Chapter 6
Keyword Extraction

ABSTRACT

Keywords are defined as phrases that capture the main topics discussed in a document. As they offer a brief yet precise summary of document content, they can be utilized for various applications. In an IR (information retrieval) environment, they serve as an indication of document relevance for users, as the list of keywords can quickly help to determine whether a given document is relevant to their interest. As keywords reflect a document's main topics, they can be utilized to classify documents into groups by measuring the overlap between the keywords assigned to them. Keywords are also used proactively in information retrieval (i.e., in indexing).

TERMINOLOGY AND NOTATIONS USED

The general terminology used in this chapter is brie y discussed in Table 1.

RELEVANCE OF KEYWORDS IN PLAIN TEXT

As Keywords reflect a document's main topics, they can be utilized to classify documents into groups by measuring the overlap between the Keywords assigned to them. Keywords are also used proactively in information retrieval i.e., in indexing. Good Keywords mostly supplement full-text indexing by assisting users in finding relevant documents (Christos, B. 2006).

DOI: 10.4018/978-1-7998-3772-5.ch006

Table 1. Terminology used for keyword extraction

Notation	Term	Meaning
D	Document	A text document consisting of a set of words
W	Word	A sequence of non-blank characters
WL	Word List	A list of meaningful words
SW	Stop Words	A collection of stop words
S	Stemmed Word	The stem of the word
FC	Frequency count of a word	The number of times the word is found in the document
T	Frequency threshold	User Input criteria to find dense words
N	Document size	Total number of words found in the document
M	Extracted words size	Total number of words from document after removing stop words from the N words.
MS	Min Support	T * (N/ M)
DW	Dense Word	A word whose frequency count (FC) in the document is greater than or equal to the Min Support (MS)
CWx	Candidate Word phrase of length x in document, D	Sequence of x Dense Words which could be the Frequent Word phrase for the document, D
FWx	Frequent Word phrase of length x in document, D	Candidate Word phrase of length x whose FC is greater or equal to Min Support and can be considered as a keyword phrase of length x for Document, D
KW [i][j]	Keyword Set	A table (2 dim array) of frequent word phrases FWi,.i.e. ith row consists of all frequent word phrases of length i. KW[i] [j] represents the jth FWi, frequent word phrase of length i

Keywords are Meant to Serve Multiple Goals

1. When they are printed on the first page of a journal article, the goal is summarization (Christos, B. and Vassilis, T.2008). They enable the reader to quickly determine whether the given article is in the reader's fields of interest.
2. When they are printed in the cumulative index for a journal, the goal is indexing. They enable the reader to quickly find a relevant article when the reader has a specific need.
3. When a search engine form has a field labeled *keyword*s, (Fan, W., Wallace, L., Rich, S. & Zhang, Z. 2006). the goal is to enable the reader to make the search more precise. A search for documents that match a given query term in the *keyword* field will yield a smaller, higher quality list of hits than a search for the same term in the full text of the document.