Document Clustering and Visualization with Latent Dirichlet Allocation and Self-Organizing Maps

Jeremy R. Millar and **Gilbert L. Peterson** and **Michael J. Mendenhall**
Air Force Institute of Technology
2950 Hobson Way, WPAFB OH 45433-7665

Abstract

Clustering and visualization of large text document collections aids in browsing, navigation, and information retrieval. We present a document clustering and visualization method based on Latent Dirichlet Allocation and self-organizing maps (LDA-SOM). LDA-SOM clusters documents based on topical content and renders clusters in an intuitive two-dimensional format. Document topics are inferred using a probabilistic topic model. Then, due to the topology preserving properties of self-organizing maps, document clusters with similar topic distributions are placed near one another in the visualization. This provides the user an intuitive means of browsing from one cluster to another based on topics held in common. The effectiveness of LDA-SOM is evaluated on the 20 Newsgroups and NIPS data sets.

Introduction

Automatic organization of document collections into clusters of related documents has been shown to significantly improve the results of information retrieval systems (Salton and McGill 1983; Deerwester et al. 1990; Kaski et al. 1996). Visualization of document collections provides users with an overview of the collection and enables them to perform exploratory data analysis (Feldman and Sanger 2007). Clustering and visualization form the basis of modern interactive information retrieval systems. Dividing a collection into clusters of similar documents improves performance and reduces cognitive load on the user. Visualization of the clusters enables intuitive search, browsing, and navigation of a document collection.

This article presents an approach to clustering and visualization of document collections based on Latent Dirichlet Allocation (LDA) and self-organizing maps. LDA is a probabilistic topic model capable of inferring a topic distribution based on word content for each document in a collection. LDA-SOM operates directly on these topic distributions rather than word histograms in order to reduce the dimensionality of the document representations. Additionally, topic distributions form a semantic space; therefore, our approach clusters documents based on content or meaning rather than word distribution alone.

Copyright © 2009, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

Self-organizing maps (SOMs) are unsupervised neural networks based on competitive learning. SOMs operate on a fixed, low-dimensional (typically two) lattice that lend themselves to a variety of visualizations. Moreover, the mapping from the original data space onto the lattice is topology preserving so that samples near one another in the data space are placed near one another on the lattice. Consequently, documents with similar topic distributions are mapped to similar lattice points.

Once the documents are mapped onto the SOM lattice, the lattice itself is clustered using k-means. Plotting the lattice along with cluster boundaries provides a simple and intuitive view of the document collection. Because of the topology preservation property of SOMs, neighboring clusters often have one or more topics in common, based on the mean topic distribution for each cluster. This property makes it easy for the user to visually identify related groups of documents.

Related Work

Clustering document collections has a unique set of challenges when compared to non-text oriented data mining tasks. Foremost, the data is textual in nature, and highly unstructured. Leveraging statistical machine learning techniques requires encoding raw text into a form consumable by traditional clustering algorithms. Second, the encoded data often has extremely high dimensionality. Many of these dimensions have little or no discriminatory value, necessitating some form of feature selection or dimensionality reduction prior to the actual clustering.

The encoding challenge is generally met by encoding the documents according to the *vector space model* (Salton and McGill 1983). In this model, each document is tokenized and encoded as an m-dimensional vector, where m is the total number of unique tokens in the entire collection and each vector component is a frequency count of that token's occurrence in a given document. Given a collection with m unique words, a document is represented as a vector $\mathbf{d}$ such that each d_i is a count of how often word i occurs in the document, i.e., the vector $\mathbf{d}$ represents a word histogram.

It is not uncommon for even a modest collection of documents to have a large number of unique words. Consequently, some form of dimensionality reduction is necessary to reduce computational run-times and effects of the curse of dimensionality. The reduction should preserve as much

information as possible with respect to the desired cluster relationships in order to maintain valid clusters. For our purposes, we are interested in clustering documents by topic; that is, documents about the same subject should be clustered together. Consequently, our dimensionality reduction must preserve topical and semantic information.

Several possibilities exist for accomplishing this dimensionality reduction, including Latent Semantic Indexing (LSI) (Deerwester et al. 1990), probabilistic LSI (pLSI) (Hofmann 1999), Latent Dirichlet Allocation (LDA) (Blei, Ng, and Jordan 2003), self-organizing context maps (Kaski et al. 1996; Kohonen 2001), and random projection (Kaski 1998; Kohonen 2001; Lagus, Kaski, and Kohonen 2004).

Kohonen et al. developed the WEBSOM method for clustering document collections (Kaski et al. 1996; Lagus et al. 1999; Kohonen 2001; Lagus, Kaski, and Kohonen 2004). Early versions of WEBSOM employed a two-layer architecture based on word context vectors. Each word in the vocabulary is assigned a random code vector and the expected values for the previous and next words across the entire collection are calculated. The code vectors for the previous, current, and next words are concatenated to form a *context vector*, i.e., the vector

$$\begin{bmatrix} E(w_{j-1}) \\ w_j \\ E(w_{j+1}) \end{bmatrix}$$

is computed for each word in the vocabulary, where w_j is the code vector for the jth word and $E(w_{j-1})$ and $E(w_{j+1})$ are the expected code vectors for the previous and next words, respectively.

A self-organizing map is then trained using the context vectors as input. Words with similar context end up near one another, resulting in a *context map*. Each document's text is mapped onto the context map, forming a histogram of hits. A second SOM is trained using the hit histograms of each document to form a document map. Documents with similar contexts cluster together on the document map. A side-effect of this approach is the creation of a visually appealing two-dimensional display of the document clusters.

Ampazis and Perantonis have developed the LSISOM method for document clustering (Ampazis and Perantonis 2004). LSISOM is quite similar to WEBSOM; the primary difference lies in the use of LSI-based term vectors rather than statistical context vectors in the generation of the context map. To compute these vectors, the document vectors are arranged row-wise to form a sparse term-document matrix $\mathbf{M}$ and the singular value decomposition (Golub and Van Loan 1996) is computed. The largest k singular values and associated singular vectors are retained while the rest are discarded in a fashion similar to principal component analysis. This process effectively collapses the original document space onto a smaller semantic vector space. The projections of the original terms into this space are clustered on an SOM to form a context map as in WEBSOM. Hit histograms for each document are computed on this context map and the resulting document representations clustered on a second SOM.

Recent document clustering work applies LDA to the document clustering problem using an implementation based on Gibbs sampling (Griffiths and Steyvers 2004). The document collection's vector space representation is transformed into a lower dimensional topical representation. Clustering is performed in this space using traditional techniques such as k-means with symmetrized Kullback-Liebler divergence or Jensen-Shannon divergence as a metric. In addition to document clustering, Griffiths and Steyvers have applied their approach to the identification of hot and cold topics and topic trends within scientific communities (Griffiths and Steyvers 2004; Steyvers et al. 2004). Unlike the WEBSOM approach, their technique has no built-in visualization component. Typically, visualizing document clusters given by this method requires some form of multi-dimensional scaling and projection pursuit.

Probabilistic Latent Semantic Visualization (PLSV) extends pLSI to include Euclidean coordinates as latent variables within the topic model (Iwata, Yamada, and Ueda 2008). This method shows considerable improvement over traditional methods such as multi-dimensional scaling and local linear embedding. PLSV is unique in that the coordinates required for visualization are explicitly accounted for in the model. Under PLSV, document generation begins by choosing a set of coordinates for the document. Topics are chosen from a multinomial distribution conditioned on the document location, and words chosen from each topic. While novel and effective, the PLSV model seems contrived since no author begins authoring a document by considering its location on the plane.

LDA-SOM

LDA-SOM combines probabilistic topic models and self-organizing maps to cluster and visualize document collections. It is similar to LSISOM, although LDA-SOM uses only one SOM and applies LDA rather than LSI as a dimensionality reduction technique. Applying LDA rather than LSI provides a document model that is both richer in semantic content and more statistically sound (Blei, Ng, and Jordan 2003). Additionally, LDA-SOM eliminates the use of an SOM to generate document contexts since document topic distributions provide context automatically. Once the documents are clustered on the SOM, k-means clustering is applied to the SOM nodes themselves to aid in visualization.

Latent Dirichlet Allocation

Latent Dirichlet Allocation (LDA) (Blei, Ng, and Jordan 2003) is a generative latent variable model for documents. Documents are modeled as distribution of topics, and each topic is modeled as a distribution of words.

Figure 1 presents a graphical representation of the LDA model. Here, shaded nodes are observed variables and unshaded nodes are latent variables. Arrows represent dependencies. The boxes (or plates) represent repeated sampling operations. The values M, N, and T are the number of documents in the collection, the number of words per document, and the number of topics, respectively. The value z is the topic from which a particular word w is drawn. The

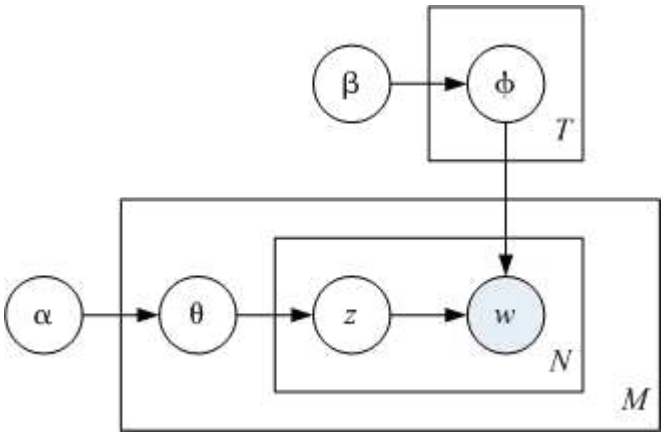

Figure 1: Graphical representation of LDA (Blei, Ng, and Jordan 2003). Shaded nodes represent observed variables; other nodes represent latent variables. Plates represent repeated operations.

per-document multinomial topic distributions are given by θ, while ϕ gives the per-topic multinomial word distributions. Dirichlet priors are placed over these distributions, parameterized by α and β.

LDA is a straightforward process:

1. Choose values for the hyperparamters α and β and the number of topics T. The values of α and β depend on T and the vocabulary size; good general choices are $\alpha = 50/T$ and $\beta = 0.01$ (Steyvers and Griffiths 2007).
2. For each document:
   (a) Choose the number of words N.
   (b) For each word:
      i. Sample z from $\theta^{(j)}$, where j is the current document index.
      ii. Sample w from $\phi^{(z)}$.

To perform document clustering using LDA, we must find $P(\mathbf{z}|\mathbf{w})$ for fixed α, β and T. In general, this problem is intractable (Blei, Ng, and Jordan 2003); common approximation techniques include variational EM (Blei, Ng, and Jordan 2003) and Gibbs sampling (Griffiths and Steyvers 2004), which is more common and followed here.

Once $P(\mathbf{z}|\mathbf{w})$ is calculated, the distributions ϕ and θ can be estimated for each topic and document. The topic distributions θ form the basis of our clustering method.

Self-Organizing Maps

Self-organizing maps (SOMs) are a form of unsupervised neural network developed by Kohonen (Kohonen 2001). SOMs consist of a fixed lattice (typically 2-dimensional) of processing elements. Each processing element has an associated (initially random) prototype vector.

Learning in the SOM takes place in a competitive fashion. For each input, the processing element with the shortest Euclidean distance (the best matching unit, or BMU) is identified. The prototype vector for this element and all other elements within a neighborhood are updated according to

$$w_j(t+1) = w_j(t) + \alpha(t)h_{ji}(t)(x_m - w_j)$$

where w_j is the prototype vector associated with the jth processing element, $\alpha(t)$ is a monotonically decreasing learning rate, $h_{ji}(t)$ is a time-decreasing neighborhood function, and x_m is the input sample. Typically, the neighborhood is a Gaussian function. Over time, the map converges to a low-dimensional representation of the original input space.

Self-organizing maps naturally cluster the input data so that inputs with similar features are mapped to the same or neighboring processing elements. Moreover, SOMs preserve the topology of the original high-dimensional input space on the lattice, i.e., relationships between samples in the high-dimensional input space are preserved on the low-dimensional mapping (Kohonen 2001; Bauer and Pawelzik 1992). These properties make the SOM an ideal tool for visualizing high-dimensional data in 2-dimensional space.

The LDA-SOM Method

The LDA-SOM approach to document clustering uses LDA for dimensionality reduction and the SOM for clustering and visualization. This approach results in a map of topical clusters. Documents within each cluster share similar topics, and neighboring clusters may have one or more topics in common. This unique layout allows users to quickly browse through a document collection. It can also indicate relationships between documents that might otherwise go unnoticed.

As in most text mining and information retrieval tasks, the process begins by preprocessing the document collection. Stop-words, i.e., definite articles, prononuns, and other words of little discriminative value are removed from the collection's vocabulary. Additionally, exceptionally rare words, e.g., those appearing in fewer than three documents, can be removed. Documents are then encoded as word histograms based on word occurrence frequency. Additional weighting schemes such as inverse document frequency may also be applied at this stage.

Once the document collection is encoded, LDA is applied to the word histograms to deduce the topics and topic distributions and to reduce the dimensionality of the input space. The number of topics is a parameter that must be set by the user; 50 to 300 topics are identified as good general purpose values (Wei and Croft 2006). Additionally, values for the α and β hyperparameters must be chosen.

Following topic modeling, an SOM is trained using the document topic distributions as input. The size of the map is determined based on the ratio of the largest two eigenvalues of the topic distributions. This ensures there are enough nodes to accurately capture the dimensions with the most variance. The SOM neighborhood function $h_{ji}(t)$ decays linearly, while the learn rate $\alpha(t)$ decays exponentially. The SOM prototype vectors are initialized randomly.

After the SOM converges, the prototype vectors are clustered using k-means. The optimal value of k is determined by minimizing the Davies-Bouldin index (Davies and Bouldin 1979) for $k = 1 \ldots \sqrt{n}$, where n is the number of

nodes in the SOM. Each node of the SOM is labeled based on its assigned cluster, and these labels are back-propagated to the document vectors.

The clustering results are visualized by rendering the SOM lattice and coloring each node according to its label. Cluster density can be visualized by rendering the alternative U-matrix for the map.

The methodology can be summarized as follows:

1. Pre-process and encode data as word histograms. Stop-words may be removed and various term-weighting schemes applied.
2. Using Gibbs sampling and the LDA model, compute $P(\mathbf{z}|\mathbf{w})$.
3. Estimate the per-document topic distributions and per-topic word distributions.
4. Using SOM, cluster the documents based on the per-document topic distributions.
5. Cluster the nodes of the SOM using k-means. Optimal k can be found by minimizing the value of the Davies-Bouldin index for values of k from 1 to the square root of the number of processing elements in the SOM.
6. Display results.

Of the related document clustering methods, LDA-SOM is closest to LSISOM, WEBSOM and PLSV. However, LDA-SOM differs from each of these in important ways. For instance, both LSISOM and WEBSOM utilize a dimensionality reduction process and two self-organizing maps to cluster documents. LDA-SOM, on the other hand, uses a single map. In all three cases, transformed document representations (context vectors in the case of LSISOM or WEBSOM, topic distributions for LDA-SOM) are ultimately clustered on a self-organizing map. This superficial similarity masks a fundamental difference in approach. The context vectors used by LSISOM and WEBSOM effectively collapse synonyms onto one another, reducing clustering errors due to differences in phrasing. However, they do not capture the topical content of a document. LDA-SOM on the other hand, explicitly computes what a document is about and clusters documents based on those topical representations.

LDA-SOM has more in common with PLSV. Both methods cluster documents based on topic models and render the results on a two-dimensional display without the clutter issues often found in traditional methods such as multi-dimensional scaling or local linear embedding. The first difference between LDA-SOM and PLSV is that the latter treats visualization coordinates as latent variables that are part of the document generation process. LDA-SOM does not include visualization as an explicit part of the document model, generating display coordinates as part of the clustering process instead.

The second difference between LDA-SOM and PLSV is that PLSV is not a clustering algorithm. It does a fine job of arranging documents on the plane, but a secondary clustering process must be applied in the absence of class labels. LDA-SOM accomplishes both the clustering and visualization tasks.

Experimental Evaluation

Evaluating the effectiveness of a clustering algorithm can be difficult and subjective at best (Feldman and Sanger 2007). However, there are some well known performance metrics for self-organizing maps. These are the quantization error and topological error.

The quantization error of an SOM is the average distance between an input vector and its best matching prototype vector, taken over the entire data set. Quantization error provides a measure of how well the map represents the training data. Small quantization errors indicate the map matches the input, while large errors indicate the map has failed to learn the training data.

Topological error is the proportion of input samples for which the first and second best matching prototypes are not neighboring map nodes. It provides a measure of how well-ordered the map is. A large topological error indicates the map is "twisted" and has failed to preserve the topology of the original high-dimensional input data. In order to achieve a meaningful visualization, the topological error must be minimized.

In addition to SOM quality metrics, we also measure the performance of LDA-SOM by the minimal Davies-Bouldin index achieved during the clustering process. The Davies-Bouldin index is a function of the ratio of within-cluster scatter to between-cluster distance. Larger values indicate a better clustering.

We applied LDA-SOM to the 20 Newsgroups (Lang 2008) and Neural Information Processing Systems (NIPS) (Chechik 2008) data sets in order to evaluate its effectiveness. The 20 Newsgroups data consists of postings from 20 Usenet newsgroups. It contains a total of 11,269 documents with 53,975 distinct words. The NIPS data set contains 1500 documents with 12,419 distinct words collected from the NIPS conference proceedings between the years 1987 and 1999.

The number of topics for LDA-SOM was fixed at 50. The LDA hyperparameters were set at $\alpha = 1.0$ and $\beta = 0.01$ as recommended by Griffiths and Steyvers.

Each data set had stop-words and words occurring fewer than 10 times removed. No word stemming was performed. Initial word histograms were based on simple word frequencies; inverse document frequency weighting was not applied. The LDA Gibbs sampler was run for 300 iterations before sampling to provide enough time for the topic distributions to stabilize.

LDA-SOM was run a total of 10 times for each data set and average error metrics were calculated. Quantization and topological error were calculated after clustering based on the converged LDA-SOM. The Davies-Bouldin index was calculated during the clustering process and is the minimum found for several values of k. Figure 2 tabulates the results of these experiments. For each metric, the mean and standard deviation are reported.

Both data sets show good behavior with respect to the target error metrics. The standard deviation is small across all data points, indicating stability in the clustering process. The 20 Newsgroups data exhibits a slightly high Davies-Bouldin index, probably because more than 20 clusters were

Data Set	Quantization Error	Topological Error	Davies-Bouldin Index
20 Newsgroups	0.0720 ± 0.0011	0.1127 ± 0.0252	0.8728 ± 0.0457
NIPS	1.2733 ± 0.0302	0.4541 ± 0.0022	0.0687 ± 0.0073

Figure 2: Mean error ± one standard deviation for the 20 Newsgroups and NIPS data sets. Quantization and topological error were calculated from the converged LDA-SOM. Davies-Bouldin index was calculated as part of the LDA-SOM clustering process.

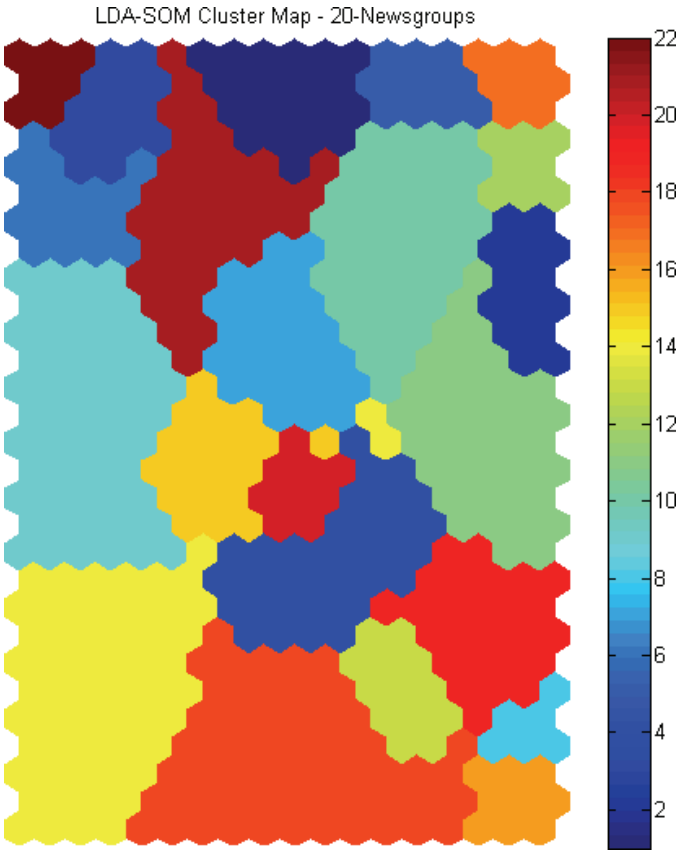

Figure 3: LDA-SOM cluster map for the 20 Newsgroups data set showing 22 topical clusters.

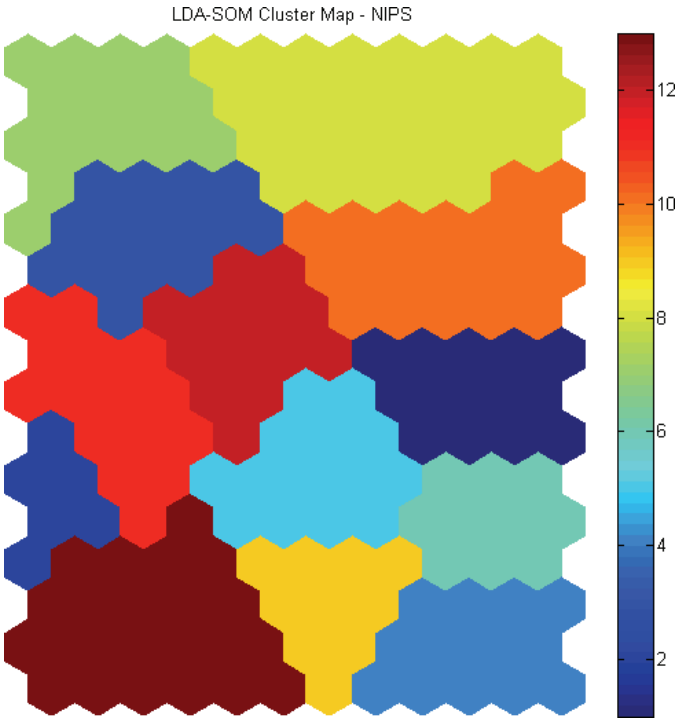

Figure 4: LDA-SOM cluster map for the NIPS data set showing 13 topical clusters.

identified. Regardless, the index is similar to that reported in (Tatti 2007). Quantization and topological error are excellent. The NIPS data shows slightly higher quantization and topological errors and a much better Davies-Bouldin index.

Figures 3 and 4 show the LDA-SOM cluster maps for the 20 Newsgroups and NIPS data sets respectively. For the newsgroup data, 22 distinct topical clusters were found. A mismatch in the number of discovered clusters and the number of newsgroups is not unexpected since Usenet discussions frequently drift off topic. The map does not exhibit any twisting or scattering of clusters.

The NIPS cluster map indicates 13 distinct clusters were found in the collection. Interestingly, there are 13 major topic areas for the NIPS conference. It is tempting to assume each cluster corresponds directly to a NIPS topic area; however, doing so is unwarranted. The LDA-SOM process clusters documents based on topic distributions discovered in the document content. Nothing guarantees that the inferred topics line up with the conference topic areas. Additionally, for most documents, these topic distributions are multi-modal. Consequently, each cluster in the LDA-SOM cluster map represents a mixture of two or more topics. This is especially true near cluster boundaries, where topic distributions begin to change. Cases of extreme topical mismatch are represented as a twisting of the map, i.e., a scattering of cluster nodes across non-contiguous areas. In addition, the topological error of a twisted map is extremely high.

Conclusion and Future Work

This article presents a document clustering and visualization method based on Latent Dirichlet Allocation and Self-Organizing Maps. The method transforms the word histogram representations of documents into topic distributions, reducing dimensionality in the process. Once the topic distributions for each document have been inferred, the documents are clustered using a self-organizing map. To aid in visualization, the resulting map is clustered using k-means, giving a colorized cluster map. The topological preservation properties of SOMs ensure that documents with similar document distributions are place near one another on the map and that neighboring clusters are more similar than clusters located far apart. This provides an intuitive means of browsing and navigating a document collection.

We have applied this method to two real-world data sets with good results. However, there are some challenges in applying this method. Most significantly, reasonable values for the LDA hyperparameters and the number of topics must be selected.

Second, we have processed our SOMs using the default Euclidean distance measure. Other measures more suitable for probability distributions such as the symmetrized Kullback-Liebeler divergence are perhaps more appropriate. We intend to address this issue in future work.

Acknowledgements

This work is supported by the Joint IED Defeat Organization (JIEDDO). The views expressed in this paper are those of the authors and do not represent the views or policies of the United States Air Force, Department of Defense, or Joint IED Defeat Organization.

References

Ampazis, N., and Perantonis, S. J. 2004. LSISOM – a latent semantic indexing approach to self-organizing maps of document collections. *Neural Processing Letters* (19):157–173.

Bauer, H.-U., and Pawelzik, K. R. 1992. Quantifying the neighborhood preservation of self-organizing feature maps. *IEEE Transactions on Neural Networks* 3(4).

Blei, D. M.; Ng, A. Y.; and Jordan, M. I. 2003. Latent dirichlet allocation. *Journal of Machine Learning Research* 3:993–1022.

Chechik, G. 2008. Nips 1–17 data. Online. http://ai.stanford.edu/ gal.

Davies, D. L., and Bouldin, D. 1979. A cluster separation measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 1:224–227.

Deerwester, S. C.; Dumais, S. T.; Landauer, T. K.; Furnas, G. W.; and Harshman, R. A. 1990. Indexing by latent semantic analysis. *Journal of the American Society of Information Science* 41(6):391–407.

Feldman, R., and Sanger, J. 2007. *The Text Mining Handbook*. Cambridge University Press.

Golub, G. H., and Van Loan, C. F. 1996. *Matrix Computations*. The Johns Hopkins University Press, third edition.

Griffiths, T. L., and Steyvers, M. 2004. Finding scientific topics. In *Proceedings of the National Academy of Sciences*, 5228–5235.

Hofmann, T. 1999. Probabilistic latent semantic indexing. In *Proceedings of the 22nd International Conference on Research and Development in Information Retrieval (SIGIR'99)*, 50–57.

Iwata, T.; Yamada, T.; and Ueda, N. 2008. Probabilistic latent semantic visualization: Topic model for visualizing documents. In *KDD '08: Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 363–371.

Kaski, S.; Honkela, T.; Lagus, K.; and Kohonen, T. 1996. Creating an order in digital libraries with self-organizing maps. In *Proceedings of WCNN'96, World Congress on Neural Networks*, 814–817.

Kaski, S. 1998. Dimensionality reduction by random mapping: Fast similarity computation for clustering. In *Proceedings of IJCNN'98, International Joint Conference on Neural Networks*, volume 1. Piscataway, NJ: IEEE Service Center. 413–418.

Kohonen, T. 2001. *Self-Organizing Maps*. Springer Series in Information Sciences. New York: Springer, 3rd edition.

Lagus, K.; Honkela, T.; Kaski, S.; and Kohonen, T. 1999. WEBSOM for textual data mining. *Aritificial Intelligence Review* 13(5/6):345–364.

Lagus, K.; Kaski, S.; and Kohonen, T. 2004. Mining massive document collections by the WEBSOM method. *Information Sciences* 163(1-3):135–156.

Lang, K. 2008. 20 newsgroups. Online. http://people.csail.mit.edu/jrennie/20Newsgroups/.

Salton, G., and McGill, M. J. 1983. *Introduction to Modern Information Retrieval*. New York: McGraw-Hill Book Company.

Steyvers, M., and Griffiths, T. L. 2007. Probabilistic topic models. In Landauer, T. K.; McNamara, D. S.; Dennis, S.; and Kintsch, W., eds., *Handbook of Latent Semantic Analysis*. Laurence Erlbaum Associates. 427–448.

Steyvers, M.; Smyth, P.; Rosen-Zvi, M.; and Griffiths, T. 2004. Probabilistic author-topic models for information discovery. In *Proceedings of the 10th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (SIGKDD'04)*.

Tatti, N. 2007. Distances between data sets based on summary statistics. *Journal of Machine Learning Reasearch* 8:131–154.

Wei, X., and Croft, W. B. 2006. LDA-based document models for ad-hoc retrieval. In *SIGIR '06: Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 178–185.