Detection of topics from newspaper and its analysis of temporal variations in regions

Taizo YAMADA[12*]
1 Historiographical Institute, The University of Tokyo
2 Collaborative Research Organization for Historical Materials on Earthquakes and Volcanoes, The University of Tokyo
E-mail: t_yamada@hi.u-tokyo.ac.jp

In the paper, we introduce a method of topic detection using topic model for Japanese newspaper and propose to visualize the time change of the detected topics.

In the study, we detected topics from newspapers published by Mainichi Newspapers from 2010 to 2015. There are about six hundred articles (number of characters: about 300 million) in the text data. The data of the article has items such as title, date, posting page and text which was written in Japanese. We performed to extracts nouns as characteristic words of the text. In the extraction, we used Mecab which is morphological analyzer for Japanese and IPAdic as a dictionary. We characterized the text with a latent topic which is hidden in the text and can be detected by LDA (Latent Dirichlet Allocation) which is one of a topic model.

In the topic detection, we assumed that there are 200 topics in the newspaper. There are very diverse topics including politics such as general resignations of the Cabinet and national elections, sports such as the Olympics and the Football World Cup, lotteries, Southeast Asian affairs, Japanese economics, academics, and so on. From them, we noticed earthquake topics and focused on them. In order to grasp the characteristics of the topic, we visualized the change of the frequency of the occurrence and the top words on a monthly basis. In January 2010, there were "Earthquake", "Haiti", "Earthquake disaster", "Victim", "Great Hanshin Earthquake" as its top words in the topic of earthquake. About the year, this topic does not appear so much. However, in March 2011, the topic appeared about 10 times or more than before. The topic in March 2011 was characterized by "East Japan great earthquake disaster", "tsunami", "disaster area", "victim", "shelter" and so on. Furthermore, we introduced a method which can search similar topics to the topic and display them. Here, in order to calculate similarity between topics, we used cosine similarity in which the frequency of word occurrence per topic was used.

Analyzing topics by region helps you to grasp the situation fluctuation in the region. If we investigate the posting position of the article and the topic variation, we can find out the importance of the topic or the article at that time. We believe that incorporating other newspapers and data on the web such as SNS and Wikipedia will enable us to grasp more sophisticated events. This leads to the importance of using big data in area studies.