

Bag-of-concepts: Comprehending document representation through clustering words in distributed representation



Han Kyul Kim, Hyunjoong Kim, Sungzoon Cho*

Department of Industrial Engineering, Seoul National University 1 Gwanak-ro, Gwanak-gu, Seoul 151-744, Republic of Korea

ARTICLE INFO

Article history:

Received 26 May 2016

Revised 28 March 2017

Accepted 22 May 2017

Available online 26 May 2017

Communicated by Y. Chang

Keywords:

Bag-of-concepts

Interpretable document representation

Word2vec clustering

ABSTRACT

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.

© 2017 Elsevier B.V. All rights reserved.

1. Introduction

With growing importance of unstructured data, text data, as one of the most common forms of unstructured data, have inevitably become an important source for data analysis. To extract interesting patterns and insights from them, most fundamental and crucial step is document representation. In order to apply various machine learning and data mining techniques, raw documents need to be transformed into numerical vectors, in which each document's defining characteristics are captured. If these document vectors can preserve proper proximity between the documents and their unique characteristics, subsequent text mining algorithms can extract more accurate and valuable information hidden in the data.

Most popular document representation methods have often relied on the bag-of-words based approaches [1,18,28], in which a document is fundamentally represented by the counts of word

occurrences within a document. For decades, this approach has been shown to be effective in solving various text mining tasks [12,15,33]. One of its major advantages is that it produces intuitively interpretable document vectors as each feature of the document vector indicates an occurrence of a specific word within a document. The bag-of-words approach, however, can be problematic when a number of documents being represented are enormous. As a number of documents increase, a number of unique words in the entire document set will also naturally increase. Consequently, not only will the generated document vectors be sparse, but their dimensions will also be huge. As the dimension and the sparsity of the document vectors increase, conventional distance metrics become ineffective in representing the proper proximity between the documents. Furthermore, the bag-of-words methods assumes that all words within the documents are independent. However, different word types such as synonyms and hypernyms usually describe similar information within the document. Thus, this word independence has adverse impact on capturing document proximity. Consequently, the text mining models constructed from the bag-of-words approach can be unsuccessful in capturing

* Corresponding author.

E-mail addresses: hank@dm.snu.ac.kr (H.K. Kim), hyunjoong@dm.snu.ac.kr (H. Kim), zoon@snu.ac.kr (S. Cho).

[Document 1]:

Arsenal legend Robert Pires has labelled another Gunners icon, Dennis Bergkamp, a maestro after naming the former Holland star in a best XI of his former teammates for The Fantasy Football Club.


The attack-minded duo lined up alongside each other for Arsenal for six years, after Pires made the move to north London from Marseille in 2000.

Arsenal enjoyed great success during that time, lifting two Premier League titles and three FA Cups.

[Document 2]:

Robert Pires has selected a dream team for Sky Sports and it features seven Frenchmen, six former Arsenal superstars, a handful of La Liga players and one of the greatest players of all time.

Decorated winger Robert Pires joined Arsenal in 2000 after winning the World Cup and the European Championship with France. It is no surprise that his ultimate XI has been filled with an abundance of Les Bleus internationals and former Gunners.



	X[1]: Arsenal	X[2]: Legend	X[3]: Robert	X[4]: Pires	...
Document 1	3	1	1	2	...
Document 2	2	0	2	2	...

Fig. 1. Document vectors generated via bag-of-words approach.

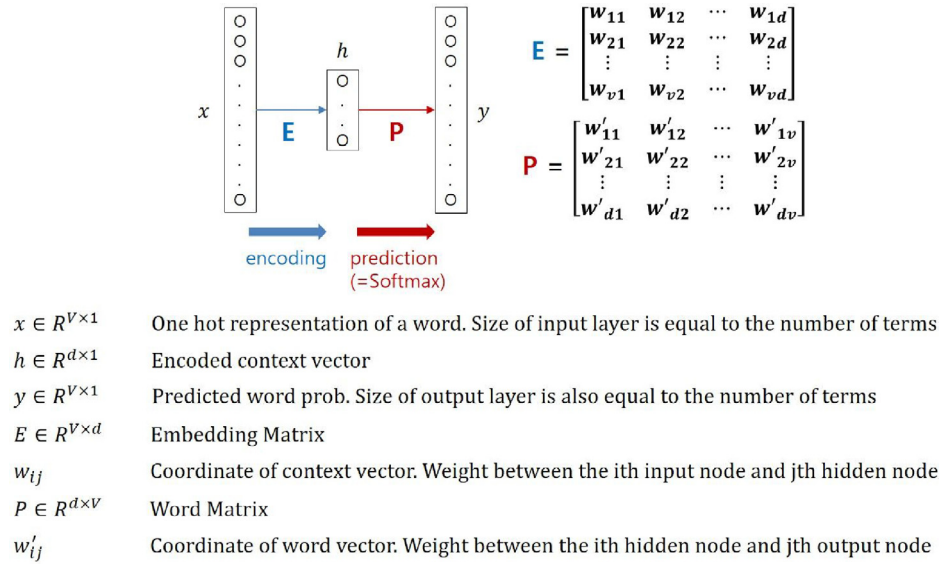


Fig. 2. Basic architecture of word2vec (Skip-gram).

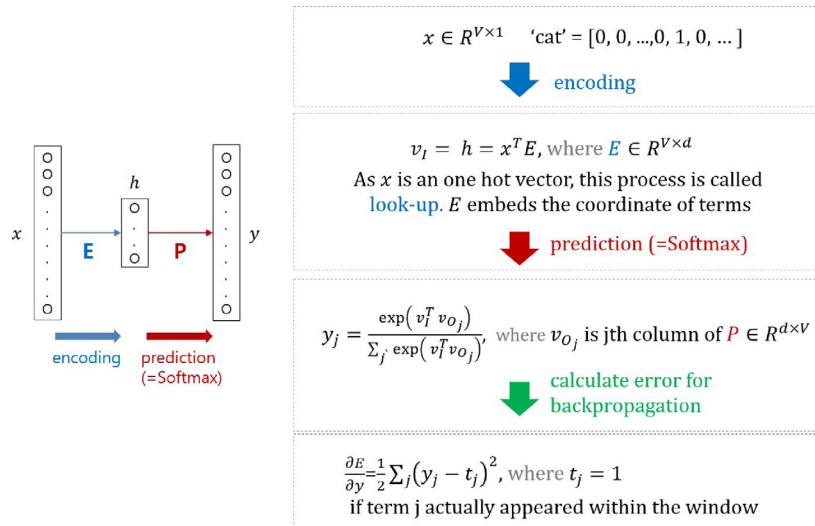


Fig. 3. Word2vec (Skip-gram) training.

Download English Version:

<https://daneshyari.com/en/article/4946948>

Download Persian Version:

<https://daneshyari.com/article/4946948>

[Daneshyari.com](https://daneshyari.com)