Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 24 (2013) 274 – 279

Procedia Computer Science

17th Asia Pacific Symposium on Intelligent and Evolutionary Systems, IES2013

Learning to Predict the Need of Summarization on News Articles

Ji Eun Lee, Hyun Soo Park, Kyung Joong Kim* , Jae Chun No

Dept. of Computer Engineering, Sejong Univ., Seoul, Republic of Korea

Abstract

Recently, we live with a huge amount of data. For example, we have great amount of news articles everyday. But there are small amount of useful information in the articles and it is hard to extract useful information manually. As a result, there are lots of news articles but, it is hard to read all of articles and find informative news manually. One of solutions on this problem is to summarize texts in the article. There are many studies on the text summarization techniques, but small number of studies to predict whether the article should be summarized or not. If we don't know about that, it is likely to waste computing resources to summarize unnecessary articles. In this paper, we propose a method to model the pattern of user's summarization needs on news articles. We perform experiments using news articles and apply data mining techniques (C4.5 and Naïve Bayes) to model common preprocessor to execute the automatic summarization. Finally, we can get some meaningful results on the "desire to summarize" prediction.

Keywords: article summarization; natural language processing; data mining; decision tree, naïve bayes

1. Introduction

Recently, many people and devices generate massive data and its size is continuously growing exponentially. One representative case is news articles. There are a lot of media companies and they produce news competitively everyday. People face with an overflow of news articles in a day. Therefore, people often fail to acquire useful

* Corresponding author.
E-mail address: apple173@naver.com (Ji Eun Lee), hspark@sju.ac.kr (Hyunsoo Park), kimkj@sejong.ac.kr (Kyung Joong Kim), jano@sejong.ac.kr(Jae Chun No).

doi:10.1016/j.procs.2013.10.051

information easily and even it is hard to distinguish useful news from unnecessary articles. So, it has been a significant problem to acquire information you want from massive news articles.

Automatic summarization is helpful to solve this problem. A summary of something is a short account of it, which gives the main points but not the details. Studies about summarization have been a popular topic to enable user access information effectively by discarding irrelevant parts. While there are many studies on the summarization methods, but we hardly see specific studies on the prediction of summarization needs [1]. If we can estimate what articles need to be summarized, there is a big advantage to save computing resources on large data. Many text summarization techniques are relatively complex and require a computational/data intensive job. Because of this, if we want to apply the text summarization on huge amount of data, probably we need a lot of computational resources. But if we can predict what article should be summarized in a selective manner and vice versa, we can expect saving of unnecessary computational resource.

Intuitively, we can guess that if an article is too long, people want to get summary. But we do not know the appropriate length of articles to be summarized and what is the good metric in order to measure the article length. In order to investigate on the issue, we collect data from users. In the experiment, we collect readers' desire (this article should be summarized/this article should not be summarized) and learn prediction models using data mining tool WEKA [2]. As a result, we can generate a prediction model which predicts with precision almost 90%.

2. Related works

Table 1. Summary of related works

Author	Year	Description
R. Barzilay, M. Elhada [3]	1997	Topic, representation (Lexical Chain)
C. Lin, E. Hovy [4]	2000	Topic, representation (Topic Signature)
O. Buyukkokten, H. Garcia-Molina, A. Paepcke [5]	2001	Summary for mobile device
J. M. Conroy [6]	2001	Selection (Hidden Markov Models)
Y. Gong, X. Liu [7]	2001	Topic, ranking method (Latent Semantic Analysis)
G. Erkan, D. R. Radev [8]	2001	Representation (Stochastic Graph)
L. Antiqueira, O. N. Oliveira Jr., *et al* [9]	2009	Representation (Complex Network)
M. A. Fattah, F. Ren [10]	2009	Selection (various Machine Learning techniques)
R. M. Aliguliyev [11]	2009	Sentence similarity measure

Text summarization is the one of the most important topics in natural language processing. Sometimes, we can regard the text summarization system as a text-to-text system [1]. This system outputs text shorter than imputed text. This system consists of three parts (1) transform inputted text to intermediate representation, (2) score each sentence and (3) select importance sentences. We can analyze many works in this frame. In most of cases, each work proposes new representation, new scoring method and/or new selection methods.

In the early days, many researchers find a topic in given text and score by its importance [3, 4, 7]. But, nowdays, many works use indicator instead topic [8-11]. In this approach, they compare the importance of each sentence directly, instead of searching for the topic or interpreting the sentences.

Table 1 is the summary of related works. This table summarized authors, published year and their main contribution.

3. The proposed method

For the experiment, we get experimental data from graduated students using news articles. There are four steps in our approach (Fig. 1). (1) We collect URL of an article selected by users for the summarization. If a student thinks this article should be summarized then presses the 'necessary' button on the screen while browsing web sites. (2) We