Optimization of LDA text microblogging recommendation algorithm based learning to rank

Yuxiang Xu

Dep. of Management, Hefei University of Technology, Tunxi Road, Hefei City, Anhui Provence, PC230009, China

346591653@163.com

Keywords: Personalized recommendation; LDA; Short text; Microblogging; TF-IDF

Abstract. With the recent rise of web3.0 hot, personalized recommendation social networks become an important aspect of research. Social networks on behalf of the domestic microblogging abnormal hot, more and more domestic and foreign research applied over microblogging. Because the characteristics of micro-Bo short text, LDA topic model is more applicable to micro-blog user's interest analysis. Firstly, the use of the network topology, 10 to find a candidate set target users interested users, and then by LDA microblogging users of potential interest analysis to get a point of interest microblogging users, which are interested in the candidate set recommended target users . Experiments show that the method based on TF-IDF with respect to micro-blog content for the word to get user interest method more effective and efficient.

Introduction

Microblogging occupy Web2.0 in a very important position, has become an important part of the Internet everyday applications, its fast and flexible manner has gradually changed the way people communicate [1, 2]. Through the microblogging people to exchange information or express their views, while microblogging text by the length limit, the information contained limited. Thus microblogging short text such as the classification study found that for hot events, personalized recommendations and other fields have great research value [3-6].

This article is for users to obtain information to help them source user recommendation information of interest to them. Because microblogging own characteristics, in a micro-Bo 140 words or less, microblogging text length is short, it contains less information. Leading to the result of traditional build user interest model method based on word frequency is user model data sparse, high latitude, it is recommended not effective, so the need for semantic short text microblogging expansion. Topic model is a generative model of unsupervised learning, to identify potentially large number of documents in the collection of information for such a short text microblogging. Firstly, through the microblogging Followers / followers is the network topology to obtain the target user's recommended candidate set, and then use the microblogging short text expansion algorithm semantic expansion target user and the candidate set of users tweets, according to LDA model Construction of the target user and user interest vector candidate set, finally recommended to the user based on matching target users interested in the results.

Related theory and technology

Text categorization

Text mining text classification is an important application [7]. It can be divided manually sorting and automatic classification. Representative manually classify the early Yahoo and now ODP, namely the Open Directory system, the largest on the Internet currently produced manually retrieve the classification system).

Another way is automatic text classification categories. Automatic classification is the first to have been marked by the label classification training set to learn, generate model data [8]. Automatic text classification from real, high accuracy, and because it is computer for training and classification and

therefore the speed faster than manually classified, is the current mainstream text classification. Automatic classification process was shown in Figure 1.

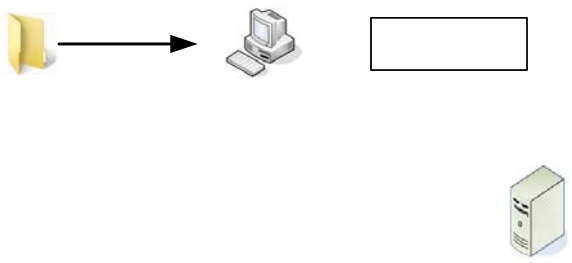

Figure 1. A schematic diagram of automatic classification process

Similarity calculation document

Text similarity is the basis of text mining. You need to text or text similarity between the distance in some research or application areas such as information retrieval, classification or clustering so. Because traditional vector spatial model (VSM) to a text; $x=(x_1, x_2 \ldots x_t)$ represented as t-dimensional vector, so butane - distance vector space model can be represented as a vector between two several kinds of computing a metric which is widely used distance measure is as follows:

Euclidean distance:

$$Dis(x,y) = |x-y| = \sqrt{\sum_{k=1}^{t}(x_k - y_k)^2} \tag{1}$$

Vector inner product:

$$Sim(x,y) = x \bullet y = \sum_{k=1}^{t}(x_k \bullet y_k) \tag{2}$$

Cosine similarity:

$$Sim(x,y) = \frac{x \bullet y}{|x| \cdot |y|} = \frac{\sum_{k=1}^{t}(x_k \cdot y_k)}{\sqrt{\sum_{k=1}^{t} x_k^{\ 2}} \cdot \sqrt{\sum_{k=1}^{t} y_k^{\ 2}}} \tag{3}$$

Category positive examples of precision and recall are defined as follows:

$$precision = \frac{TP}{TP+FP} \tag{4}$$

$$recall = \frac{TP}{TP+FN} \tag{5}$$

LDA model

LDA for the text implicit semantic analysis, unsupervised purpose is to find a method of learning implicit semantic dimension from the text, that topic. It is based on the same principal characteristics of workers in the text to find the words so the topic structure of the text. It is a typical bag of words model, it considers each document that contains a lot of words, by-word, the word is not between

words and the order of a document can contain many themes, and text 60 profile every word is by the subject document contains generated. LDA generation process was shown in Figure 2.

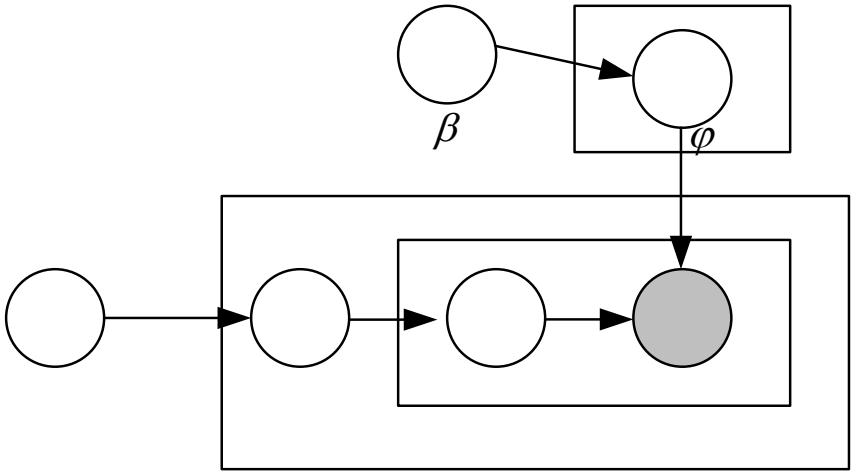

Figure 2. LDA generation process

Experiments and results

Experiments for information request user recommendation information source users, who need to obtain access to information users, based on such evaluation to select the target user, that person access to information.

$$IsInformationSee\ker(u) = \frac{\frac{followers(u) - followees(u)}{followers(u) + followees(u)} + 1}{2} \tag{6}$$

This experiment target user, you should try to select Information Seeker value of less than 0.5. According to the index selection target user 200 people in Sina Weibo, the micro-Bo quantity, number of followers, and the number of fans that features 200 targets users as follows (Figure 3):

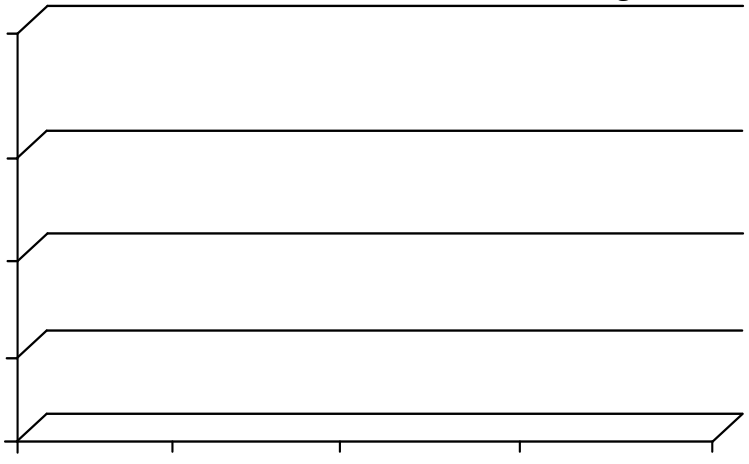

Figure 3. Target user feature map

Use Sina Weibo API and user interface to get the attention of the user target user's attention all published microblogging. While taking advantage of the data in the classroom microblogging manual classification obtained is training to be trained LDA model. Due to the use of distractors API data obtained microblogging complex, the need for micro-blog data obtained pretreatment, including the removal of word less than 10 characters in it, and stop words, emoticons, @ goals.

For the above preset number of themes, selection of training samples of the highest classification accuracy at each of the three groups of parameters modeling topics, which classify the test data, and the results are shown in Table 1, Table C and gamma are to 2 to the end of a few.

Table 1. The result of parameters space searching

Subject number	Log2 (C)	Log2 (gamma)	Classification accuracy of training samples(%)	Test sample classification accuracy(%)
4	13	1	69.4942	80.719
4	5	3	69.3872	80.9368
4	-1	3	69.3604	80.61
6	9	1	71.7635	80.2832
6	11	1	71.7368	80.6922
6	3	3	71.7368	79.6296
8	7	3	71.1508	80.9368
8	9	1	70.9909	81.4158
8	9	3	70.9643	80.9368
10	7	3	71.4971	81.5904

Summary

Leading to the result of traditional build user interest model method based on word frequency is user model data sparse, high latitude, it is recommended not effective, so the need for semantic short text microblogging expansion. Topic model is a generative model of unsupervised learning, to identify potentially large number of documents in the collection of information for such a short text microblogging. 50 Firstly, through the microblogging Followers / followers is the network topology to obtain the target user's recommended candidate set, and then use the microblogging short text expansion algorithm semantic expansion target user and the candidate set of users tweets, according to LDA model Construction of the target user and user interest vector candidate set, finally recommended to the user based on matching target users interested in the results.

References

[1] Chen H, Cui X, Jin H. Top-k followee recommendation over microblogging systems by exploiting diverse information sources[J]. Future Generation Computer Systems, 2016, 55: 534-543.

[2] Shin D, Cetintas S, Lee K C, et al. Tumblr blog recommendation with boosted inductive matrix completion[C]//Proceedings of the 24th ACM International on Conference on Information and Knowledge Management. ACM, 2015: 203-212.

[3] Ifrim G, Shi B, Hurley N J. Learning-to-Rank for Real-Time High-Precision Hashtag Recommendation for Streaming News[C]//25th International World Wide Web Conference, Montreal, Canada, 11-15 April 2016. ACM, 2016.

[4] Li L, Zhang R. Recommended Study of the Flow of Information based on TF-IDF[J]. International Journal of Hybrid Information Technology, 2015, 8(8): 191-200.

[5] Yu J, Zhu T. Combining long-term and short-term user interest for personalized hashtag recommendation[J]. Frontiers of Computer Science, 2015, 9(4): 608-622.

[6] Doulamis N D, Doulamis A D, Kokkinos P, et al. Event Detection in Twitter Microblogging[J]. 2015.

[7] Creamer G G. Can a corporate network and news sentiment improve portfolio optimization using the Black–Litterman model?[J]. Quantitative Finance, 2015, 15(8): 1405-1416.

[8] Wu M S. Modeling query-document dependencies with topic language models for information retrieval[J]. Information Sciences, 2015, 312: 1-12.