International Congress of Information and Communication Technology (ICICT 2017)

Web Topic Analysis of the Two-child Policy in China

Yonghe Lu[a,*], Menghui Zheng[b]

[a]School of Information Management, Sun Yat-Sen University, Guangzhou 510006, China
[b]School of Information Management, Sun Yat-Sen University, Guangzhou 510006, China
* Corresponding authors: zsuluyonghe@163.com

Abstract

Since the Two-child Policy was enacted by The Chinese government in October 2015, two children in a family have become an important question related to people's livelihood and have triggered a wide discussion. As an important platform to discuss and monitor public opinion changes, Sina Weibo plays an important role in public opinion analysis. This paper chooses two-child policy as search topic to crawl 19751 related microblogs in Sina Weibo, uses topic over time (TOT) model of the microblogs to make topic analysis, to extract what people focus on the Internet. The results of analysis show that most people cause extensive concern of the country's follow-up policy, including economy, education and other aspects of the follow-up policy.

Keywords: The two- child policy, Web topic, Public opinion, Topic analysis, Topic over time model

1. Introduction

In October 2015，Chinese government promulgated the two- child policy, which means that every couple can have two children. The policy was formally implemented in January 1, 2016. Widespread concerns and discussions are caused among public because the policy is closely related to the livelihood of people in all aspects. Therefore, it is very significant to analyze public opinions of the policy. By analyzing the contents of microblogs, this paper aims to extract main topics of the policy and pays attention to what people discuss on Sina Weibo.

Nowadays, the analyses of two-child policy mainly focus on theoretical side, such as demography[10], advantages and disadvantages of the policy[11], impact on the policy[12], the case analysis[13], etc. There are rare papers using automatic text analysis to accomplish their researches.

2. Introduction of the algorithm: ToT model

2.1. Related research

In natural language processing, the topic can be regarded as a lexical item probability distribution. The topic model is used to simulate the document generation process, and the parameter estimation is used to get the topic. Therefore, the topic model is to analyze the words in the unstructured text, not only to find out the correlation and similarity between documents, but also to find the implied semantic association between words. Topic model can facilitate us to understand and grasp the Web public opinions of two- child policy, and concerned about its dynamic changes.

Probabilistic Latent Semantic Indexing proposed by Hofmann in 1999[0], and the LDA model (Latent Dirichlet Allocation) proposed by Blei in 2003[0] are the commonly used topic models at this stage. After introducing time parameters into the basic model, some new models were formed. ToT (topic over time) model 错误!未找到引用源。 is one of the improved model, which combines the time, text, and term as observation data, and uses the Beta distribution to model the subject intensity varying over the time to analyze the change of topic intensity. DTM (dynamic topic models)[0] is another one. This model firstly discretizes time into slices so that the word probability distribution and the topic probability distribution under each time slice are dependent on the state of the previous time slice. Thus, the continuity between topics is fully utilized, and the change of the topic content and intensity can be analyzed. Compared with DTM, which is directly finding the dynamic characteristics of the theme, suitable for analysis of online dynamic analysis of the document, ToT model is a static discovery of the subject and time, suitable for analysis of static documents. According to the data collected in this paper, this paper selects the ToT model for topic research.

ToT model is extended by adding author parameter (to generate aTOT model[0]), emotional parameter (to generate eToT model [0]) and other parameters, in order to analyze the relationship between more factors and time.

2.2. Analysis process of ToT model

ToT model is a generation model based on timestamps and the vocabulary of timestamps. It not only determines the model based on the co-occurrence of words, but also takes the time factor into account. The model assigns a timestamp to each word, namely, each word in each document (each microblog) has the same timestamp, and each word in microblog has the same timestamps as its microblog.

The generation process of ToT model can be described by the graph model of Gibbs sampling based on parameter estimation as follows[7]:
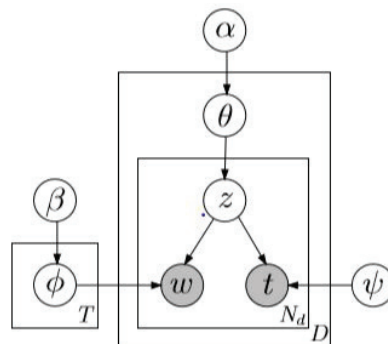

Fig. 1. The graph model of the generation process of ToT model

Where T represents the number of topics, in this paper, T equals 10, that is, this paper sets 10 topics. D represents the number of documents, in this paper, D equals 19751. N_d is the number of words per document. α 、 β are parameters, both of which obey the Dirichlet distribution, this paper sets α = 50 / T, β = 0.1 [7].