Abstract

It is common that text documents are characterised and classified by keywords and that the authors use to give and name these text characteristics. Visa et al. have, however, developed a new methodology based on prototype matching. The prototype is an interesting document or a part of an extracted, interesting text. This prototype is matched with the existing document database or the monitored document flow. Our claim is that the new methodology is capable of extracting meaning automatically from the contents of the document. To verify this hypothesis a test was designed with the Bible. Two different translations, one in English and another in Finnish, were selected as test text material. Verification tests that included the search of the ten nearest books to every book of the Bible were performed with a designed prototype version of the software application. The interesting test results are reported in this paper.

The new methodology is based on a hierarchy of Self-Organizing Maps (SOM) and on a smart encoding of words. The words of a text document are encoded. The encoded words are represented as word vectors. The word vectors are clustered by the SOM and this process creates a word map. The words of a text document are replaced with the addresses on the word map. Now the document consists of a sequence of addresses. The document is considered sentence by sentence. These sentence vectors are clustered by SOM. This process creates a sentence map. Now the sentences of the text document are replaced with addresses on the sentence map. After that the document consists of a sequence of addresses. These addresses contain information of different types of sentences. The document is then considered paragraph by paragraph. The paragraphs are considered as context vectors and clustered by SOM. The created map is called a context map. The paragraphs are replaced with the addresses on the context map. The document consists finally of a sequence of addresses on the context map. The more detailed description of the methodology can be found in several proceedings [10, 7, 11, 9, 8].

The test hypothesis was that the words, the word order in the sentences and the order of sentences in paragraphs can grasp higher level of information than ordinary word based searches. Two tests were designed. It was important to find a text that is well translated at least into two languages. The Bible was selected. Each book of 66 books in the Bible was selected as a prototype both in English and in Finnish versions. A window of ten closest books was considered. The window size ten was selected to guarantee a statistical significance. In the first test the number of books in the window that matched with other books in the Old Testament, respectively in the New Testament, was counted for each book. In the second test the same books within the window in English and in Finnish versions were considered. The results from these tests are statistically significant. The methodology is capable of understanding the contents of the document at least on a certain level.