Presentation Submission

Vocabulary Size and Its Effect on Topic Representation for Informetric and Information Retrieval Data Processing

Kun Lu
School of Library and Information Studies, University of Oklahoma,
401 West Brooks, Norman, OK 73019, USA Email: kunlu@ou.edu

Xin Cai
School of Information Studies, University of Wisconsin-Milwaukee,
P.O. Box 413, Milwaukee, WI 53201, USA Email: xincai@uwm.edu

Isola Ajiferuke
Faculty of Information and Media Studies, University of Western Ontario,
London, ON N6A 5B7, Canada Email: iajiferu@uwo.ca

Dietmar Wolfram
School of Information Studies, University of Wisconsin-Milwaukee,
P.O. Box 413, Milwaukee, WI 53201, USA Email: dwolfram@uwm.edu

Introduction

Topic modeling (Hoffman, 1999; Blei, Ng, & Jordan, 2003) is a machine learning technique applied to text corpora that was initially developed to reduce computational overhead. It has been widely used in information retrieval, and more recently has found application in informetrics research (Ding, 2011; Lu & Wolfram, 2012). By reducing the "dimensionality" of the computational task that could potentially involve hundreds of thousands 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 reduced. Forms of topic modeling have been applied in informetric study over the past decade, where similar issues of computational overhead are becoming more common as a result of larger, full text datasets now available. One form of topic modeling that has been used in informetric studies relies on Latent Dirichlet Allocation (LDA) (Blei, Ng, & Jordan, 2003). The LDA model treats a document as a mixture of topics and a topic as a mixture of terms. Rosen-Zvi et al. (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. If the vocabulary size could be reduced during training without significantly affecting the nature of the topics or document comparisons, this could also reduce the computational overhead in identifying the topics.

The present study explores the following research questions:

1) What impact does the removal of frequently 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 or infrequently occurring terms affect topic distributions in documents and the distinctiveness of the trained topics using entropy and pairwise topic similarity?

Method

Three datasets were used in the study: Ohsumed, TREC Genomics Track 2006 and SIGMET. The Ohsumed collection is a subset of MEDLINE with 348,566 bibliographic records. The Genomics collection includes 162,259 full-text articles published in biomedical journals. Since the Ohsumed and Genomics datasets are very large and would require much computation, a random sample of about 10% of each of the datasets was used in the study. This resulted in 34,846 bibliographic records in the Ohsumed subset and 16,201 full-text articles in the Genomics subset. SIGMET dataset was provided by Elsevier, Inc. consisting of 56,620 bibliographic records from 118 Elsevier Arts & Humanities journals. The LDA model was implemented in Java using the Gibbs Sampling inference method. The parameters are set as follows: alpha equals $50/K$ (K is the number of topics), beta equals 0.01 and the number of iterations is 1000. Topic modeling was applied using pre-specified numbers of topics (10, 20, 30, 40, 50, 100).

The datasets with the full vocabulary representation served as a baseline. Outcomes using several assessment measures were compared for the removal of singly occurring terms (removing the most term types), and the top 0.5%, 1% and 5% most frequently occurring terms, which remove the greatest number of term tokens.

To compare the document space generated from different vocabulary selection strategies or topic quantities, Document Space Density (DSD) was introduced to indicate the degree of scatter for documents. Here, each document was regarded as a point in a multi-dimension space and the position was determined by its topic distribution. The lower the DSD value, the more scattered the document space and the more distinctive each document is.

Information entropy was applied to interpret how topics are distributed in documents. A lower entropy indicates the documents concentrate on a few topics and a higher entropy means the topic distribution is more uniform.

Pairwise topic similarity was used to compare the topics by examining the vocabulary distribution in topics. A lower pairwise topic similarity means the topics are more distinctive. In other words, these topics are more differentiable. Cosine similarity was used to measure the pairwise topic similarity.

Results

Only a small representation of the findings can be presented here. The number of topics does not make a difference for the DSD, but the removal of larger numbers of terms increased the density, thereby decreasing the ability to discriminate between documents (Figure 1). For entropy, the number of topics does make a difference, with the lowest numbers of topics having the lowest entropy (Figure 2). The vocabulary size does not play a big role, so singly occurring and the very high frequency terms may be removed without greatly affecting the topic distributions in documents. Using pairwise topic similarity, both the number of topics and vocabulary size play a big role in the topic distinctiveness, where higher numbers of topics provide more distinctiveness (Figure 3). Although removing more terms also reduces topic similarity, the corresponding increase in DSD must be kept in mind. The outcomes were quite similar across the datasets for the different measures.

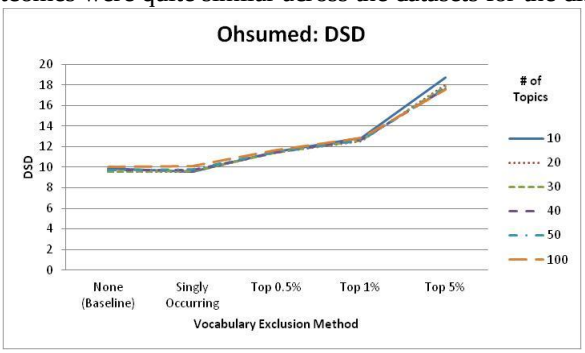

Figure 1 – Document Space Density Outcome

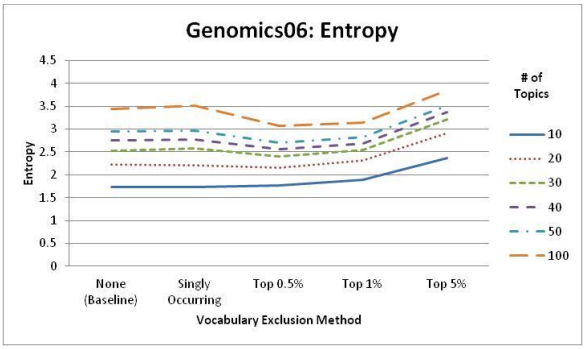

Figure 2 – Entropy Outcome

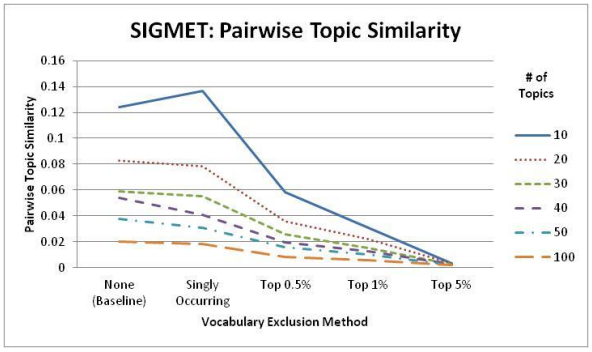

Figure 3 –Pairwise Topic Similarity Outcome

Conclusion

When using topic modeling techniques such as LDA for metrics research it is possible to reduce the vocabulary size for text corpora of interest without the loss of topic or document distinctiveness, depending on the number of topics selected. Applications extend to metric comparisons among entities of interest (e.g., authors, journals, research units), where full text or extended bibliographic text datasets are available.

References:

Blei, D. M., Ng, A. Y., & Jordan, M. J. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research*, *3*, 993-1022.

Ding, Y. (2011). Topic-based PageRank on author cocitation networks. *Journal of the American Society for Information Science and Technology*, 62(3), 449-466.

Hofmann, T. (1999). Probabilistic latent semantic indexing. In *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (ACM SIGIR'1999) (pp. 50-57).* Berkeley, California, USA: ACM.

Lu, K., & Wolfram, D. (2012). Measuring author research relatedness: A comparison of word-based, topic-based, and author cocitation approaches. *Journal of the American Society for Information Science and Technology*, 63(10), 1973-1986.

Rosen-Zvi, M., Chemudugunta, C., Griffiths, T., Smyth, P., & Steyvers, M. (2010). Learning Author-Topic models from text corpora. *ACM Transactions on Information Systems*, *28*(1), 1-38.