Performance Analysis of Semantic Indexing in Text Retrieval

Bo-Yeong Kang, Hae-Jung Kim and Sang-Jo Lee

Department of Computer Engineering, Kyungpook National University,
Sangyuk-dong, Puk-gu, Daegu, 702-701, KOREA
comeng99@hotmail.com

Abstract. We developed a new indexing formalism that considers not only the terms in a document, but also the concepts to represent the semantic content of a document. In this approach, concept clusters are defined and a concept vector space model is proposed to represent the semantic importance of words and concepts within a document. Through experiments on the TREC-2 collection, we show that the proposed method outperforms an indexing method based on term frequency.

1 Introduction

To intelligently retrieve information, many indexing methods such as term frequency (TF), inverse document frequency (IDF), the product of TF and IDF have been proposed and tested [1]. Most of TF-based methods have difficulties in extracting semantically exact indexes that express the topics of a document. Consider the sample text below, the important terms that could be topics of the text are *anesthetic* and *machine*(*device*). However, the TF weight of the word *machine* is 1, which is the same as that of semantically unimportant words such as *rate* and *blood*. Thus, the TF approach fails to discriminate the degree of semantic importance of each word within the text.

"Dr. Kenny has invented an anesthetic machine. This device controls the rate at which an anesthetic is pumped into the blood."

Linguistic phenomenon such as *lexical chain*[2], which links related words in a text, have been used to enhance the indexing performance[3]. In the sample text, we obtain two representative chains, *anesthetic*−*anesthetic* and *machine*−*device*, which correctly indicates that the focus words of the text are *anesthetic* and *machine*/*device*. In the present study, we propose a new semantic approach based on lexical chains for extracting words from a text and assigning them importance degrees, and analyze the performance of the proposed semantic indexing.

2 Semantic Indexing

Documents generally contain various concepts, and we must determine those concepts if we are to comprehend the aboutness of a document. In accordance