Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 35 (2014) 929 – 936

Procedia Computer Science

18th International Conference on Knowledge-Based and Intelligent Information & Engineering Systems - KES2014

Classifying homographs in Japanese social media texts using a user interest model

Tomohiko Harada[a], Kazuhiko Tsuda*[b,]

[a] Graduate School of Systems and Information Engineering, University of Tsukuba, Bunkyo-ku, Tokyo 112-0012, Japan
[b] Graduate School of Business Sciences University of Tsukuba, Bunkyo-ku, Tokyo 112-0012, Japan

Abstract

The analysis of text data from social media is hampered by irrelevant noisy data, such as homographs. Noisy data is not usable and makes analysis, such as counting estimates, of the target data difficult, which adversely affects the quality of the analysis results. We focus on this issue and propose a method to classify homographs that are contained in social media texts (i.e. Twitter) using topic models. We also report the results of an evaluation experiment. In the evaluation experiment, the proposed method showed an accuracy improvement of 8.5% and a reduction of 16.5% in the misidentification rate compared with conventional methods.

Peer-review under responsibility of KES International.

Keywords: Social media, Twitter, Homographs, Semantic analysis, Topic modelling, Latent Dirichlet allocation

1. Introduction

In recent years, there has been a rapid increase in efforts to generate business strategies, such as marketing strategies and business improvements, by collecting and analysing Big Data. The social network service 'Twitter' has attracted attention as a source of such information. In messages called 'tweets', which can only be 140 characters or less, a user can post thoughts or day-to-day experiences. When a user posts a tweet, the tweet is transmitted from person to person; tweets can be shared among many users. Tweets often include user impressions of purchased products and services, as well as the criteria used to select those products and services. It is increasingly becoming important for businesses to collect and analyse such useful information. However, there is a common problem with noisy data that is contained in results (e.g. tweets) that are collected by keyword searches in the analysis and study of social media, such as Twitter. Such noisy data is unusable for targeted analysis and affects the accuracy of the analytical results. For example, when performing analysis of corporate reputations, if there are homographs, such as the name of another company with the same name, that are included in the results of a keyword search of a company name, this becomes a factor in the analysis, and accuracy is often reduced. In Table 1,847 tweets containing the keyword

* Corresponding author. Tel.: +81-3-3942-6869 ; fax: +81-3-3942-6829.
E-mail address: s1230165@u.tsukuba.ac.jp.

Peer-review under responsibility of KES International.
doi:10.1016/j.procs.2014.08.168

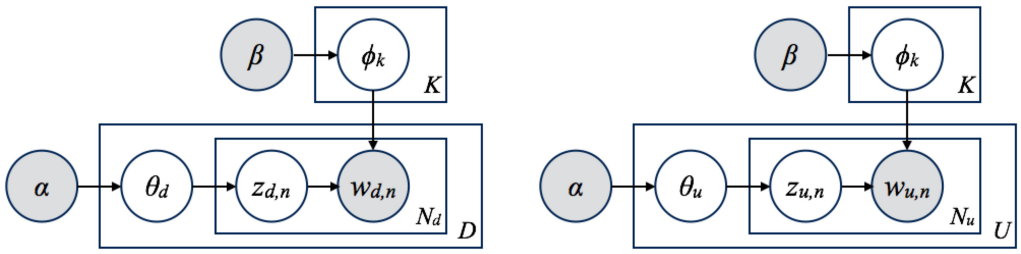

Fig. 1. (a) Simple graphical model of LDA; (b) Graphical model of our LDA.

'apple' in Japanese katakana characters are classified by subject. The Japanese tweets shown were posted from 4 to 11 January 2014. When counting the number of tweets about 'Apple Inc.', the digital consumer electronics and computer manufacturer, after searching for tweets using the keyword 'apple', it is common to encounter tweets with 'apple' used with the intended meaning of fruit, such as 'apple tea' or 'apple juice'. In addition, other companies that may include the word 'apple' will be included in such search results. In general, these unrelated tweets are noisy data. Therefore, it is necessary to classify tweets of interest from search results that include noisy data. In this paper, we focus on this issue and propose a method to classify homographs contained in the text of social media using topic models. We also report the results of evaluation experiments.

Table 1. Examples of the tweets' subjects containing the keyword 'apple'.

Subject	% tweets
Apple Inc.	70
apple tea	4
apple juice	2
other	24
Total	100

2. Topic modeling

Topic modelling [1] has attracted attention as a statistical modelling method that is used to acquire knowledge from large-scale and heterogeneous data. In topic modelling, one document is represented as a mixture of multiple topic information. It has been confirmed that topic modelling can model documents with higher accuracy than a mixture of multinomial distributions represented by one topic one document [2]. In this section, we review latent Dirichlet allocation (LDA) [2], which is a representative topic model that is known to work well. We then review representative studies of applying LDA to the Twitter data analysis.

2.1. LDA: Latent dirichlet allocation

Blei et al. [2] proposed LDA, a technique in which the Dirichlet prior distribution is taken as a prior distribution of the multinomial distribution that represents the topic of a document. The potential of topic modeling has recently attracted attention, and LDA is known to work well. Based on the idea that a document is represented as a random mixture over latent topics, where each topic is characterized by a distribution over words, LDA infers the probability distribution of the topic.

Fig.1.(a) shows the graphical model of LDA, where random variables and parameters are represented by a vertex; their dependencies are represented by a directed edge. The shaded vertex indicates observed variables; the other vertices indicate latent parameters or latent variables. The number written at a rectangle's corner indicates the iterations of the variable in the rectangle. D is the number of documents, K is the number of topics, and N_d is the word count in