Information Processing and Management 53 (2017) 653–665

Contents lists available at ScienceDirect

Information Processing and Management

journal homepage: www.elsevier.com/locate/infoproman

Vocabulary size and its effect on topic representation

Kun Lu[a,*], Xin Cai[b], Isola Ajiferuke[c], Dietmar Wolfram[b]

[a] School of Library and Information Studies, University of Oklahoma, 401 West Brooks, Norman, OK 73019, USA
[b] School of Information Studies, University of Wisconsin-Milwaukee, P.O. Box 413, Milwaukee, WI 53201, USA
[c] Faculty of Information and Media Studies, University of Western Ontario, London, ON N6A 5B7, Canada

ARTICLE INFO

Article history:
Received 16 February 2016
Revised 26 September 2016
Accepted 9 January 2017

Keywords:
Information retrieval
Informetrics
Topic modeling
Latent Dirichlet allocation
Vocabulary size
Term frequency

ABSTRACT

This study investigates how computational overhead for topic model training may be reduced by selectively removing terms from the vocabulary of text corpora being modeled. We compare the impact of removing singly occurring terms, the top 0.5%, 1% and 5% most frequently occurring terms and both top 0.5% most frequent and singly occurring terms, along with changes in the number of topics modeled (10, 20, 30, 40, 50, 100) using three datasets. Four outcome measures are compared. The removal of singly occurring terms has little impact on outcomes for all of the measures tested. Document discriminative capacity, as measured by the document space density, is reduced by the removal of frequently occurring terms, but increases with higher numbers of topics. Vocabulary size does not greatly influence entropy, but entropy is affected by the number of topics. Finally, topic similarity, as measured by pairwise topic similarity and Jensen-Shannon divergence, decreases with the removal of frequent terms. The findings have implications for information science research in information retrieval and informetrics that makes use of topic modeling.

Published by Elsevier Ltd.

1. Introduction

Topic modeling (Blei, Ng, & Jordan, 2003; Hoffman, 1999) is a machine learning technique applied to text corpora that was initially developed to reduce computational overhead inherent in high-dimensional spaces found in environments such as information retrieval (IR). Topic modeling adopts computational approaches (e.g. unsupervised learning or matrix factorization) to unveil latent topic structures, an unobservable layer between documents and terms, from a text corpus. The topics are generally represented as a mixture of semantic coherent terms (with different probabilities) that concentrate on some interpretable themes. This technique has been widely used in information retrieval, and more recently has found application in informetrics research (Ding, 2011a; Lu & Wolfram, 2012). By reducing the "dimensionality" of the computational task–that could easily involve millions of index terms in large document corpora–down to at most several hundred latent topics, the computational burden associated with the processing of the textual data may be largely reduced. The trained topics are also believed to be a more parsimonious representation of the semantic space than the traditional bag-of-words representations, and could therefore be more advantageous (Efron, 2005). Forms of topic modeling have been applied in informetric studies over the past decade, where similar issues of computational overhead are becoming more common as a result of larger, full-text datasets that are now available. One form of topic modeling that has been used in informetric studies relies on latent Dirichlet allocation (LDA) (Blei et al., 2003). The LDA model treats a document as a mixture of topics

[*] Corresponding author.
E-mail addresses: kunlu@ou.edu (K. Lu), xincai@uwm.edu (X. Cai), iajiferu@uwo.ca (I. Ajiferuke), dwolfram@uwm.edu (D. Wolfram).

http://dx.doi.org/10.1016/j.ipm.2017.01.003
0306-4573/Published by Elsevier Ltd.

and a topic as a mixture of terms. Rosen-Zvi, Chemudugunta, Griffiths, Smyth, and Steyvers (2010) extended the original LDA model to include authors and proposed the author-topic model, which has direct application in informetrics for author-based comparisons. More recently, Lu and Wolfram (2012) proposed using LDA to compare how similar authors' oeuvres are to each other.

Although topic modeling techniques, such as LDA, can reduce the computational burden of comparing entities once topics have been trained, the training process itself can be time consuming and computationally intensive. The complexity of the LDA algorithm is proportional to the number of documents, the number of topics, and the number of terms in the corpus. The vocabulary size can be huge as increasingly large text corpora are compiled. If the vocabulary size could be reduced during training without significantly affecting the nature of the topic representations or document comparisons, this could also reduce the computational overhead in identifying the topics. The purpose of this study is to systematically investigate the impact of vocabulary size on the outcomes of topic models. The present study explores the following research questions:

1) What impact does the removal of frequently and/or infrequently occurring terms for topic training have on the ability to discriminate documents in a text corpus based on the document space density?
2) How does the removal of frequently and/or infrequently occurring terms affect topic distributions in documents and the distinctiveness of the trained topics using several measures (entropy, pairwise topic similarity and Jensen–Shannon divergence)?

Experiments on three datasets are carried out to examine the impact of different vocabulary selection strategies on the outcomes of topic models. A thorough understanding of the research questions contributes to more efficient use of the topic modeling technique for information science research, including information retrieval and informetric studies where large text corpora are used.

2. Literature review

The present study builds on earlier research that has examined issues of dimensionality reduction in information retrieval environments. The following sections review the literature on the development of topic modeling, applications of topic modeling, and computational overhead in topic modeling.

2.1. Development of topic modeling

The vector space model is a classic mathematical model that describes the relationship that exists between documents and queries as well as between the documents themselves. Documents and queries are represented as vectors in a high-dimensional space whose dimensionality is determined by the number of terms indexed by the system. Document and query relationships are determined by various calculated distance or angle-based measures within the space. As the number of documents indexed by an IR system (n) increases, the computational burden associated with processing documents and queries also increases. With potentially hundreds of millions of terms (m) and many millions of documents, the computational challenges associated with processing an n x m space become apparent.

A method of dimensionality reduction to lower this computational overhead without appreciable loss in the observed relationships that exist among documents and queries would benefit the assessment and retrieval of documents. A pioneering method for document indexing that employed dimensionality reduction was developed by Deerwester, Dumais, Furnas, Landauer, and Harshman (1990) who proposed Latent Semantic Indexing (LSI). LSI represented an early form of topic models that relied on singular value decomposition. A more sophisticated approach that built on LSI, called probabilistic LSI (pLSI), was developed by Hofmann (1999). It used a statistical latent class model based on the likelihood principle and defined a generative data model. The retrieval performance was found to be superior to LSI. More recently, Blei et al. (2003) have proposed latent Dirichlet allocation. Like pLSI, it has the advantage over LSI in explicitly modeling latent topics, and over pLSI in solving the over-fitting problem (i.e. a model with too many parameters). Another line of development in topic modeling involves relaxing the assumptions of the model as the original LDA model assumes that topics are independent of each other. For example, correlated topic modeling has been introduced to allow correlations among topics (Blei & Lafferty, 2007) and hierarchical topic modeling has been developed to model hierarchical relationships among topics (Teh, Jordan, Beal, & Blei, 2012) or to associate topic models with numeric variables to represent latent events (Park, Lee, & Moon, 2015). Topic modeling has also been extended to multilingual environments to accommodate the need to process multilingual text resources (Vulić, De Smet, Tang, & Moens, 2015).

2.2. Applications of topic modeling

LDA-based topic modeling has become very popular in a variety of contexts involving natural language processing in information science, including information retrieval, informetrics and other applications.

Wei and Croft (2006) investigated how LDA-based document models may be used to improve ad-hoc retrieval. They demonstrated that the LDA model consistently outperformed a cluster-based approach and performed comparably to the Relevance Model that used pseudo-relevance feedback. Similarly, Yi and Allan (2009) compared different topic models and