Information Processing Letters 136 (2018) 44–48

Contents lists available at ScienceDirect

Information Processing Letters

www.elsevier.com/locate/ipl

Background feature clustering and its application to social text

Chuangying Zhu, Junping Du *

Beijing Key Laboratory of Intelligent Telecommunication Software and Multimedia, School of Computer Science, Beijing University of Posts and Telecommunications, Beijing 100876, China

ARTICLE INFO

Article history:
Received 26 April 2017
Received in revised form 24 March 2018
Accepted 24 March 2018
Available online 4 April 2018
Communicated by Jinhui Xu

Keywords:
Algorithms
Social media
Short text
Spatio-temporal characteristics
Feature fusion

ABSTRACT

The demand for and dependence on social networks make the virtual world returning to real life along with real time, actual space and concrete events. To create joints from online topics to offline activities, a spatio-temporal and structure feature model is established by fusing the background information, and then the topics are investigated by clustering the keywords. Compared with the traditional methods, background future clustering keeps the constrains caused by data sparseness and spatio-temporal dependence off, and can be used for unpredictable activities discovery.

© 2018 Elsevier B.V. All rights reserved.

1. Introduction

Text semantic analysis aims to get the topics' distribution and the associated structure distribution from the contents. As a major technique focusing on semantic analysis, the topic method represents the structure information of a document through words frequency distribution got by keywords quantification using bag of words model, and gets the meaning of the content by mapping the high-dimensional vector into low-dimensional semantic space, to bridge the gap between the syntax and the semantics of the document finally [1]. For popularizing semantic analysis technique in wider field, a variety of extended topic models have been investigated according to different environments [2–4].

Nevertheless, the traditional models and improved approaches train the topics classifier are often based on pre-collected corpus data, which limits the application of topic model in social media. Because we cannot know what will happen, and what new keywords will occur in social network in the future. Furthermore, the social text is too short to be extracted semantics by the conventional method.

Considering that, social contents depending on time and their structure, various of solutions based on their associated information have been proposed [5–8], which aims to solve the issue of semantic sparsity. This paper focuses on the time-dependent problem, the sparsity problem and the unpredictable problem, further proposes a short text clustering method based on background feature fusion.

This paper is organized as follows: Section 2 introduces associated feature extraction and feature fusion method. In Section 3 some experiments for social texts research will be displayed. Finally, we make a conclusion for entire paper.

2. Data feature extraction and fusion

In detail, process of data pretreatment is Chinese word segmentation, part-of-speech extraction and stop words removing. Then the keywords can be extracted from original data. Then, the frequency distribution of the keywords in time and space, and the structure relations among the keywords are extracted as the background features of the

* Corresponding author.
E-mail addresses: zhuchuangying@qq.com (C. Zhu), junpingdu@126.com (J. Du).

https://doi.org/10.1016/j.ipl.2018.03.017
0020-0190/© 2018 Elsevier B.V. All rights reserved.

social texts. Finally, the social texts are represented as a unified form by background feature fusion.

2.1. Background features fusion

(1) Time feature extraction

Social data own obvious characteristics of time dependence, whose keywords have different semantic distribution in different time slots. It is easy to get time information from the original social contents, and further divide into slots according to requirement, such as several hours or a month as a time slot. Temporal distribution of the short text, expressed as a matrix: $X \in \mathbf{R}^{m\times t}$, can be obtained by counting the number of keywords obtained by preprocessing in different time slots. Where, m denotes the dimension of keywords, t denotes the dimension of time slots, meanwhile, the matrix X is treated as one of the background data for feature fusion.

The same keyword may appears in the same content repeatedly. However, the purpose of time feature extraction aims at getting the temporal distribution characteristics of the keywords, so the same word in the same text needs to be counted only once.

(2) Location feature extraction

Offline activities that are closely related to social contents have obvious characteristics of geographical spatial that bridge the offline life with online topics. The Topics related to space can be obtained by semantic analysis of the social contents with location feature.

Location information in the social network is usually labeled with GPS coordinates or location-tags, such as "main building-BUPT". It is easy to obtain a unified GPS location with the help of Google API interface. Offline activity concentration region can be obtained by GPS location clustering. It may be simplest to divide the map into grids according to the latitude and longitude, and merge the neighbored grids with more data above a threshold to get hot region.

Location distribution of the social text, expressed as a matrix: $Y \in \mathbf{R}^{m\times d}$, can be obtained by counting the number of keywords located in different hot regions. Where, m denotes the dimension of the keywords, d denotes the number of hot regions. Similarly to matrix X, matrix Y is seen as one of the background data for feature fusion.

Hot regions extracted by grids division cannot completely cover the area that the data falls in. Nevertheless, it does not need to calculate the hot regions accurately, because the purpose of location feature extraction aims to get the geographic distribution characteristics of keywords. In this case, only the approximate regions are enough.

(3) Structural association feature extraction

The keywords distribution reflects the content attributes, the relations among keywords reflect the complex attributes of the content, and the co-occurrence of keywords is the main form of text topic distribution. Furthermore, single text in social network is too short to be extracted semantic information.

Two adjacent texts can be treated as one unit that has the same topic, to extract structure association feature, due to the behaviors of response, comment and retransmission in social network are continuation and expansion of the same topic. It should be noted that two different texts with same predecessor might belong to different topics.

The structure relations of keywords expressed as a symmetric matrix: $Zc \in \mathbf{R}^{m\times m}$, can be obtained by calculating the co-occurrence times based on combination of adjacent texts. Where, the element $Zc_{i,j}$ is the co-occurrence times between word i and j.

Structural correlation between keywords i and j can be calculated by the Pearson Correlation coefficient based on matrix Zc.

$$sim(c_i, c_j) = \frac{\sum_k (c_{ik} - \overline{c_i})(c_{jk} - \overline{c_j})}{\sqrt{\sum_k (c_{ik} - \overline{c_i})^2}\sqrt{\sum_k (c_{jk} - \overline{c_j})^2}} \tag{1}$$

where, $c_i = \{Zc_{i,1}, \ldots, Zc_{i,i-1}, Zc_{i,i+1}, \ldots, Zc_{i,m}\}$, i.e. the ith row of matrix Zc except the ith element. $c_j = \{Zc_{j,1}, \ldots, Zc_{j,i-1}, Zc_{j,i+1}, \ldots, Zc_{j,j-1}, Zc_{j,i}, Zc_{j,j-1}, \ldots, Zc_{i,m}\}$, $(i < j)$ or $c_j = \{Zc_{j,1}, Zc_{j,j-1}, Zc_{j,i}, Zc_{j,j-1}, \ldots, \ldots, Zc_{j,i-1}, Zc_{j,i+1}, \ldots, Zc_{i,m}\}$, $(i > j)$, i.e. the elements in the jth row of matrix Zc with the jth element replaced by the ith.

The structure relations among social texts, which is expressed as a matrix $Z \in \mathbf{R}^{m\times m}$, can be calculated using formula (1), which is also a limitation to the following clustering algorithm.

2.2. Data feature representation

The spatio-temporal and structural representation of the keywords can be obtained by fusing the above background data.

A three-dimensional tensor $A \in \mathbf{R}^{m\times t\times d}$ that is derived by collaborative tensor decomposition is used to characterize the feature distribution of keywords. It is assumed that the feature tensor A does exist, and can be decomposed into the product of one core tensor $C \in \mathbf{R}^{k\times k\times k}$ and three matrices $M \in \mathbf{R}^{m\times k}$, $T \in \mathbf{R}^{t\times k}$ and $D \in \mathbf{R}^{d\times k}$ ($A = C \times_M M \times_T T \times_D D$) on the basis of Tucker tensor decomposition theory [9], as shown in Fig. 1.

In ideal status, the background data X, Y and Z can be composed by the following formulas on the basis above: $X = M \times T^{\mathrm{T}}$; $Y = M \times D^{\mathrm{T}}$; $Z = M \times M^{\mathrm{T}}$. To keep the original characteristics of the background data, the core tensor C be set as an identity tensor, i.e. $C_{i,i,i} = 1$ if $1 < i < k$; $C_{i,i,i} = 0$ otherwise. Therefore, the matrix gained by reducing the dimensions of A from any direction is isomorphic to the corresponding background matrix.

Optimization theory, such as gradient descent is used for the solution of matrices M, T and D. Then, the approximations of matrices X, Y and Z can be obtained by mean of different objective functions: $X \approx M \times T^{\mathrm{T}}$, $Y \approx M \times D^{\mathrm{T}}$, $Z \approx M \times M^{\mathrm{T}}$. The loss function to quantize the error of collaborative tensor decomposition is defined as follow:

$$\begin{aligned} f = {} & 0.5\lambda_1 \left\| X - M \times T^T \right\|_F^2 + 0.5\lambda_2 \left\| Y - M \times D^T \right\|_F^2 \\ & + 0.5\lambda_3 \left\| Z - M \times M^T \right\|_F^2 \\ & + 0.5\lambda_4 \left(\|M\|_F^2 + \|T\|_F^2 + \|D\|_F^2 \right) \end{aligned} \tag{2}$$

where, $\|\cdot\|_F$ denotes Frobenius norm. Parameters λ_1, λ_2 and λ_3 are used to control the contribution of each

Download English Version:

https://daneshyari.com/en/article/6874161

Download Persian Version:

https://daneshyari.com/article/6874161

Daneshyari.com