{pn}\right)$, if d_p contains tag l_j then w_{pj}=1, w_{pj}=0 otherwise. The vector $\boldsymbol{V}_i^{'}=\left(w_{i1}^{'},w_{i2}^{'},\ldots,w_{in}^{'}\right)$ is the updated tag weight vector for user u_i. We predefine a threshold $\gamma_{i,}$ if $f(u_i,d_p)>\gamma_i$, then the microblog d_p will be recommended to the user u_i.

3 Experiments and Results

In this section, we show the experimental results on the dataset collected from Sina microblog platform. Our dataset includes 532 users who posted a large number of microblog from March 21th to March 25th, 2014. We preprocess a series of experimental data, and then the final experimental dataset is constructed. The dataset consists of 39,886 train and 9,100test datain14 categories, as shown in Table 2.

Table 2. Number of training/test data in the 14 categories

Category	#Train	#Test	Category	#Train	#Test
Sports	3481	800	Military	1880	300
Technology	2860	600	Parenting	2400	600
Estate	2650	600	Environmental protection	2880	600
Stock	2200	600	Health	2650	600
Emotion	3550	800	Travel	3260	800
Entertainment	4562	800	Medicine	2403	600
Political	1880	300	Commodity	3230	800

The experiments include two parts: 1) Comparison with that of 4 other algorithm on tag retrieval; 2) The impact analysis on the parameter to examine the overall recommending performance of our model.

Our method is denoted as *TF*Clarity* to demonstrate the effectiveness of our proposed approach, we compare it against the following four methods: *TF*, *TF*IDF*, *TF*Clarity*, *TF*IDF*Clarity*. For each scheme, we select top {1, 3, 5, 7, 9, 11} words with the highest scores as users' tags.

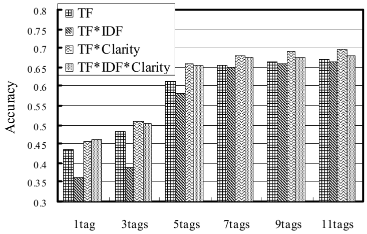

Fig. 1. Accuracy on F1 scores

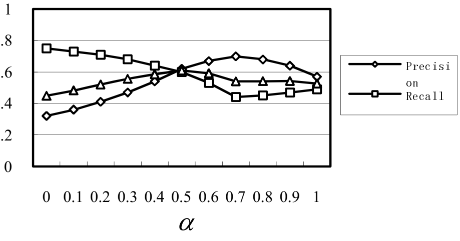

Fig. 2. Different α for our algorithm accuracy

From Figure 1 we can make the following observations. First, more tags (from 1 tag to7 tags) lead to better recommendation algorithm accuracy. When more than 7tags are chosen the improvement becomes minor. Second, among all tag selection methods, our method outperforms the others in most runs with three tags or more labels are marked by user. For one tag, *TF*IDF*Clarity* is the best scheme. However, if only one tag is obtained, the recommendation accuracy is very poor. Therefore, *TF * Clarity* is chosen as our method for tag retrieval.

Parameters α is the most important parameter in our method. It is used to balance the effects of inner and outer tag correlations. Figure 2 reveals that the experimental results match favorably with our hypotheses and encourage us to further explore the reasons. First, the performance of our algorithm is greater than that of either considering inner or outer correlation in isolation, indicating that that an optimal performance comes from an appropriate combination of both the inner and outer correlation among tags. Second, when parameter α is 0.5, our recommendation algorithm shows best performance, which means multi-tag inner correlation is as important as outer relationship. Third, when parameter α is 1, the performance of our algorithm is better than the value of parameter α is 0.

4 Conclusions and Future Work

In this paper, we explore the performance of a multi-tag correlation based approach for recommending microblog posts. Given the short nature of the posts and no background knowledge source, both inner and outer correlations among tags are investigated. Nevertheless, microblogs (and possibly other short texts as well) offer several other information that we have not yet discussed or explored. Future work aims at finding proper ways of adding different information.

Acknowledgement. This work is supported by the National Natural Science Foundation of China (No.61363058, 61163039,61165002), Youth Science and technology support program of Gansu Province (145RJZA232, 145RJYA259) and the open fund of Key Laboratory of intelligent information processing Institute of computing technology of Chinese Academy of Sciences (IIP2014-4).

References

1. Sun, A.: Short text classification using very few words. In: Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 1145–1146. ACM, Portland (2012)
2. Sun, Y.X., Ma, H.F., Shi, Y.K., Cui, T.: Self-adaptive microblog hot topic tracking method using term correlation. Computer Applications **34**(12), 3497–3501 (2014)
3. Tang, J., Wang, X., Gao, H., Hu, X., Liu, H.: Enriching short text representation in microblog for clustering. Frontiers of Computer Science **6**(1), 88–101 (2012)
4. Zhou, X., Wu, S., Chen, C., Chen, G., Ying, S.S.: Real-time recommendation for microblogs. Information Sciences **279**, 301–325 (2014)