Automatic Extraction and Learning of Keyphrases from Scientific Articles

Yaakov HaCohen-Kerner, Zuriel Gross, Asaf Masa
Department of Computer Sciences, Jerusalem College of Technology (Machon Lev)
21 Havaad Haleumi St., P.O.B. 16031, 91160 Jerusalem, Israel
{kerner, zuriel, masa}@jct.ac.il

Abstract. Many academic journals and conferences require that each article include a list of keyphrases. These keyphrases should provide general information about the contents and the topics of the article. Keyphrases may save precious time for tasks such as filtering, summarization, and categorization. In this paper, we investigate automatic extraction and learning of keyphrases from scientific articles written in English. Firstly, we introduce various baseline extraction methods. Some of them, formalized by us, are very successful for academic papers. Then, we integrate these methods using different machine learning methods. The best results have been achieved by J48, an improved variant of C4.5. These results are significantly better than those achieved by previous extraction systems, regarded as the state of the art.

1 Introduction

Summarization is a process reducing an information object to a smaller size, and to its most important points [1, 18]. Various kinds of summaries (e.g.: headlines, abstracts, keyphrases, outlines, previews, reviews, biographies and bulletins) can be read with limited effort in a shorter reading time. Therefore, people prefer to read summaries rather than the entire text, before they decide whether they are going to read the whole text or not. Keyphrases, which can be regarded as very short summaries, may help even more. For instance, keyphrases can serve as an initial filter when retrieving documents. Unfortunately, most documents do not include keyphrases.

Moreover, many academic journals and conferences require that each paper will include a list of keyphrases. Therefore, there is a real need for automatic keyphrase extraction at least for academic papers. There are a few such systems. However, their performances are rather low. In this paper, we present a system that gives results significantly better than those achieved by the previous systems.

This paper is organized as follows: Section 2 gives background concerning extraction of keyphrases. Section 3 describes a few general kinds of machine learning. Section 4 presents our baseline extraction methods. Section 5 describes our model. Section 6 presents the results of our experiments and analyzes them. Section 7 discusses the results, concludes and proposes future directions.

2 Extraction of Keyphrases

A keyphrase is an important concept, presented either in a single word (unigram), e.g.: ‘learning’, or a collocation, i.e., a meaningful group of two or more words, e.g.: