Association Rule Generation and Mining Approach to Concept Space for Collective Documents

Ken Nittono*
Hosei University, Tokyo, Japan

In recent years, the amount of exchanged and accumulated online documents has been growing with the diversification of computer network services and terminal device technology. And many kinds of publications which have been published in paper-based style such as books, proceedings and transcripts become to have more opportunities to be addressed in digitized text data. Based on those backgrounds, attempts to analyze such large amounts of data and to obtain useful knowledge effectively have been attracted attention, and various methods are proposed in many areas. In these circumstances, this study aims to mine collective documents for significant terms or contexts and extract particular information. Finding the desired information from the documents in such cases as books or journals that are systematically edited by the author or editor is relatively straightforward, however, it is rather difficult to find it with later reading for the whole data which is recorded in accordance with passage of time such as communication log in some network service or transcript of interview because they often contain redundant expression, incomplete sentence or irrelevant context. The latter case of text data is dealt with in this study and a method for extracting essential part from large text data is proposed. At first, the documents are basically formulated as a term-document matrix. And some characterizing terms or phrases are mined throughout the documents by the application method based on association rules. For the generation of association rules, the meanings for traditional criteria of support, confidence and lift should be given, and here, they are considered as rules to obtain common, precious and rare information, respectively. The amount of information to be extracted from original documents as a result depends on adjustment of several parameter values for the generation of the rules. Second, essential contexts are extracted from original documents by the use of latent semantic analysis (LSA). The whole documents are translated in a concept space by singular value decomposition (SVD) and the space enhances latent meanings contained by the documents. Mining this space with terms or phrases obtained by the above method enables to extract essential contexts or paragraphs. For example, latent semantic indexing (LSI) is a method to evaluate distance between the concept space and the terms. Note that some approaches such as extracting all of sentences that contains some particular terms throughout documents are not suitable for this case.

Key Words: Text mining, collective intelligence, latent knowledge, big data