Chapter VIII

KEA:
Practical Automatic
Keyphrase Extraction

Ian H. Witten
University of Waikato, New Zealand

Gordon W. Paynter
University of Waikato, New Zealand

Eibe Frank
University of Waikato, New Zealand

Carl Gutwin
University of Saskatchewan, Canada

Craig G. Nevill-Manning
Google, Inc., USA

Abstract

Keyphrases provide semantic metadata that summarize and characterize documents. This chapter describes Kea, an algorithm for automatically extracting keyphrases from text. Kea identifies candidate keyphrases using lexical methods, calculates feature values for each candidate, and uses a machine-learning algorithm to predict which candidates are good keyphrases. The machine-learning scheme first builds a prediction model using training documents with known keyphrases, and then uses the model to find keyphrases in new documents. We use a large test corpus to evaluate
Kea’s effectiveness in terms of how many author-assigned keyphrases are correctly identified. The system is simple, robust, and available under the GNU General Public License; the chapter gives instructions for use.

Introduction

Keyphrases provide a brief summary of a document’s contents. As large document collections such as digital libraries become widespread, the value of such summary information increases. Keywords and keyphrases are particularly useful because they can be interpreted individually and independently of each other. They can be used in information retrieval systems as descriptions of the documents returned by a query, as the basis for search indexes, as a way of browsing a collection, and as a document clustering technique.

In addition, keyphrases can help users get a feel for the content of a collection, provide sensible entry points into it, show how queries can be extended, facilitate document skimming by visually emphasizing important phrases, and offer a powerful means of measuring document similarity (e.g., Gutwin et al., 1999; Witten, 1999).

Keyphrases are usually chosen manually. In many academic contexts, authors assign keyphrases to documents they have written. Professional indexers often choose phrases from a predefined “controlled vocabulary” relevant to the domain at hand. However, the great majority of documents come without keyphrases, and assigning them manually is a tedious process that requires knowledge of the subject matter. Automatic extraction techniques are potentially of great benefit.

There are two fundamentally different approaches to the problem of automatically generating keyphrases for a document: keyphrase assignment and keyphrase extraction. Both use machine learning methods, and require for training purposes a set of documents with keyphrases already attached.

Keyphrase assignment seeks to select the phrases from a controlled vocabulary that best describe a document. The training data associates a set of documents with each phrase in the vocabulary, and builds a classifier for each phrase. A new document is processed by each classifier, and assigned the keyphrase of any model that classifies it positively (e.g., Dumais et al., 1998). The only keyphrases that can be assigned are ones that have already been seen in the training data.

Keyphrase extraction, the approach used here, does not use a controlled vocabulary, but instead chooses keyphrases from the text itself. It employs lexical and information retrieval techniques to extract phrases from the docu-