CS 533 - Information Retrieval Systems
Spring 2013

PROJECT PLAN REPORT

Topic Extraction from Turkish News Articles

Group Members

Anıl ARMAĞAN
Fuat BASIK
Fatih ÇALIŞIR
Arif USTA

21.03.2013

1. Introduction

Increasing volume of the online data has been important in the daily lives of humans. People have access to a huge variety of data via web, which increases knowledge can be gathered from this data while hardened to reaching specific data. For example, reading newspapers from the web is possible while one can reach online versions of the almost all newspapers. However, reading all of them will take hours and if you want to have a quick look on the news, it is hard to see a summary of them. To provide this, we want to come up with an algorithm and its' implementation, which summarizes what happened yesterday, by looking news provided to the system.

2. Description of the problem

Summarizing what happened yesterday requires text analyzing and text mining. In the literature, this issue is addressed as topic extraction or tag extraction. Goal is, to define what topic is related with the given text document or which tags can be assigned to the document. To be able to solve this problem, developer has to score the terms, and find most important terms inside the documents. Importance of a term is not only related with the frequency of it inside the document but also appearances of it in the inverse documents should be considered. In the methodology section, different term scoring algorithms will be explained.

As it will be explained in the motivation part, this project will be focused on the news articles in Turkish languages. This issue should be introduced here, since structure of the Turkish language requires extra processing, for example stemming. Stop word removal, repeating letter removal, correcting typos are some other parts of the problems should be solved to create algorithm will be designed.

3. Motivation and Importance

Many people who uses web are trying to cope with huge amount of new data every day. People do not have enough time to keep following the news. Therefore, there is a need to mine the news data for usage of people. Those people only like to be able to find the important news as fast as possible.

Agglutinating structure of Turkish language, made mining the Turkish texts harder. This language requires a lot of preprocessing steps, which might be costly. There are many other works using text mining for English but we are planning to mine an agglutinating language. In this way, we hope we can contribute to text mining research in Turkish language

Another motivation for this project is, as long as we continue to use text based search engines, text mining and text processing research will stay a hot topic. Therefore, learning text processing, research about this area we believe will be beneficial for us and might open new research areas.

4. Methodology

Before mining the texts, basically news articles, it is better to have some preprocessing, since without preprocessing, number of the terms resulted from the news articles, gets higher and meaning of the terms might be meaningless or ambiguous, which makes latter text mining process useless. Hence, as preprocessing steps, stop word removing (such as “ve”,“ya da”, etc.) and stemming following that are to be used in this project. In addition to those techniques, we also plan to search over other preprocessing techniques such as tokenizing or generating n-grams.

After making the collection to be ready for processing, what is important is to assign weight to each term that is not eliminated from pre-processing steps. In order to do that, there are some algorithms such as TF-IDF or TF-PDF. TF-IDF is a well known method to assign weights to each term. It is shortly taking term frequency in a particular document and term rareness in the collection into account in an equal manner. TF-PDF is same with TF-IDF considering both having term frequency in a particular document as factor in the formulas. Yet, PDF is the part that makes approach different. Rather than assigning high weight values to each term that is not appearing much in the collection, PDF (Proportional Document Frequency) assigns more weight to terms that are coming from multiple sources and appearing in most of them as well.

After assigning weights to each term, what is left is to try the find most important news in a particular day by looking the terms in each news. The news having terms with higher weights will be the most important ones to be serviced.

For the project, the dataset to be used is the Turkish news articles for a period of 28 days gathered by Çağrı Toraman.

5. Expected results

We expect to be able to develop an algorithm and implement it, given the daily news which returns the user, a summary of the events addressed in the provided list, sorted with respect to the importance of them.

Another result we hope to find, by while creating the algorithm we guess that, we will use different term scoring algorithms. We hope to find a comparison between these algorithms and we might be able to find which feature scoring algorithm is better than the others in terms of efficiency and effectiveness.

6. References

[1] K.B. Khoo and M. Ishizuka: “Emerging Topic Tracking System” In: Proc. of Web Intelligent (WI 2001), LNAI 2198 (Springer), pp. 125-130, Maebashi, Japan. 2001.

[2] Bun, Khoo K., and Mitsuru Ishizuka. "Topic Extraction from News Archive Using TF*PDF Algorithm."

[3] J. Allan, R. Papka, and V. Lavrenko: “Online New Event Detection and Tracking” In: Proc. of SIGIR ’98: 21st Ann. Int’l ACM SIGIR Conf. Research and Development in Information Retrieval, ACM Press, New York, 1998, pp. 37-45

[4] G. Salton and C. Buckley : “Term-Weighting Approached in Automatic Text Retrieval” In: Information Processing and Management, Vol. 4, No. 5, pp. 513-523 1989.