Chapter 4
Methodologies and Technologies to Retrieve Information From Text Sources

Anu Singha
South Asian University, India

Phub Namgay
Sherubtse College, Royal University of Bhutan, Bhutan

ABSTRACT

A tool which algorithmically traces the effectiveness of the text files would be helpful in determining whether the text file have all the characteristic of important concepts. Every text source is build up on key phrases, and these paramount phrases follow a certain grammatical linguistic pattern widely used. An enormous amount of information can be derived from these key concepts for the further analysis such as their dispersion, relationship among the concepts etc. The relationship among the key concepts can be used to draw a concept graphs. So, this chapter presents a detailed methodologies and technologies which evaluate the effectiveness of the extracted information from text files.

INTRODUCTION

Before the advent of internet, text files formed the only source to get the content knowledge. Text files are read as a part of academics for a university student or for leisure, who fancy reading. With the rise of modern technology, paper based text files is being replaced by electronic version and is easily available online. Though

DOI: 10.4018/978-1-5225-2805-0.ch004

the readers embrace comfort and portability provided by electronic version of the text files, paper based text files are still popular. Unlike decades back, text files in any field of study is readily available now. With the demand for text files rising, the qualities of the text files are compromised. With the rate reading habits on decline, impatient readers hardly spent some time to evaluate the content of the file. A technique to analyze the effectiveness by evaluating the text files in detail will be of great importance. This ultimately would help the readers in finding the text file most suitable to his / her needs and comprehension capability.

A tool which algorithmically traces the effectiveness the text files would be helpful in determining whether the text file have all the characteristic of a good source. Every file is build up on key concepts, and these key concepts form the foundation of a good source. The text sources that contain concepts that share some common properties and semantically related are more lucid and intelligible than those text sources which contain many unrelated concepts. These paramount phrases follow a certain grammatical linguistic pattern widely used. An enormous amount of information can be derived from these key concepts for the further analysis such as their dispersion across the file, relationship among the concepts, etc. Such analysis will help in better assessment of the text file. The relationship among the key concepts can be used to draw a concept graphs. Since we live in an increasingly visual society, pictorial representation of the key concepts as a graph would help the readers in easily judging the text source and their content.

The goal of this chapter is to confer the methodologies on the key concepts retrieval from the text files. The authors investigate the techniques for examining the key concepts in the text files. This chapter also presents some of the different tools used in natural language processing. Their uses and implementation methods are explicitly discussed. The key concepts in our context correspond to the terminological noun phrases. These extracted set of phrases can be further analyzed to check the credibility of the text file in conveying the required set of information to the readers. It is based on the intuition that a source which contains right set of related key concepts is more beneficial and comprehensible. The set of key concept which form the cornerstone of a text file can be further used to draw concept graphs. The noun phrases from the candidate set of extracted phrases form the nodes of the graph and the relationship that exists between the nodes can be denoted by a link between the concept pairs. The ‘in’ degree and ‘out’ degree of each vertex of the graph i.e., the noun phrases can be used to determine the most important key concepts. Such representation of the source in a visual form helps readers in easily judging and grasping the key concepts. This ultimately serves as a preface to the text files and reduces the cognitive burden of the readers.