Word Community Allocation: Discovering Latent Topics via Word Co-Occurrence Network Structure

Michael D. Salerno Christine A. Tataru Michael R. Mallory

December 9, 2015

1 Introduction

The increasing ubiquity of digitally stored knowledge has motivated significant amounts of research into statistical models for discovering abstract topics occurring within collections of documents. This family of models, commonly known as topic models, has applications within a variety of fields involving collections of data; canonical examples come from the field of text mining, however, data from domains such as content-based image retrieval, collaborative filtering, and bioinformatics have also been conducive to topic modeling. That said, difficulties in applying topic modeling techniques persist by virtue of heterogeneity between the structure and content of modern document collections. In this paper, we consider the problem of latent topic discovery on collections of sparse documents, commonly referred to as short text.

Short text has become a pervasive form communication due to the surge of web applications and mobile devices that have accompanied the development of the internet and cellular networks. Examples include micro-blogging, search queries, forum messaging, instant messaging, cellular Short Message Services (SMS), and comments on internet content. These media produce document collections that are of interest to a wide range of stakeholders. Consider Twitter, a popular micro-blogging service with - as of 2015 - 307 million active users producing over 500 million messages everyday. Targeted collection and analysis of these messages has been used for tasks such as assessing the efficacy of marketing campaigns, tracking the spread of social unrest, better understanding terrorist recruitment tactics, and monitoring trending topics. However, despite the significant advances in the field of topic modeling, short text continues to pose a challenge due to the sparsity of information contained in each document. State-of-the-art topic models such as Latent Dirichlet Allocation (LDA) and extensions to LDA construct topics by using document-word co-occurrence information [1]. LDA is designed to perform well on document collections in which documents are reasonably long and have little or no topic skew. In the case of short text, word-by-document spaces end up being extremely sparse resulting in poor topic quality. In addition to document sparsity, topic skew is also a common difficulty in topic modeling using short text. Except for very specialized collections, short text collections tend to over-represent more popular topics. This topic imbalance

will result in most modern topic models failing to detect relatively obscure topics because their objective functions tend to maximize the probability of the observed data over a set of topics [9].

We propose Word Community Allocation (WCA), a network-based alternative to the LDA family of topic models. WCA constructs a network based on word co-occurrences and then uses community detection to discover groups of words belonging to latent topics within the document collection. The advantages of WCA include robustness to sparse document collections and robustness to topic skew. These properties are attributable to the fact that, unlike standard topic modeling techniques, WCA searches a word-by-word space instead of a word-by-document space for groups of words representing topics. This is described in more detail in section 3.

2 Related Work

Latent Dirichlet Allocation (LDA) is considered the state-of-the-art in topic modeling as it provides the first fully probabilistic model for text modeling [1], improving upon previous models such as Latent Semantic Indexing (LSI) [4] and Probabilistic Latent Semantic Indexing (pLSI) [5]. Many extensions to LDA and pLSI have been proposed such as dynamic topic model, author-topic model, author-topic-community model, and social topic model. These extensions are similar to LDA, however, they each incorporate some specialized information about the document collection in order to enhance performance, such as social relationships and authorship.

Research efforts into topic modeling on short text has also focused on using ancillary information about document collections. For example, Hong et al. [10] improve the performance of topic models trained on tweets by aggregating tweets that share the same word. Blei et al. [11] have also proposed a supervised topic model for use with labeled document collections. Zuo et al. [3] have proposed a word-network-based topic model called Word Network Topic Model (WNTM) which generates a corpus of pseudo-documents on which to apply standard Gibbs sampling, as in LDA, in order to generate the topic model.

In contrast to most previous approaches, and using an intuition similar to WNTM, we seek to handle sparse and imbalanced texts without relying on special properties or assumptions on the document collection. WCA differs from WNTM in that community detection methods are used in place of Gibbs Sampling in order to assign words to topics and also in that WCA embeds word-proximity information into the edge weights.

Infomap is one of the leading community detection algorithms, as determined by a study by Gnce Keziban Orman, Vincent Labatut, and Hocine Cherifi in On Accuracy of Community Structure Discovery Algorithms [12]. In order to compare efficiency and performance of the most popular community detection algorithms, Orman et al. sampled well known algorithms from the following categories : link-centrality-based algorithms, modularity optimization algorithms, spectral algorithms, random-walk-based algorithms, information-based algorithms, and other algorithms. One or two best known algorithms from each category was run on realistic simulated data comparable to those modeled by Newman and Girvan [13, 14] to benchmark each respectively. Of the eleven

algorithms tested, the top performer proved to be Infomap, which outperformed all others in every case, even in those with high mixing coefficients, followed by Walktrap, SpinGlass and Louvain.

Wallach et al. [6] provide three different algorithms for the evaluation of topic models. These algorithms will each be considered for use in the evaluation phase of this project.

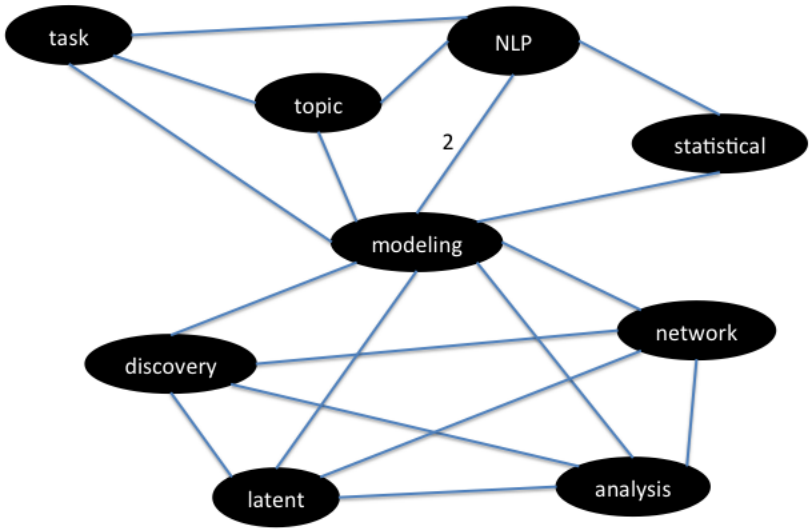
3 Word Community Allocation

WCA constructs a network out of a document collection in which nodes represent individual words in the vocabulary and links represent the co-occurrence of a pair of words within a document. Links are weighted based on frequency of co-occurrence and the context of co-occurrences. Each time a pair of words occurs together within a document, the link between their nodes is up-weighted by an amount determined by the context of their co-occurrence; for example, a co-occurrence within a title or within the same sentence carries more weight than only co-occurring within the same paragraph. In this manner, information about the proximity of word co-occurrences is embedded into the network. WCA then uses Louvain's community detection algorithm to discover groups of words belonging to latent topics within the document collection. The salience of the words belonging to each topic can then be assessed using community-level network statistics. WCA is robust to sparse documents because it constructs topics using word-word co-occurrence information. Unlike the word-by-document representation used by the LDA family of models, word-by-word spaces remain relatively dense in spite of sparse documents. The issue of topic skew is also mitigated because the number of words belonging to rare topics is generally much greater than the number of documents belonging to rare topics. In addition to its robustness to the characteristics of short text, embedding co-occurrence proximity information into network edges allows WCA to take advantage of the additional contexts provided by longer documents.

The following is an illustration of the basic idea behind how the WCA network is generated. Consider three documents:

document	body
doc1	Topic modeling and NLP tasks
doc2	Statistical modeling and NLP
doc3	Latent topic discovery via network analysis

When a pair of words co-occurs in a document, a link is formed between them in the graph. If a link between them already exists, then the existing link is up-weighted. The corresponding graph is displayed below. The weight between the words "modeling" and "NLP" is 2 because they co-occurred in doc1 and in doc2. WCA will determine how much to up-weight the edges based on the context in which the words co-occurred.



3.1 Louvain's Algorithm

We use Louvain's Algorithm as a means of discovering latent topics from the structure of the WCA word co-occurrence network. In Louvain's greedy method, every node starts off in its own module. Every node is then considered sequentially, and merged with the module of the neighboring nodes that would most increase the modularity of the system, defined as follows:

$$Q = \sum_{vw} \left[\frac{A_{vw}}{2m} - \frac{k_v k_w}{(2m)(2m)} \right] \delta(c_v, c_w)$$

Where k_v and k_w are the degrees of nodes v and w respectively, and $\delta(c_v, c_w)$ is 1 if v and w are in the same community, and 0 otherwise. If there is no merge that would yield an increase in modularity, the node remains by itself. This continues until every node has been viewed once. This process concludes phase one.

In phase two, a new network is formed, with each module of nodes obtained from the first round is considered its own large node. Weighted edges are placed based on how many nodes in each new module have edges connecting them to nodes in another module. The algorithm is run again, forming the second hierarchical layer of clustering. This continues until no merges would increase the modularity of the system.

4 Data Collection

For initial testing, we pulled datasets from the University of California Irving Machine Learning Repository, specifically their Bag of Words dataset, which was designed for clustering and topic modeling. For demonstration purposes, we used their smallest dataset, which is a collection of 3.4 thousand blog articles from dailykos.com, with over 6,906 words. Word counts where filtered to only include words that appeared more the 25 times in a document. These datasets have been stripped of all document metadata, document structure, word ordering, and labeling.

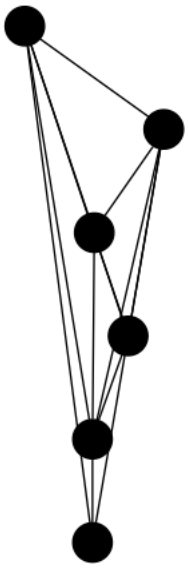
Subsequently, we acquired the BBC Sports dataset from the Insight Centre for Data Analytics. The BBC Sports dataset is a collection of sport related documents broken down into 5 different categories, Athletics, Football, Tennis, Rugby, and Cricket. It contains 742 documents that are unevenly spread across the 3.2 MB dataset. Because of its small size, it was perfect for testing multiple approaches. The fact that the corpus contains documents of various lengths, many small enough to be considered short text, also makes it conducive to testing for robustness against sparse documents.

5 Results

Here we present the results of applying the WCA work-flow to our datasets. Initial tests were used to develop a general sense of the interpretability of WCA's results. Indeed, topic modeling applications generally involve human interpretation of word clusters for topic determination. Subsequent testing was focused on quantitatively measuring WCA's performance.

5.1 Initial Testing

A word co-occurrence network was constructed using a set of articles by the blog kosdaily.com. Edges are currently based on co-occurrence frequency. Once the words had each been assigned to a community, we aggregated the nodes within each community to produce the following simple visualization which suggests the existence of six major topics within the document collection.



Bag of Words KOS

The following are the five most frequently occurring terms in each community:

1: Bush Cheney, NPR, Kerry Edwards, Iraqi, Ward

2: Dean Clark, Kerry Edwards, Eerily, Clarke, Democratic Leaning

3: Blacks, Whites, Peoples, Bush Cheney, Races

4: Gov, Generally, Kerry Edwards, Peoples, Elections

5: Bunnings, Senate Governors, Races, Monitor, Republicans

6: Boxfeed Listing, Dress, Republicans, GOPs, Senate Governors

The dataset these lists where generated from was a set of articles by the blog kosdaily.com, a political blog. The above lists contain the most frequently occurring terms in each community. Considering the dataset was acquired soon after the 2008 election, it is unsurprising that Bush/Cheney and Kerry/Edwards are high on every list, as these names are likely to occur multiple times each article. To create the communities, we used the info words algorithm as a simple demonstration of latent topic discovery. Even though the lists contain a number of common terms, in their entirety, the communities are fairly heterogeneous. These short lists already suggest the discovery of coherent topics; for example, (3) appears to represent the topic of race in the general election, (2) seems to represent the topic of the democratic primary elections, and (5) contains words related to the senatorial elections.

5.2 Experiments

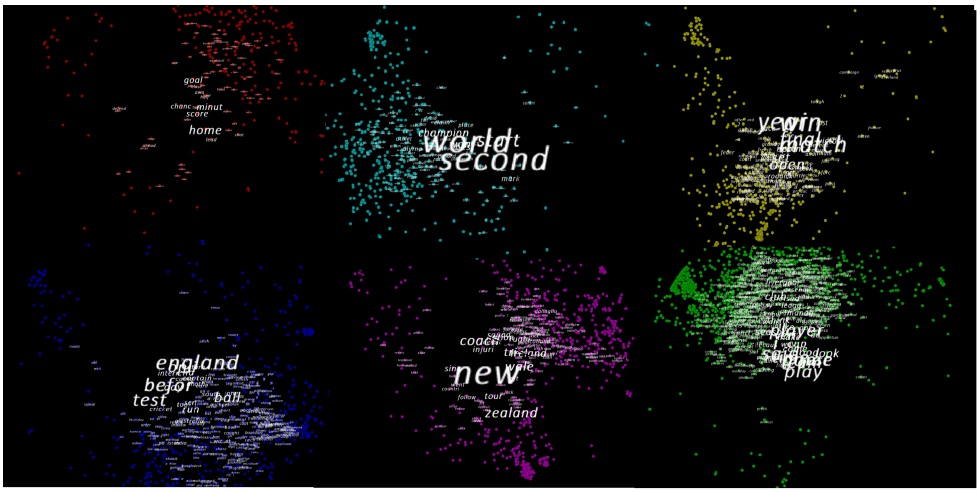
In order to quantitatively assess WCA's ability to discover latent topics, we applied it to a training subset of BBC Sports dataset and then used the clustering discovered by WCA to label the documents in the holdout set. The error rate was then compared to the results achieved by two null models; one which classifies documents randomly and another which classifies documents based on the most common label in the training set. Most topic discovery applications involve unlabeled document sets; thus, a favorable comparison with the null models suggests that the algorithm would be capable of detecting useful signal in such a scenario.

Classification of a document using WCA's clustering was performed as follows:

1. Create a Boolean vector of all possible words, $\mathbf{w}_\theta$, one for each cluster θ. An index is 1 if that word appears in that cluster, 0 otherwise.
2. Take the word count vector for the document, $\mathbf{c}$, in which each index contains the number of times the corresponding word occurred in the document.
3. Compute the cluster assignment using $\arg\max_\theta(\mathbf{w}_\theta \cdot \mathbf{c})$
4. Classify the document as the label that is most represented in the assigned cluster.

Using this method, a misclassification rate of 31.43% was achieved. Classifiers which assigned a random labeling and labeled all documents based on the most common training label achieved 78.09% and 64.28% misclassification rates, respectively.

A visualization of the topic clustering produced during this experiment is displayed below:



5.3 Conclusions

We find that WCA is capable of producing cohesive and interpretable word clusters. Furthermore, we found that these clusters were useful in determining document labels during experiments involving documents which were labeled by topic. This indicates that the clusters discovered by WCA are suggestive of the true topics expressed by the document corpus. These results were achieved in spite of the presence of short text in the corpus that was used for experiments; this suggests that the WCA is capable of alleviating the difficulty of latent topic discovery in the presence of sparse documents.

Future work for WCA will involve more thoroughly examining the properties of the resulting clusters by considering the significance of network statistics such as measures of centrality in the context of topic discovery. Such measures may provide useful information about the salience of the words belonging to each topic. Additionally, thorough experiments involving different types of document sets must be performed in order to precisely determine WCA's performance and utility for other use cases. It is also likely that WCA can be improved by considering community detection algorithms other than Louvain's method for topic search, such as InfoMap. Lastly, a probabilistic version of WCA should be considered by exploring ways to fit a model to the WCA network using maximum likelihood or Bayesian methods.

References

[1] D.M. Blei, A.Y. Ng, M.I. Jordan, "Latent Dirichlet Allocation", *Journal of Machine Learning Research*, 2003.

[2] M. Newman, "Detecting community structure in networks", *The European Physical Journal B - Condensed Matter and Complex Systems,Volume 38(Issue 2), Pp 321-330.*, 2004

[3] Y. Zuo, J. Zhao, K. Xu, " Word Network Topic Model: A Simple but General Solution for Short and Imbalanced Texts", *Knowledge and Information Systems*, 2014.

[4] C. Papadimitriou, P. Raghavan, H. Tamaki, S. Vempala, "Latent Semantic Indexing: A probabilistic analysis", *Proceedings of ACM PODS*, 1993.

[5] T. Hofmann, "Probabilistic Latent Semantic Indexing", *Proceedings of the Twenty-Second Annual International SIGIR Conference on Research and Development in Information Retrieval*, 1999.

[6] H. M. Wallach, I. Murray, R. Salakhutdinov and D. Mimno, "Evaluation Methods for Topic Models", *Proceedings of the 26th International Conference on Machine Learning (ICML)*, 2009.

[7] Reuters Corpora (RCV1, RCV2, TRC2)," 13 8 2015. [Online]. Available: http://trec.nist.gov/data/reuters/reuters.html.

[8] D.D. Lewis, "The LYRL2004 Distribution of the RCV1-v2 Text Categorization Test Collection", 12 4 2004. [Online]. Available: http://www.ai.mit.edu//projects/jmlr/papers/volume5/lewis04a/lyrl2004_rcv1v2_README.htm.

[9] J. Jagarlamudi, H. Daumé III, R. Udupa, "Incorporating Lexical Priors into Topic Models", *EACL*, 2012.

[10] Hong, L., Davison, B.D. Empirical study of topic modeling in twitter. In: SOMA, (2010)

[11] Mcauliffe, J.D., Blei, D.M. Supervised topic models. J. Platt, D. Koller, Y. Singer, S. Roweis Advances in Neural Information Processing Systems 20, (2008)

[12] Gnce Keziban Orman, Vincent Labatut, and Hocine Cherifi, On Accuracy of Community Structure Discovery Algorithms, Journal of Convergence Information Technology 6(11):283-292, 2011

[13] Michelle Girvan, Mark Newman, Community structure in social and biological networks, Proceedings of the National Academy of Sciences of the United States of America, vol. 99, no. 12, pp. 7821-7826, 2002.

[14] Mark Newman, Fast algorithm for detecting community structure in networks, Physical Review E, vol. 69, no. 6, p. 066133, 2004.