JAIST Repository
https://dspace.jaist.ac.jp/

Title	テキストマイニングに関する調査研究 [課題研究報告書]
Author(s)	藤井, 晃
Citation	
Issue Date	2017-03
Type	Thesis or Dissertation
Text version	author
URL	http://hdl.handle.net/10119/14183
Rights	
Description	Supervisor:白井　清昭, 情報科学研究科, 修士

JAIST
JAPAN
ADVANCED INSTITUTE OF
SCIENCE AND TECHNOLOGY

Japan Advanced Institute of Science and Technology

Survey of Text Mining Studies

Akira Fujii (1210907)

School of Information Science,
Japan Advanced Institute of Science and Technology

March, 2017

Keywords: Text Mining, Data Mining, Natural Language Processing

In recent years, the methods for acquiring useful knowledge from a huge amount of data, called "Big Data", are paid attention much as up-to-coming technology all over the world. Although there are a lot of types of big data, this project report focuses on text data. Text mining is a method to process a huge amount of unstructured text data and extract useful information and knowledge from it. It is supported by techniques of natural language processing and data mining. In this Internet era, text data rapidly increase day by day. There might be a lot of useful but unknown knowledge in the text data in the world. Therefore, we can expect that text mining will be further studied and developed in future. It is important to summarize the past text mining studies and to understand the current trend of text mining, since it would be helpful to decide future directions.

A goal of this project report is to analyze and summarize the past studies of text mining. Especially, we pay much attention to applicability of text mining technologies. We try to analyze the trend of text mining with respect to "topic" and "medium" of the text used for mining. In this project report, "topic" stands for a domain or a genre of text such as medical, management, economy domain and so on. The previous papers of text mining are classified in terms of the topics of the text used for finding new knowledge, so that we can clarify what kinds of knowledge has been acquired by text mining. On the other hand, "medium" refers to a medium on which the texts are published, such as news on Web, Social Network Services (SNS), news paper article, questionnaire survey, academic literatures and so on. The previous studies are also classified in term of the media, so that we can clarify what types of texts were analyzed by text mining.

Google Scholar is used to obtain a set of the past papers about text mining. The papers written in Japanese ("Japanese papers" hereafter) are searched with a query "TEKISUTO-MAININGU" (a Japanese word that means text mining) by Google Scholar and the top 100 papers are used for the survey. The

Copyright © 2017 by Akira Fujii

papers written in English (“English papers” hereafter) are also searched with a query “text mining” and the top 100 papers are chosen. To classify the papers, twenty-five categories of the topics and twenty-nine categories of the media are defined. First, for each paper, one category of the topics and one category of the media are manually chosen with respect to the text from which the paper attempted to acquire knowledge. Second, distributions of the topics and media in 100 Japanese papers and 100 English papers are investigated. Then we explore difference of the topics and media in the papers of different languages (Japanese or English) as well as the change of the topics and media year by year.

In the Japanese papers, the most frequent topic is “medical and life science”. The number of medial papers is 10, which is 10% of the Japanese papers. Furthermore, there are 7 topic categories that only one paper belongs to, and 5 topic categories that only two papers belong to. Thus we cannot find any bias of the topic distribution. It means that Japanese papers of text mining focused on the text of various topics. As for media, “questionnaire” is the most frequent media (21 papers), followed by “Internet community” (17 papers) and “report” (10 papers). It seems consistent with the fact that there is no bias in the topic distribution, because there are various topics in the “questionnaire”, “Internet community” and “report”.

On the other hand, in the English papers, the most frequent topic is same as in Japanese, that is “medical and life science”. The number of the papers in “medical and life science” is 33, which is 3.3 times of that of Japanese papers. In this category, not only the papers that proposed new methods but also ones that reported the survey and tools for text mining are included. Therefore, we can conclude that text mining on medical science was mainly focused in English papers. As for the media, the most frequent category is “literature” (21 papers). Note that the category “literature” includes the papers that tried to acquire knowledge from the database of technical papers like PubMed and MEDLINE. There is no other category that more than 10 papers belong to. That is, strong bias is found in the distribution of the media. With the fact that the most frequent topic in English papers is “medical and life science”, we can conclude that many English papers of text mining tackled a problem to mine knowledge from medial papers and abstracts in the database. One of the reasons may be fast growth of the number of papers in medical domain. A certain statistics reported that over 500,000 papers were published every year.

Next, we discuss the change of the number of published papers year by year. The largest number of Japanese papers published in a year is 13 in 2004; that of English papers is 11 in 2005 and 2008. The study of text mining was the most active in these years. Even in the active period of the text mining, the number of the papers per year may be unstable. To examine the change of the number of the papers more accurately, for each year, we accumulate the number of published papers during 5 years (between two year before and after), and then check the change of them by years. The highest number of English papers is 45 at the year 2006 (between 2004 and 2008). After that, the number of the papers is decreased. The number of Japanese papers is 40 at the year 2006, 35 at 2007, 41 at 2008, 41 at 2009 and 40 at 2010.

These numbers are higher than in other years. We can find that text mining research in Japan was active in a relatively long period.

Initially we assumed that there were a lot of studies that attempt to automatically acquire meaningful knowledge from huge texts on the Internet, which is regarded as a typical big data, and the domains of the text to be mined were economy, management, social research and so on. By contrast, the text mining was mainly applied for medical and life science database, especially in English papers. Nevertheless, there are huge numbers of English text mining papers. Google Scholar shows that 2,390,000 English papers are hit by the query “text mining”, while 3,010 Japanese papers. When more English papers (more than 100) are investigated, a wide variety of the topics and media might be found.

In future, we will examine more Japanese and English papers for the survey of the topic and medium distribution, since 100 papers might be insufficient. We will also investigate the contents of the papers and explore the fundamental techniques used for the text mining. Another future work is to conduct a survey of the text mining tools and example cases of the usage of them.