Abstract

This paper proposes a novel and efficient approach to calculate bigram frequency which uses E-VSM as basis to represent text document. E-VSM: Enhanced-Vector Space Model is nothing but an extension to simple VSM which stores positions of tokens in addition to their frequency in document. Many recent methodologies in Information Retrieval and Text Mining have used bigram along with unigram since bigram gives more information gain than unigrams. Also recent efforts to provide more richer text document representation than simple 'Bag of Words' have also used bigram along with unigram. Proposed approach to calculate bigram frequency outperforms state-of-art in terms of time complexity. Analysis show that proposed approach improves time complexity to significant extent.

References

- Matthew A. Russel, "Mining the Social Web", O'Reilly (2011), chapter 7, pp 224-229
Efficient Approach to find Bigram Frequency in Text Document using E-VSM

- Yashodhara Haribhakta, Arti Malgaokar and Dr. Parag Kulkarni, "Unsupervised Topic Detection Model and Its Application in Text Categorization"; 2012 ACM 978-1-4503-1185-4/12/09

Index Terms

Computer Science  
Artificial Intelligence

Keywords

E-VSM  bigram  trigram  n-gram  frequency count