■ 연구논문 요약문

논문제목	Bag−of−concepts: Comprehending document representation through clustering words in distributed representation
게재정보	Neurocomputing 266 (2017) 336-352
개요	Two document representation methods are mainly used in solving text mining problems. Known for its intuitive and simple interpretability, the bag−of−words method represents a document vector by its word frequencies. However, this method suffers from the curse of dimensionality, and fails to preserve accurate proximity information when the number of unique words increases. Furthermore, this method assumes every word to be independent, disregarding the impact of semantically similar words on preserving document proximity. On the other hand, doc2vec, a basic neural network model, creates low dimensional vectors that successfully preserve the proximity information. However, it loses the interpretability as meanings behind each feature are indescribable. This paper proposes the bag−of−concepts method as an alternative document representation method that overcomes the weaknesses of these two methods. This proposed method creates concepts through clustering word vectors generated from word2vec, and uses the frequencies of these concept clusters to represent document vectors. Through these data−driven concepts, the proposed method incorporates the impact of semantically similar words on preserving document proximity effectively. With appropriate weighting scheme such as concept frequency−inverse document frequency, the proposed method provides better document representation than previously suggested methods, and also offers intuitive interpretability behind the generated document vectors. Based on the proposed method, subsequently constructed text mining models, such as decision tree, can also provide interpretable and intuitive reasons on why certain collections of documents are different from others.
연구결과	With appropriate weighting scheme such as concept frequency−inverse document frequency, the proposed method provides better document representation than previously suggested methods, and also offers intuitive interpretability behind the generated document vectors.
활용분야 및 기대효과	Based on the proposed method, subsequently constructed text mining models, such as decision tree, can also provide interpretable and intuitive reasons on why certain collections of documents are different from others.