Topic Modeling Approaches for Text Categorization in Micro-blogs

Maxwell Tetteh

Amrita Vishwa Vidyapeetham – Amritapuri

Word Count: 1300

Authors	Origin	Purpose	Research design	Conceptual framework	Methodology / Technique	Metric Evaluation	Framework proposed	Major Themes
Toni et al. (2016)	USA	Comparing the ability of LSA to Latent Dirichlet Allocation (LDA) to categorize large size patent data into meaningful groups.	Experimental research	No	Evaluating the algorithms in direct juxtaposition to one another	Document-similarity matix	No	Latent Semantic Analysis, Patent based design, PCA
Mehrotra et al. (2013)	Australia	To improve topics learned from Twitter content without modifying the basic machinery of LDA	Quantitative	Tweet aggregation based on similarity	Comparison of Pooling schemes	Purity And NMI PMI Score	Hashtag-based Pooling Schemes	Tweet Pooling, Automatic Labeling, LDA
Lin et al. (2014)	China	Correct sparsity in short text using Dual-sparse topic models	Quantitative	'Spike and jab' prior to decouple document sparsity and smoothness	Applying Smoothing Prior and DsparseTM on short text	PMI Score, Dual-sparsity ratio	Dual-Sparse Topic Model	Sparsity, Document Pooling, Sparsity enhabced topic models
Sriram et al. (2010)	USA	To use a small set of domain-specific features extracted from the user profile and texts to better process short-text	Correlational study	8 Features (8F)	BOW BOW-A 7F+BOW	K-Fold Cross Validation	8F coupled with BOW (with authorship)	Micro-blogging, short text, Authorship
Qiang et al.(2019)	China	Conduct a comprehensive review of various short text topic modeling techniques	Descriptive Case-study	GS-DMM GPU-DMM	PLDA LDA	Metric Purity And NMI	Dirichlet multinomial mixture (DMM) based methods, Global word co-occurrences based methods.	Dirichlet Multinomial Mixture based Methods, Parameter Setting, Topic Coherence

Topic Modeling Approaches for Text Categorization in Micro-blogs

The rate of creation of content on current social media platforms signifies an insurgence in textual data than has ever been available. This quantity of data, however, poses a bigger problem for the administrators of these 'creator' accounts who have to manually sort through countless amounts of user feedbacks in order to try and understand how their content is perceived by other viewers. This literature survey focuses on micro-blog platforms, making use of different topic modeling algorithms to discover latent topics in user posts.
It will draw on studies conducted by other related researches to demonstrate how a particular set of data or chosen parameters affect the outcome of classification or categorization in generating topics from a given text corpus.

Additionally, it will examine the difference between Short Text documents and more structured blocks of documents such as news, blogs, and research papers. These findings will further help us understand the concept of micro-blogging, the main limitations of working with short text corpus, and how classical topic models can be optimized without affecting the primary operating mechanism of these algorithms.

Although the implementation of topic modeling algorithms to discover concurrence and latent similarities between documents is considered a very efficient computational text processing technique, there are a number of factors at play to determine which modeling algorithm may be well suited for an intended task. This primarily depends on the type of the text or document repository that is being analyzed. Toni [1] examined the efficiency of the Latent Semantic Analysis algorithm, comparing it with the Latent Dirichlet Allocation (LDA) approach. Preliminary studies were focused on LSA-which was used to categorize patent documents in an existing mechanical patents database. A major constraint with the algorithm, however, was that it was used to train a smaller scale of data (about 100 Patent documents). Increasing the size of the repository to an exponential amount implied significant reduction in the relevance and similarity of the patents. Despite the dissimilarity in accuracy as the quantity of documents changes, the repository is a text-rich data source and hence would not require any significant modifications to the algorithm or data in order to enhance efficiency of categorization [2].

MICROBLOGGING AND SHORT-TEXT TOPIC MODELS

Current social media platforms have introduced the concept of micro-blogging where users can post or comment a limited number of text characters to express their thoughts or emotions about events or their daily activities. The concept of limited words means users get to convey a lot of information with few words as possible leading to sparseness in the text that is generated

[3]. One challenge with analyzing these sparse documents is the lack of concurrency in words. This makes the outcome of generic topic modeling techniques vague as the changes of terms re-appearing in a document or cluster of documents is significantly less compared to when working with Long-text topic modeling.

Sriram et al. [4] illustrates a very interesting approach to tackling the problem of short-text classification. As compared to TweetStand (Sankaranarayanan et al.) which was only able to classify tweet data as news or non-news, this method had a lot more label options. The model used classified tweets (short-text) into these categories: News (N), Events (E), Opinions (O), Deals (D), and Private Messages (PM). The sorting process is a reverse strategy that focuses on authorship since empirical results have proven that different tweets from the same user tend to have the same category.

In Lin [3], we see a new way to reduce sparseness in short text called the dual-sparse topic model. Heuristic methods to improve word co-occurrence in documents include document pooling, conceptualization, and self aggregation [5]. By merging short texts into long pseudo-documents before topic inference, the level of word co-occurrence information in the documents significantly increases producing better performance with the topic model. Unlike all the other quantitative analysis on text data in topic modeling, Lin [3] argued that instead of reducing the sparsity in the Topic or Term of a corpus, we could drastically offset sparsity by targeting both the topic and terms in the document. This approach of dual-sparsity, though not very popular, was able to establish a clearer understanding of the context of short-text data by decoupling the sparsity and the smoothness of the document-topic and topic-word distributions using a ‘Spike and Slab’ prior.

CONCLUSION:

Traditional topic models have been very efficient in classifying and mining latent topics in large, properly curated text documents. This is due to the size of these ‘long text’ corpora as well as the range of vocabulary that is used. Unlike well-structured long text documents, the semantic structure of short text sources from micro-blogs and social media platforms seem have a high level of sparseness coupled with low word co-occurrence and is therefore a main cause of sub-optimal performance of classical topic modeling algorithms in short texts. Though smoothing and word-concurrence strategies could be applied to these short text in order for them to run LDA, PLDA or LSA algorithms, it has been established that sparsity could only generally be reduced in only one parameter while working with traditional topic models. Sparsity-enhancement topic models such as Sparse Topic Coding (STC) and Dual-Sparse TopicModel

(DsparseTM) have rather proven to be able to discover meaningful latent topics in short text more efficiently [3].

REFERENCES

[1] Cvitanić, Tonči, Bumsoo Lee, Hyeon Ik Song, Katherine K. Fu and David W. Rosen. “LDA v. LSA: A Comparison of Two Computational Text Analysis Tools for the Functional Categorization of Patents.” *ICCBR Workshops* (2016).

[2] Mehrotra, Rishabh, Scott Sanner, Wray L. Buntine and Lexing Xie. “Improving LDA topic models for microblogs via tweet pooling and automatic labeling.” *Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval* (2013): n. pag.

[3] T. Lin, W. Tian, Q. Mei, C. Hong, The dual-sparse topic model: mining focused topics and focused terms in short text, in: International Conference on World Wide Web, 2014.

[4] Sriram, Bharathi, David Fuhry, Engin Demir, Hakan Ferhatosmanoğlu and Murat Demirbas. “Short text classification in twitter to improve information filtering.” *Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval* (2010): n. pag.

[5] Qiang, Jipeng, Qian Zhenyu, Li Yun, Yuan Yunhao and Wu Xindong. “Short Text Topic Modeling Techniques, Applications, and Performance: A Survey.” *IEEE Transactions on Knowledge and Data Engineering* (2020): n. pag.