Chapter II
Deriving Document Keyphrases for Text Mining

Yi-fang Brook Wu
New Jersey Institute of Technology, USA

Quanzhi Li
Avaya, Inc., USA

ABSTRACT

Document keyphrases provide semantic metadata which can characterize documents and produce an overview of the content of a document. This chapter describes a Keyphrase Identification Program (KIP), which extracts document keyphrases by using prior positive samples of human identified domain keyphrases to assign weights to the candidate keyphrases. The logic of our algorithm is: the more keywords a candidate keyphrase contains and the more significant these keywords are, the more likely this candidate phrase is a keyphrase. To obtain human identified positive inputs, KIP first populates its glossary database using manually identified keyphrases and keywords. It then checks the composition of all noun phrases extracted from a document, looks up the database and calculates scores for all these noun phrases. The ones having higher scores will be extracted as keyphrases. KIP’s learning function can enrich the glossary database by automatically adding new identified keyphrases to the database.

INTRODUCTION

As textual information pervades the Web and local information systems, text mining is becoming more and more important to deriving competitive advantages. One critical factor to successful text mining applications is the ability of finding significant topical terms for discovering interesting patterns or relationships. Discourse representation theory (Kamp, 1981) shows that a document’s primary concepts are mainly carried by noun phrases. We believe that noun phrases (NPs) as textual elements are better
suited for text mining and could provide more discriminating power than single words, which are still used by many text mining applications. Since not all NPs in a document are important, we propose using them as candidates and identifying keyphrases from them. Document keyphrases are the most important topical phrases for a given document. They provide a concise summary of a document’s content, offering semantic metadata summarizing a document.

Previous studies have shown that document keyphrases can be used in a variety of applications, such as retrieval engine (Li, Wu, Bot & Chen, 2004), browsing interface (Gutwin, Paynter, Witten, Nevill-Manning & Frank, 1999), thesaurus construction (Kosovac, Vanier & Froese, 2000), and document classification and clustering (Jonse & Mahoui, 2000). For example, they may be utilized to enrich the metadata of the results returned from a search engine (Li et al, 2004). They may also be used to efficiently classify or cluster documents into different categories (Jonse & Mahoui, 2000). Some documents have a list of author-assigned keyphrases, but most documents do not have them. The keyphrases assigned by domain experts, such as indexers or authors, may be chosen from a document or a controlled vocabulary. However, manually assigning keyphrases to documents is costly and time-consuming, so it is necessary to develop an algorithm to automatically generate keyphrases for documents. Automatic keyphrase generation can be executed in two ways: keyphrase extraction and keyphrase assignment. Keyphrase extraction methods choose phrases from the document text as keyphrases. Keyphrase assignment methods choose phrases from a controlled vocabulary as document keyphrases. The problem with keyphrase assignment is that some potentially useful keyphrases will be ignored if they appear in the document but not in the controlled vocabulary. In some domains, such controlled vocabularies might not even be available, and most controlled vocabularies are not updated frequently enough. Therefore, automatic keyphrase extraction method is more desirable and popular.

In this chapter, we describe our Keyphrase Identification Program, KIP, and its learning function. KIP utilize a keyphrase extraction technique: the keyphrases generated by KIP must appear in the document. The algorithm considers the composition of a noun phrase. To analyze a noun phrase and assign a score to it, KIP uses a glossary database, which contains manually pre-identified domain-specific keyphrases and keywords, to calculate scores of noun phrases in a document. The noun phrases having higher scores will be extracted as keyphrases. KIP’s learning function enriches its glossary database by automatically adding new keyphrases extracted from documents to the database. Consequently, the database will grow gradually and the system performance will be improved.

The remainder of this chapter is organized in the manner described below. Previous studies are first presented. Then our methodology and system architecture are described, followed by KIP’s learning function. Finally, the evaluations of KIP’s effectiveness and its learning function are presented.

RELATED WORK

Several automatic keyphrase extraction techniques have been proposed in previous studies. Krulwich and Burkey (1996) use some heuristics to extract significant topical phrases from a document. The heuristics are based on documents’ structural features, such as the presence of phrases in document section headers, the use of italics, and the different formatting structures. This approach is not difficult to implement, but the limitation is that not every document has explicit structural features.

Zha (2000) proposes a method for keyphrase extraction by modeling documents as weighted undirected and weighted bipartite graphs. Spectral graph clustering algorithms are used for partitioning