ISSN 2616-5775 Vol. 4, Issue 2: 76-82, DOI: 10.25236/AJCIS.2021.040214

Video topic detection on Micro-Blog using Relational Topic Model

Yiping Wang*, Tong Wu, Gaoxu Li, Qing Wan

Media Engineering College, Communication University of Zhejiang, Hangzhou, China
**Corresponding author email: 3242782698@qq.com*

Abstract: *To enhance the performance of the personalized tag recommendation method, a microblog video topic discovery algorithm based on common knowledge atlas analysis was proposed. Firstly, graphical form was used for expressing latent local expression in the microblog video topic and top-k similar user discovery of users represented by user topic distribution. Then, the frequency of all the tags occurring in these users was calculated and the tags most relevant with users were recommended. Next, to mine potential topic information, enhanced cosine similarity RTM model with penalty term was used for naming the tag of topic for microblog video, which enhanced the impact of united modeling on tag generation for potential topic and might discover the relationship between global tab and topic. At last, the real experiment result showed that the recommended method was superior to typical tag recommendation algorithms of several selected microblog video topics. Meanwhile, the validity of the algorithm was verified.*

Keywords: *video topic detection; relation topic model; Micro-Blog*

1. Introduction

In the public opinion analysis, video topic detection is an arduous task [1]. Through video topic detection about topic information of microblog video, we can get much useful information [2~3]. For example, in e-commerce, the company can popularize the product via website, blog or social network [4]. Each transaction is carried out online. Whenever a new product is issued, people will view the information, give comments to express their opinions. Thus, video topic detection plays an increasingly important role in mining network information [5].

Some video topic detection methods have been proposed for the visual language. Literature [6] used behavior modeling approach for classification and recognition of sarcasm language information of microblog video topic. The author analyzed complicated expression form of topic expression. Literature [7] extracted subjective data from the customer's comments at e-commerce website of Amazon, and built a model distinguishing non-sarcasm and sarcasm on a basis of six models including POS n-grams, interest analysis, positive/negative spectrum analysis, video topic detection, joy analysis. Literature [8] looked for sarcasm sentences and extracted clue according to the three concept layers from low level to high level and the discourse features: signature characteristic (pertinence, anti-truth, space-time compression), topic scene (activation, image and joy) and burstiness (time disequilibrium and context disequilibrium). However, another research focuses on the comprehension of metaphor. Literature [9] adopted two-step analysis method for recognizing metaphor. The first step included the construction of concept and acquisition of semantic theme concept of the statement. The second step was the full interpretation of semantic theme concept. Literature [10] followed different methods and used non-supervision method for finding out metaphor expression correlation. Through using verb and noun clustering processing, knowledge was extracted for detecting the similarity of metaphor in larger area. Different from the above algorithm that could only detect the specific type of metaphor language, Literature [11] proposed a solution of analyzing various types of metaphor language, including humor and sarcasm, which adopted text theory to add more functions so as to express favorable and unfavorable irony context, such as the scenes of preliminary, structural ambiguity, syntactic ambiguity, semantic ambiguity, polarity, burstiness and emotion.

The above research focuses on the interest in solving the problem and emphasizes on the vocabulary level. Thus, the objective of this study is to find a new method to identify the concept of image. Here, a statistical method is used for providing an universal model. It is easily expanded to the description of other types of graphic device.

ISSN 2616-5775 Vol. 4, Issue 2: 76-82, DOI: 10.25236/AJCIS.2021.040214

2. Video topic detection frame on microblog

Let $G = (V,E)$ be an undirected diagram of microblog video topic [12~13], or the network and peak set be V and the edge set be E. For example, if several sides are connected to the same peak, it is called multiple edges, and Fig. G is called multigraph. The side is added with a peak and returned itself, which is called a circulation. The simplified graph has no such circulation, or multiple sides. The number of peak of G is usually expressed as n, and it is sorted and numbered. The number of sides of G is usually expressed with m and the side usually has corresponding size. The degree k_i of peak is equal to the number of peak i in the local scope where it is located (microblog video topic). In other words, it closes to the number of nodes. The peak with degree 1 and the connection edge of its sole event are called leaf nodes. The density of network d without circulation or multiple edges is the proportion of the number of edges accounting for the largest quantity of possible edges:

$$d = \frac{2m}{n(n-1)} \tag{1}$$

Let U be the sub-network of G, the cut set of U is the set of all the edges of connection endpoint in the network edge set E located at U. If $U = \varnothing$ or $U = V$, the cut set of U can be ignored. When all the negligible cut set is void, the graph is called to be connected.

Whereas the microblog video topic community detection has deficiency, Literature [14] has built a microblog video topic network of R-C model. The definition is as below [15]:

Definition 1: (X community) If the community object is X, the community is called X community. For example, for microblog video topic community, if the community is a user set, it is a user community; if the community element is node, it is a node community, etc.

Microblog video topic community usually includes three components: user set U, relation set L and related contents T generated by U (involving in topic and comments of microblog video). Microblog video topic community is usually expressed as $S = (U,L,T)$. Fig. 1 shows the content and corresponding relationship diagram of true microblog video topic: $U = \{U_1,U_2,U_3\}$ is the user set, $L = \{L_1,L_2\}$ is the relation set, $T = \{T_1,T_2,T_3\}$ is the content set of microblog video topic, and it is the transfer bond of microblog video topic content T.

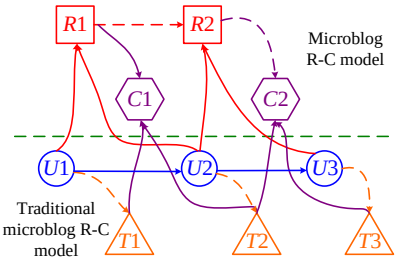

Fig. 1. Schematic diagram of microblog video topic model

Definition 2: (weight set) For the given set $A = \{a_1,a_2,\cdots,a_m\}$, the above set elements all include weight. For example, the weight of individual a_i for element I is w_{a_i}, then A is called weight set, which can be expressed as $A = \{\{a_1,w_{a_1}\},\{a_2,w_{a_2}\},\cdots,\{a_m,w_{a_m}\}\}$.

Definition 3: (weight intersection) If the network has two-weight form: $A = \{\{a_1,w_{a_1}\},\{a_2,w_{a_2}\},\cdots,\{a_m,w_{a_m}\}\}$ and $A = \{\{b_1,w_{b_1}\},\{b_2,w_{b_2}\},\cdots,\{b_m,w_{b_m}\}\}$, then the intersection of Set A and Set B is: $A \cap B = \{(c,w_c)\}$. If $c = a_j = b_j$, then, we get $w_c = \min(w_{a_i},w_{b_j})$.

For microblog video topic network, if the interest set of user U_i is expressed based on weight set I_i, then the characteristic C_x of relation R_x between user U_i and user U_j can be calculated based on Definition 3, and the form is:

$$C_x = I_i \cap I_j \tag{2}$$

Community detection is carried out based on the microblog video topic model of R-C network. The

ISSN 2616-5775 Vol. 4, Issue 2: 76-82, DOI: 10.25236/AJCIS.2021.040214

model may consider the user content and user relation comprehensively so as to enhance the community discovery performance. Moreover, it fully considers the interest characteristic problem of community. The community detection algorithm frame of microblog video topic with R-C model adopted is shown in Fig. 2.

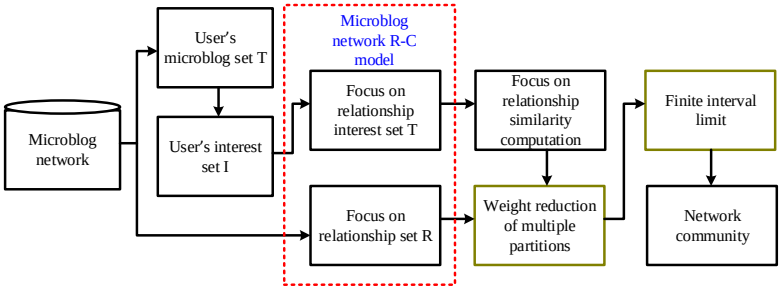

Fig. 2. Video topic detection algorithm frame on microblog

Fig. 2 shows the topic community detection algorithm frame of microblog video with R-C model adopted, and the multi-partition weight reduction and the limit mode of finite interval used in the frame will be expressed in the next section.

3. RTM model for Micro-blog video

3.1 RTM model

A hierarchical model about probability topic was early proposed by Chang et al[16], which is named as relation topic model, recorded as RTM model briefly. Not only does the model consider the content of tag to be recommended, but also incorporate tag relation link into the model, such as the mutual transmission and citation between microblog video topics.

The vector distribution of tag topic is $\theta^{(m)}$, and the vector distribution of API tag topic is $\theta^{(a)}$. The relation link variable between API tag and tag is expressed as $y_{m,a}$, which can be determined according to smoothing parameter λ and vector distribution of tag topic. In the training process, not only does RTM tag recommendation model consider the information relevance between APIs tag and tag *s*, but also considers the vector distribution of topic tag linked with the existing tag.

To depict the relation link between APIs document and tag s more accurately, an enhanced formula of similarity cosine calculation is designed, and the specific form is:

$$\phi\left(y_{m,a}=1\right)=\frac{\sum_{k=1}^{T}\left(z_k^{(m)}z_k^{(a)}\Big/e^{\lambda\left|z_k^{(m)}-z_k^{(a)}\right|}\right)}{\sqrt{\left(z_1^{(m)}\right)^2+\cdots+\left(z_T^{(m)}\right)^2}\sqrt{\left(z_1^{(a)}\right)^2+\cdots+\left(z_T^{(a)}\right)^2}} \tag{3}$$

In the enhanced computing form of similarity cosine index shown in Formula (3), penalty term was increased for amplifying the element difference between vector topic $z^{(m)}$ and $z^{(a)}$. The greater the element difference is, the value of penalty term shown in Formula is higher. Penalty parameter λ is used for representing the penalty degree in the above formula. If $\lambda=0$, the computational formula of enhanced similarity cosine index shown in Formula 2 is transformed to the computational formula of similarity cosine index in a common form. The relation combination of APIs document and its tag, and tag s is taken as the training input of RTM model and the training result of model parameters $\theta^{(m)}$, $\theta^{(a)}$, ϕ, $z^{(m)}$ and $z^{(a)}$ can be obtained by sampling using Monte Carlo-Markov chain form.

3.2 Inference of RTM for topic of microblog video

Through RTM, simple algorithm can be realized in the high-dimension model like topic model, so RTM mode can also use in the RTM model of microblog video topic tag for algorithm simplification. On the premise that *T*, *W*, *H* and *G* are given, the posterior distribution of latent variable needs to be export. *W* means all the words in the dataset, *T* means all the topics of data concentration, H means the tag of all the microblog video topics in dataset, and *G* refers to the flag mark for tag of all the

ISSN 2616-5775 Vol. 4, Issue 2: 76-82, DOI: 10.25236/AJCIS.2021.040214

microblog video topics in the dataset. Thus, the approximation of the following two distribution conditions is given in the example. In addition, the following formula doesn't include hyper-parameter.

Algorithm1. Video topic detection algorithm based on RTM model

Input: preprocessed dataset

Output: list of communities and nodes contained in each community

a) Classify all microblog video datasets according to their users, and form different topic contents. Collection items, user collections nodes, user concern relationship sets, and will be used at the same time. The network composed of items is regarded as a community

b) For each I in items

Using word segmentation software to divide I

Using LDA (latent Dirichlet allocation) model to construct interest set of microblog users uinterests

End For

c) For each I in nodes

For each j in relations

Extract the interest feature set of J from uinterests

For each k in interests of j

Calculating interest weight of K

End For

Get J's interest feature set rinterests

End For

Building RTC model

End For

d) For each m in relations

For each n in relations

If (m,n has common user)

Calculate the interest weight of M, n potential concern relation RW

Else

Continue

Building weighted non directional network community

End For

End For

e) using CNM (Clauset A, Newman MEJ, Moore C) algorithm to make attention relationship of community discovery

Community R with a relationship of concern

f) For each Ri in R

Map directly to node Mi

Form user community U

Output community id vector at this time: Mi

End for

ISSN 2616-5775 Vol. 4, Issue 2: 76-82, DOI: 10.25236/AJCIS.2021.040214

4. Experimental analysis

4.1 Experiment setting

Microblog video topic information usually contains text, picture or video, the message is finite and has 140 characters at most. The paper mainly studied the text of microblog video topic information. ID list of 8000 pieces of microblog video topic information obtained using HTTP request is taken as the training set. There are three types: 5000 groups of sarcasm information, 1000 groups of information with sarcasm meaning and 2000 groups of metaphor information. Because of the nature of information, most sarcasm information, information with sarcasm meaning and metaphor information is negative. The microblog video topic information can be expressed as the following model:

$$Z = \{< t, s > \mid s \in [-5,5]\} \tag{4}$$

Where, Z is a group of microblog video topic information set in the training set, t is the microblog video topic information and s is the scoring of microblog video topic information. The term set extracted from Z can be expressed as below:

$$T_z = \bigcup_{i=1}^{n} t_i = \bigcup_{i=1}^{n} \{\omega_j \mid \omega_j \in t_i\}_{j=1}^{m} \tag{5}$$

In the formula, T_z is the term set extracted from Z, n is the quantity of microblog video topic in the training set, ω_j is term, m is the number of terms extracted from Z.

Cosine similarity is used for evaluating the performance manifestation of the proposed system. For the scoring measured from cosine similarity, the value interval is [0,1], meaning the similarity between our result and the expected result. Firstly, the above two results can be expressed in the form of vector:

$$\begin{cases} R = \{r_1, r_2, \cdots, r_n\} \\ E = \{e_1, e_2, \cdots, e_n\} \end{cases} \tag{6}$$

Where, R is the result obtained using the proposed in the paper, E is the expected result, and n is the quantity of information to be evaluated. The similarity index can be defined as:

$$sim(R,E) = \frac{\sum_{i=1}^{n}(R_i \times E_i)}{\sqrt{\sum_{i=1}^{n}(R_i^2) \times \sum_{i=1}^{n}(E_i^2)}} \tag{7}$$

4.2 Performance evaluation

The above acquired data about microblog video topic information are divided. Because of privacy problem, some microblog video topic information cannot be downloaded. There are 4927 microblog video topics which are divided into two parts. Where, dataset 1 includes 927 microblog video topics, and dataset 2 includes 4000 microblog video topics. Dataset 1 only contains visual microblog video topic information, and it is used for evaluating the recognition capability of visual language. Dataset 2 includes visual and non-visual microblog video topic information. The comparison algorithm selects image video topic detection of content-based microblog video topic, the image video topic detection of decision tree. The comparison result between comparative algorithm and the algorithm proposed in the paper is shown in Fig. 3.

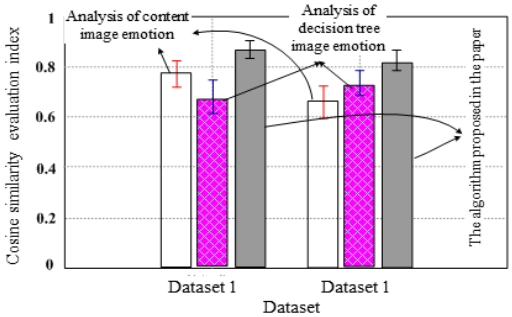

Fig. 3. Algorithm comparison based on cosine similarity index

ISSN 2616-5775 Vol. 4, Issue 2: 76-82, DOI: 10.25236/AJCIS.2021.040214

According to the comparison result shown in Fig. 3, it is known, for cosine similarity index, the algorithm proposed in the paper is superior to selecting image video topic detection of content-based microblog video topic and selecting decision tree case. In Dataset 1, for visual language, the analysis effect of adopting microblog video topic is superior to the image video topic detection method of decision tree. However, in dataset 2, for mixed information, the effect of decision tree is superior to that of image video topic detection method of content-based microblog video topic. In addition, in the algorithm stability, the algorithm proposed in the paper is superior to two comparison algorithms. In dataset 1, for visual language, the stability effect of adopting content-based microblog video topic is superior to that of decision tree image video topic detection method. However, in Dataset 2, for mixed information, the stability effect of decision tree image video topic detection method is superior to that of adopting image video topic detection method of content-based microblog video topic.

The comparison between term scoring result and actual scoring result of dataset 1 and dataset 2 is shown in Fig. 4 and Fig. 5, and the comparison algorithm is selected from Literature [4].

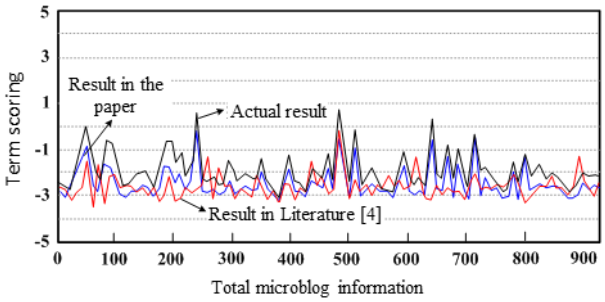

Fig. 4. Comparison of term scoring about Dataset 1

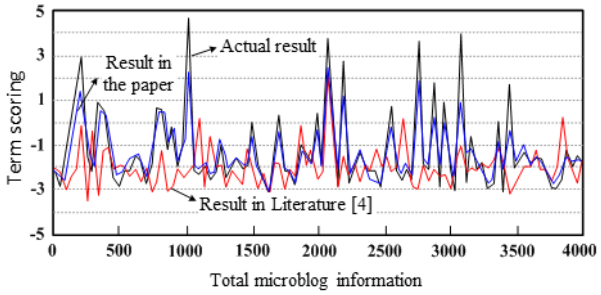

Fig. 5. Comparison of term scoring about dataset 2

According to Fig. 4 and Fig. 5, it is known, in the comparison result of term scoring, the term scoring result of algorithm proposed in the paper approximates more to the true scoring result of term than algorithm in Literature [4], which reflects the advantage of the proposed algorithm in the term scoring result.

5. Conclusion

An interactive community algorithm of microblog video topic community based on dynamic adaptive incremental model. Optimizing the interaction of members in each community and using greedy algorithm for searching the candidate optimal community effectively can measure the difference in intra-community and inter-community interaction. The proposed algorithm enhances the performance of the algorithm. The next step of research direction is: (1) the optimization and improvement of greedy algorithm; (2) algorithm performance testing in a larger dataset; (3) online and real-time analysis of data using software development.

Acknowledgement

Supported by “College Students Innovation and Entrepreneurship Training Program” (Grant No.202011647015, No.202011647017)

ISSN 2616-5775 Vol. 4, Issue 2: 76-82, DOI: 10.25236/AJCIS.2021.040214

References

[1] Lim KH, Datta A. Following the follower: Detecting communities with common interests on Twitter. In: Proc. of the 23rd ACM Conf. on Hypertext and Social Media. New York: ACM Press, 2018. 317-318. [doi: 10.1145/2309996.2310052]

[2] Liben-Nowell D, Kleinberg J. The link-prediction problem for social networks. Journal of the American Society for Information Science and Technology, 2017, 58(7):1019-1031. [doi: 10.1002/asi.20591]

[3] Dourisboure Y, Geraci F, Pellegrini M. Extraction and classification of dense communities in the Web. In: Proc. of the 16th Int'l Conf. on World Wide Web. New York: ACM Press, 2017. 461-470. [doi: 10.1145/1242572.1242635]

[4] Tang L, Wang X, Liu HF. Uncoverning groups via heterogeneous interaction analysis. In: Proc. of the Ninth IEEE Int'l Conf. on Data Mining. Miami: IEEE, 2019. 503-512. [doi: 10.1109/ICDM]

[5] Asur S, Parthasarathy S, Ucar D. An event-based framework for characterizing the evolutionary behavior of interaction graphs. ACM Trans. on Knowledge Discovery from Data (TKDD), 2019, 3(4):16. [doi: 10.1145/1281192.128129]

[6] Richardson M, Domingos P. Mining knowledge-sharing sites for viral marketing. In: Proc. of the 8th ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining. New York: ACM Press, 2016. 61-70. [doi: 10.1145/775047.775057]

[7] Iyer G, Soberman D, Villas-Boas JM. The targeting of advertising. Marketing Science, 2015, 24(3):461-476. [doi: 10.1287/mksc.1050.0117]

[8] Kaplan AM, Haenlein M. Two hearts in three-quarter time: How to waltz the social media/viral marketing dance. Business Horizons, 2017, 54(3):253-263. [doi: 10.1016/j.bushor.2011.01.006]

[9] Larsson AO, Moe H. Studying political microblogging: Twitter users in the 2010 Swedish election campaign. New Media &Society, 2018, 14(5):729-747. [doi: 10.1177/1461444811422894]

[10] Lim KH, Datta A. Finding twitter communities with common interests using following links of celebrities. In: Proc. of the 3rd Int'l Workshop on Modeling Social Media. New York: ACM Press, 2018. 25-32. [doi: 10.1145/2310057.2310064]

[11] Jansen BJ, Zhang MM, Sobel K, Chowdury A. Micro-Blogging as online word of mouth branding. In: Proc. of the 27th Int'l Conf. on Extended Abstracts on Human Factors in Computing Systems. New York: ACM Press, 2019. 3859-3864. [doi: 10.1145/1520340.1520584]

[12] Kleinberg JM. Authoritative sources in a hyperlinked environment. Journal of the ACM (JACM), 2017, 46(5):604-632. [doi: 10.1145/324133.324140]

[13] Blei DM, Ng AY, Jordan MI. Latent dirichlet allocation. Journal of Machine Learning Research, 2015, 3:993-1022.

[14] Steyvers M, Smyth P, Rosen-Zvi M, Griffiths T. Probabilistic author-topic models for information discovery. In: Proc. of the 10th ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining. New York: ACM Press, 2014. 306-315. [doi: 10.1145/1014052.1014087]

[15] McCallum A, Wang XR, Corrada-Emmanuelm A. Topic and role discovery in social networks with experiments on Enron and academic email. The Journal of Artificial Intelligence Research, 2017, 30:249-272.

[16] Pathak N, Delong C, Banerjee A, Erickson K. Social topic models for community extraction. In: Proc. of the 2nd SNA-KDD Workshop 2008. Las Vegas: ACM Press, 2018.