Advanced Science and Technology Letters
Vol.82 (SoftTech 2015), pp.6-10
http://dx.doi.org/10.14257/astl.2015.82.02

Recommended model of information flow based on TF-IDF

Liuqing Li, Rui Zhang

Huanghuai University, Henan, China
Henan Agricultural University,Henan,China
E-mail: 943360673@qq.com

Abstract. This article constructed personalized recommendation models in on-line social streams based on ties strength, topic relevance and trust dimensions. The experiments on the Sina blogs data showed that the proposed method could reduce the ranks of irrelevant tweets effectively and achieve better performance than several baseline methods based on cosine and hash tags

Keywords: Information filtering, Personalized, TF-IDF, Micro-blogs

1 Introduction

Micro-blog as a network platform of information sharing and communication, has been widely used in recent years. How to let users get micro-blog content of interest in massive micro-blog has become the new research directions of a micro-blogging platform. Most of the current recommendation system for a number of micro-blog users to recommend , and for the micro-blog platform , because of the shorter length of the text micro-blog, diverse interests of user, therefore the effect of recommend is not ideal for the user[1] . Micro-blog lets people obtain real-time and vast amounts of information through a virtual network. Now popular micro-blog website, Twitter abroad, domestic such as sina micro-blog etc.. There are many study on the micro-blog, such as finding the influence of the highest users in micro-blog; advertise propaganda and investment in the micro-blog; authenticity test micro-blog information, so as to prevent the spread of rumors; and classification of micro-blog information .

Unlike previous social recommendation always use social networking features to recommend new content or links, information retrieval model presented in this paper is recommended for users micro-blog information to meet their interests and preferences and shows a new homepage [2]. Users can find other related micro-blog information, and they can also discover new related user in this homepage.

ISSN: 2287-1233 ASTL

2 Problem and research framework

Social media not only faces the problem of information overload, also faces challenges of information hiding. Social search engines can provide a simple information retrieval approach, however, these engines cannot predict what users like and information of interest to the exhibition. In order to solve this problem, social search engine and personalized recommendation algorithm combining is necessary.
Personalized micro-blog recommendation is to estimate how much a micro-blog interest value for each user, so how do you handle user's interest is one of the key issues in this article. Users are mainly 3 actions on micro-blog: follow other users, publish micro-blog and forwarded to other users of micro-blog. The user's interests can be found by analyzing the user's behavior, while micro-Bo of user is the most direct reflection of users ' interests, it is also the focus of this paper [3].

3 Commendation based on information content characteristic

Commendation based on information content characteristic is a method which is often used by full text commendation system as well as socialized net information stream commendation. The most important aspect of it is Topic Relevance calculating. When calculating topic relevance, firstly, we should establish characteristic vector of user interest model, that is, to any of the user u, his interest vector can be shown as the formula:

$$V_u = \{v_u(w_1), v_u(w_2), .v_u(w_i)\} \quad (1)$$

i —the extractive topic number of the information in all the microblog users released

$v_u(w_i)$ —the interest degree of user u to the topic w_i

The expression $v_u(w_i)$ can be calculated by the arithmetic TF-IDF:

$$v_u(w_i) = TF_u(w_i) \square IDF_u(w_i) \quad (2)$$

In the formula, the expression $TF_u(w_i)$ shows the frequency of the key word w_i in information user u released.

$$IDF_u(w_i) = \log\left(\frac{U}{u}\right) \quad (3)$$

U —representing the total number of all the users

u —representing the number of the users who have used the key word w_i at least once

Similarly, we can establish topic vector of given information. What the difference is TF have been expressed to show the frequency of certain key word mentioned in all microblog information. At last, we can use cosine similarity arithmetic to calculate the similarity between user interest vector and topic vector of information content.

4 Information retrieval model

In the information retrieval model for arbitrary user u and micro-blog $t(t \in T)$, the values $I_u(t)$ that are used to gauge user u interest in micro-blog t ,user's interests can be drawn from the historical micro-blog information.

Topics related measure has been widely used in the recommendation system, the flow of information in social networks have begun to successfully use the sort field , and superior to the classical metrics topic model LDA based on the type of text processing micro-blog according to research topics $TF-IDF$ based on the correlation metrics [4]. In order to calculate the user's topic and a micro-blog information flow correlation vector, first get a topic from the user 's micro-blog post , and then build a similar vector reaching the user 's micro-blog information flow, the final score is calculated on correlations choice the most suitable micro-blog recommended to the user[5].This part of the collection R_u published by the similarity, and calculate the user's micro-blog and micro-blog flow of information in any subset of T to estimate the value $I_u(t)$. First, it should be given a random sequence micro-blog X, and define a bag of words vector $BT(X) = w_1, w_2, \cdots, w_n$, the words W_i should be contained in at least one t in micro-blog. However, a micro-blog information flow usually contain a small amount of micro-blog and each micro-blog are less than 140 words, the bag of words vector set up in this way is very sparse, and the similarity cannot be compared. Based on the bag of words in the vector by merging some words make the bag of words Vector final richer. Thus , the model also defines a sequence of words vector $BP(X) = P_1, P_2, \cdots, P_n$, at least ,sequence p_i pi words appear in the same micro-blog t . For a micro-blog sequence X in the case of words and word sequences given two scoring formulas.

Definition 4 is given a word w and a micro-blog information flow $Xscore(w, X)$ represent w value in a given X.

$$\begin{cases} tscore(w, X) = TF_T(w, X) \cdot IDF_T(w) \\ IDF_T(w) = \log \dfrac{|T|}{DF_T(w)} \end{cases} \tag{4}$$

$TF_T(w, X)$ is micro-blog number, which contains the words w in micro-blog sequence X, and DF_T is micro-blog number, which contains the words w in micro-blog information flow T. $TF-IDF$ is a statistical method used to evaluate a set of words for a file or a corpus in which the importance of a document . $TF-IDF$ can be seen in the experimental section in dealing with some long blog , which has more advantages. Definition 5 given a set p and a micro-blog X, $pscore(p, X)$ represent the word p in a given x.

$$\begin{cases} pscore(p, X) = TF_T(p) \cdot IDF_P(p) \\ IDF_p = \log \dfrac{|T|}{DF_P(p)} \end{cases} \tag{5}$$

TF is the number of micro-blog in information micro-blog flow, which contains word sequences of p, $DF_P(p)$ is the number of micro-blog in information micro-blog flow T, which contains word sequences of p. Micro-blog is usually very short document which contains a few sentences, but a few words or a combination of the words appear at the same time, there is a high probability in two micro- Bo. Finally we combine the definitions 4 and definitions 5 to get the score formula of estimated two Micro-blogs sequence similarity.

Definition 6 give two micro-blog information flow X_1 and X_2, we define the similarity as $\text{int}_A(X_1, X_2)$.

$$\text{int}_A(X_1, X_2) = (1-\lambda)\sum_{w \in BT(X_1) \cap BT(X_2)} tscore(w, X_2) \quad \lambda \sum_{p \in BP(X1) \cap BP(X_2)} pscore(p, X2) \tag{6}$$

In the calculation of similarity, if $\lambda = 0$ then only consider a single word, on the contrary, if $\lambda = 1$ only consider the situation of the words. In the calculation of int_λ, which can use different λ values to determine the optimal scheme. After a given user u and the published micro-blog set R_u, for any value of $(\lambda_1, \lambda_2, \cdots, \lambda_i)\lambda \in [0,1]$, the definition of $I_u(X) = \text{int}_\lambda(X, R_u)$,which can estimate the interest degree of user u for micro-blog sequence X. So that this can be calculated the value int_λ in the maximum for R_u to choose the micro-blog information flow T, and micro-blog k is recommended to the user.

5 Conclusion

This paper presents a method to measure user's interest in micro-blog, and combine collaborative filtering for micro-blog users to recommend more interesting kind of micro-blog information, and a recommendation system can be applied to many different scenarios, which provides users with more interesting home pages, and, experimental results show that the recommended method in this paper has a high accuracy rate. In the future, we will consider the user or micro-blog features to further improve the precision of micro-blog recommended [6]. In the future, we will further study the relationship between micro-blog users and micro-blog cold start System.

References

1. Sebastian, F.: Machine Learning in Automated Text Categorization. ACM Computing Surveys (2002), Vol.34 (1), pp: 24-4 (2002)
2. Zhuang, H.: The Knowledge grid. World Scientific. Singapore. (2004).
3. Ya, J.: Semantic Link Network Builder and Intelligent Browser. Concurrency and Computation：Practice and Experience.(2004).vol.16(14).pp:1453-1476.

4. Chen, J., Chi, E.: Speak little and well: Recommending Conversations in Online Social Streams. Proceedings of the 2011 annual Conference on Human Factors in Computing Systems. ACM, (2011), pp: 217-226.
5. Joinson, A. N.: Looking at, looking up or Keeping up With People: Motives and Use of Facebook. Proceedings of the SIG-CHI conference on Human Factors in Computing Systems. ACM, (2008) ,pp:1027-1036.
6. Mika. P., Tununarello, G.: Web Semantics in the Intelligent Systems, IEEE, (2008), Vol. 23.pp:82-87

