LDAExplore: Visualizing Topic Models Generated Using Latent Dirichlet Allocation

Ashwinkumar Ganesan,
Kiante Brantley
Dept. Of Computer Science &
Electrical Engineering,
University Of Maryland
Baltimore County
{gashwin1, bran4}@umbc.edu

Shimei Pan
Dept. Of Information Systems,
University Of Maryland
Baltimore County
shimei@umbc.edu

Jian Chen
Dept. Of Computer Science &
Electrical Engineering,
University Of Maryland
Baltimore County
jichen@umbc.edu

ABSTRACT

We present *LDAExplore*, a tool to visualize *topic* distributions in a given document corpus that are generated using *Topic Modeling* methods. Latent Dirichlet Allocation (LDA) is one of the basic methods that is predominantly used to generate *topics*. One of the problems with methods like LDA is that users who apply them may not understand the *topics* that are generated. Also, users may find it difficult to search correlated topics and correlated documents. *LDAExplore*, tries to alleviate these problems by visualizing topic and word distributions generated from the document corpus and allowing the user to interact with them. The system is designed for users, who have minimal knowledge of *LDA* or *Topic Modelling* methods. To evaluate our design, we run a pilot study which uses the *abstract*'s of 322 *Information Visualization* papers, where every abstract is considered a *document*. The topics generated are then explored by users. The results show that users are able to find correlated documents and group them based on topics that are similar.

Author Keywords

Topic Modeling, visual analytics, Latent Dirichlet Allocation, statistical analysis, Parallel Coordinates, text analytics.

ACM Classification Keywords

H.5.m. Information Interfaces and Presentation (e.g. HCI): Miscellaneous

INTRODUCTION

Topic Modeling tries to automate the process of extracting topics from documents while also annotate them with semantic information [3]. It is made up of a set of statistical algorithms that extract correlated words from documents. These extracted word sets are called *Topics*. Later, users can annotate the topic with semantic information. For example, consider a word set extracted which contains words such as *visualization, sets, clusters, infoviz, interfaces* from the document corpus. Knowing the words associated with this *topic*, we have a general notion that the word set represents the topic *Information Visualization*. The "topic" generated is the word set and *Information Visualization* is the semantic "topic name" annotated by the user. Latent Dirichlet Allocation (LDA) [4] is one of the common methods to perform topic modeling on a given corpus of documents.

LDA generates two types of distributions i.e. the topic distribution for each document in the set and the word distribution for each topic. These distributions can be changed by tweaking the hyper-parameters. *LDAExplore*, tries to give visual cues about how these distributions look, and how the topics and documents are interrelated at the corpus level, between groups or individual documents. It is designed for users who may not know what topic modeling algorithms do. They are concerned with understanding the document corpus, finding the hidden topics in the corpus. LDAExplore gives the users the ability to search for documents using keywords from topics.

One of the basic requirements of the design is that visual should scale for a large set of documents while providing the ability to see individual and group relations. In comparison to the set of documents, we assume that the number of topics is a much smaller set.

One of primary problems with topic modeling methods is that the "topic" generated by them, may not be clearly understood by the user [8]. One of the solutions to this problem, is to introduce a human-in-the-loop paradigm where users can interact with the algorithm, providing feedback such that the underlying model can be modified to generate "better" topics.

Our contributions in this paper are:

- Creating a comprehensive set of tasks required to design visualizations for text analysis using LDA.
- A visualization that show correlations between topics and documents. The design can be scaled up to a larger set of

documents without applying any aggregation method such as clustering to the documents.

- Combining visual search & filtrering to allow users to filter parallely across multiple topics in our design, so that users can easily filter a large document corpus.
- A pilot study which shows the usability of each part of the tool.

RELATED WORK

There are many different ways to visualize Topic Models. The methods include the use of force-directed graphs, parallel coordinates, matrices & tree designs.

Using Matrix Designs

Matrix or tabular designs are easily understood by users. *Termite* uses a tabular layout to promote comparison of terms both within and across latent topics. The primary visualization design used in it, is a matrix view where rows correspond to terms and columns to topics[2]. Another visualization [7] uses a navigator to explore LDA generated topics. It shows the words in each topic and uses a tabular form to show which documents are associated. Once a specific document is selected the use can navigate to its Wikipedia article. These matrix layouts, show the correlation between terms and topics but have a difficulty scaling to a large number of documents. They do not show how the topic is correlated to the document, what its likelihood is.

Using Trees & Hierarchial Clustering

RoseRiver is another analytics system used to visualize how topics evolve [10]. The system uses a tree cut approach with a combination of a word cloud. The word cloud is a standard method to show a word set with word sizes varying according to their frequency (or probability). Similarly, *Overview* [6] is a technique designed for analyzing large text corpora using TF-IDF. It generates document clusters by hierarchically clustering these distances and encoding the result as a topic tree. *VarifocalReader* [15] uses hierarchy structure such that it shows chapters, sub-chapters, and pages , and then lines of text. It again uses TF-IDF as one of its techniques. *Serendip* [1] is a technique designed for analyzing large text corpora using LDA. They present the results in a rectangle where each row means a document and each column means a topic.

Tree layouts can scale in design, but when combined with hierarchial clustering increase the amount of preprocessing that is necessary to present the tree structure. We choose not perform any clustering because user defined clusters and annotations, change the nature of the distributions generated by LDA.

Using Force Directed Graphs & Parallel Coordinates

UTOPIAN [9], uses a force-directed graph to represent topics and is a semi-supervised system. It uses non-negative matrix factorization. Some other visualization include *iVisClustering* [16] and *ParallelTopics* [11]. Another parallel coordinate design which works with words in the topic directly is *ThemeDelta*[12]. Both have a design which uses parallel coordinates and *iVisClustering* uses a force directed graph approach. Topic-based, interactive visual analysis tool (TIARA) [17] shows topic distributions across documents across time. Force-directed graphs are advantageous in terms of easy understanding & usability, but they consume a lot of visual / screen space and may not scale for large number of terms or documents.

There are number of challenges *LDAExplore* tries address i.e. a method that can scale for an increasing set of documents, a way to quantify the relation between documents and topics and provide and easy way to filter documents & topics, while preserving their interrelations. In *LDAExplore*, we try to reduce the amount of preprocessing necessary to display the topics, so that changes to the underlying model can be performed efficiently. We use parallel coordinates for showing document and topic relations while integrating keyword search to make it easy for the user to search for patterns using words from topics. This makes the filtering process more intuitive for the user.

DESIGN CONSIDERATIONS

Task Analysis

As described in the previous section, results got from LDA, can be unintuitive. One of the main purpose of the visualization is to provide users with the options to explore the document set and give them the ability to provide feedback about topics to the system and which topics are correlated to documents, so that word and topic distributions are more intuitive and insightful. Following are the set of tasks which form the basis for our design.

Visualize Topics

LDA generates a set of topics, each having its own individual word list. Each word has a probability of being associated with the topic. As shown in figure 3, when the user interacts with a specific topic, the top words in the topic will be displayed, showing their probability. This is one way for the user to identify, what the topic is.

Overview of Document - Topic Relations

Once the user has a generic idea of the semantics of the topics, the user can then focus on the correlations between documents and topics. The main purpose is to be able to see which topics are ranked higher over a large subset of documents, while which topics have a lower rank. The number of documents that rank a topic high, gives the user an idea as to which topics are more important.

Remove Topics from the Visual

When the user knows which topics are important, excess topics from the visual can be removed.

Filtering Documents

When the size of the corpus is large, the user will try to understand the distribution within a subset of the documents. Hence the user can filter the documents based on various criteria:

1. By top words in each document that associated with the topics having the highest probability of being related to the document.

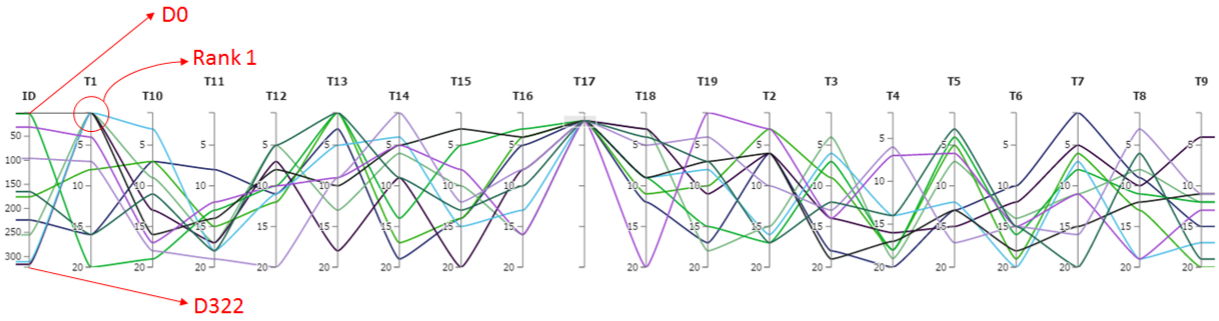


Figure 1: **Displaying the Topic Distribution**

2. The user may not know the documents, but may be able to identify some based on the *name* or *title* of the document. Thus, the user can filter out individual documents.

Perform set operations
While working with a groups of documents, the user requires the ability to perform set operations such as *Include* and *Exclude*. *Include* gives the user, the ability to add a filtered set of documents to future filters. *Exclude* does the exact opposite, which is to remove the document set from any filter operations in the future. Once the document corpus has been explored, the user can export the filtered data for any post processing that is performed separately.

Show & Cluster Similar Topics
Once the word distributions for each topic are known, the user will look for topics that are similar. Topic similarity can be measured using methods such KL-Divergence [20]. Similar topics can be grouped together, so that the number of topics on the visual display can be reduced and hierarchical collections of topics can be formed.

Perform Cluster Operations
Similar topics, will lead to a new group of documents which have these topics. The main task for the visual is to enable the user to define groups of documents and topics based on their own knowledge.

Annotating Topics
User should have the option to annotate the topics and documents or the respective clusters they lie in.

Prototype Design

As described before, various sets, groups and LDA visualizations are available. They are able to display a very basic summarization of the document corpus and reveal basic relationships that occur. They provide the ability to explore individual documents, but they lack the ability to show document correlations across the dataset. When the correlations are shown, the design may not scale for a large number of documents. *LDAExplore* creates a design to handle a sizeable number of documents while providing the capability to filter and visualize topics for individual documents.

In its present form, *LDAExplore* implements a subset of the tasks:

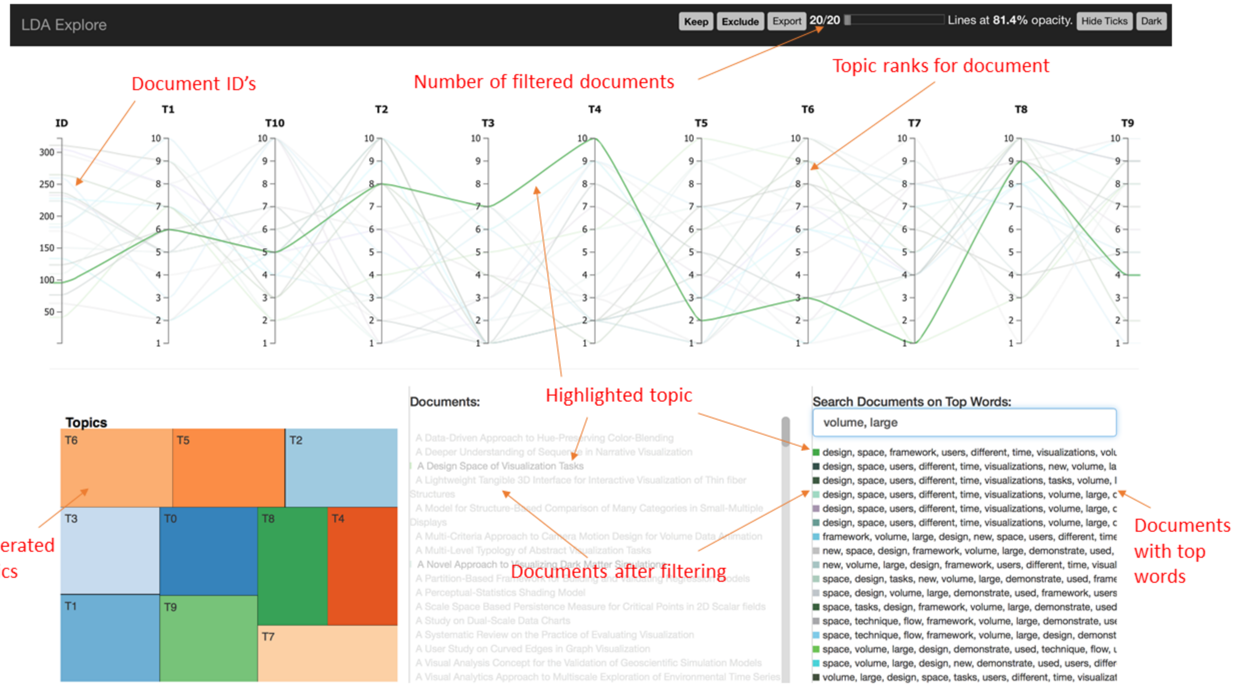
- Visualizing topics
- Filtering & Set operations
- Displaying the overview of documents and topics correlations
- Removing topics.

Based on the set of tasks defined, we describe how *LDAExplore* works.

Visualizing Topics
We use a *treemap* to visualize the topics. Figure 2, shows how the topics are displayed. Each rectangle is a topic and is given an ID like $T_1 ... T_n$ where there are n defined topics. In figure 1, the total number of topics are 20. In the treemap, the size of the rectangle defines the probability of the topic being associated to the document. In the current design, the likelihood is defined for each individual topic with respect to each document, rather than the document collection as a whole. Hence in figure 2 the topics are represented by rectangles of equal area (but possibly different dimensions, depending on the number of topics) showing that all topics have an equal likelihood.

When the user clicks on a specific topic in the *treemap*, the top 10 words associated with the topic are revealed. Figure 3, shows an example word distribution. The word with the highest probability has the largest area. In this example, the word *data* has the largest probability for topic *T4*. In the current design, the number of words displayed per topic is maintained constant at 10. The user can traverse back to the *Topics*, by clicking on the *Topics* tab at the top of the treemap. Whenever the user drills down, into the word distribution, the parent is always known.

Relating Topics & Documents
We use parallel coordinates [14][5] to show the correlation between topics and documents. The parallel coordinates have two types of axes. The first one represents the document. The second type of axis is for topics. For example, consider Figure 1, which shows a total of 21 axes. The first axis *ID* shows the all the documents in the corpus and ranges from document ID D_0 to D_{322}. The next 20 axes each represents a topic. Once LDA generates the probabilities of each topic in a document, we rank the topics. The topic axes shows the



Documents:

A Data-Driven Approach to Hue-Preserving Color-Blending
A Deeper Understanding of Sequence in Narrative Visualization
A Design Space of Visualization Tasks
A Lightweight Tangible 3D Interface for Interactive Visualization of Thin fiber Structures
A Model for Structure-Based Comparison of Many Categories in Small-Multiple Displays
A Multi-Criteria Approach to Camera Motion Design for Volume Data Animation
A Multi-Level Typology of Abstract Visualization Tasks
A Novel Approach to Visualizing Dark Matter Simulations
A Partition-Based Framework for Building and Validating Regression Models
A Perceptual-Statistics Shading Model
A Scale Space Based Persistence Measure for Critical Points in 2D Scalar fields
A Study on Dual-Scale Data Charts
A Systematic Review on the Practice of Evaluating Visualization
A User Study on Curved Edges in Graph Visualization
A Visual Analysis Concept for the Validation of Geoscientific Simulation Models
A Visual Analytics Approach to Multiscale Exploration of Environmental Time Series

Search Documents on Top Words:

volume, large

- 🟩 design, space, framework, users, different, time, visualizations, volu
- ■ design, space, users, different, time, visualizations, new, volume, la
- ■ design, space, users, different, time, visualizations, tasks, volume, l
- ■ design, space, users, different, time, visualizations, volume, large, c
- ■ design, space, users, different, time, visualizations, volume, large, c
- ■ design, space, users, different, time, visualizations, volume, large, c
- ■ framework, volume, large, design, new, space, users, different, time
- ■ new, space, design, framework, volume, large, demonstrate, used,
- ■ new, volume, large, design, framework, users, different, time, visual
- ■ space, design, tasks, new, volume, large, demonstrate, used, frame
- ■ space, design, volume, large, demonstrate, used, framework, users
- ■ space, tasks, design, framework, volume, large, demonstrate, used
- ■ space, technique, flow, framework, volume, large, demonstrate, use
- ■ space, technique, flow, framework, volume, large, design, demonst
- 🟩 space, volume, large, design, demonstrate, used, technique, flow, u
- ■ space, volume, large, design, new, demonstrate, used, users, differ
- ■ volume, large, design, space, tasks, users, different, time, visualizat

Figure 2: ***LDAExplore*: Filtering Mechanisms**
Various sections of *LDAExplore* are shown. The parallel coordinates shows the topic distribution, the treemap (which shows the word distribution for each topic on drill down), search tab used for filtering. A single filter is applied across the axis T17 to isolate documents ranked having T17 at a higher rank.

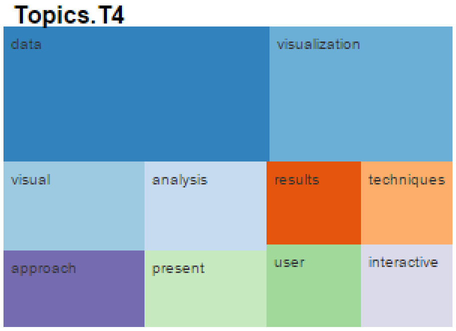


Figure 3: **Top 10 words in Each Topic**

rank of the topic with respect to the document. So in figure 1, document D_0 has Topic T_1 at rank 1, while others have it at rank $4, 7, 16\&20$ respectively. Each document has a different color, so that the user can distinguish between the graphs of various documents. The ranks are ordered in ascending order with 1 showing the highest degree of correlation and n (where n is the number of topics), showing 0 or the least correlation. Topic ranks give the user which topic is the most correlated and which is the least.

In *LDAExplore*, we decided to use parallel coordinates in order to be able to display the maximum amount of topics without the screen being too cluttered. When we tried to design other types of visualization, we quickly realized that scalability is an issue (by doing a quick claim analysis). In addition to parallel coordinate, we have a treemap to show the word in each topic. Being able see to the words in each topic is important because it allows the user to easily understand what that topic is about.

Filtering

LDAExplore provides a range of filtering options. Figure 2 provides a detailed overview of the filtering features. There are three main types of filtering:

1. *Filter by Range* - Each axis can be used to filter the documents by selecting a specific range on the axis. The parallel coordinates displays the curves, only for the selected documents. Each axis can filtered simultaneously, thus creating a *filter-chain*.

2. *Filter by Searching* - A search bar is provided to the user, to search the documents using top n words associated with each document. The current design, allows the user to search using any of the top 10 words.

3. *Filter by Selecting Individual Documents* - The documents column, lists each document in the corpus by its title. The user can click on a specific document to filter it out. Whenever a document is filtered out, its color is changed to downplay the document.

Once the user has filtered the documents, the user can highlight a specific document within the filtered document set by using the mouse to hover over the document - word set (column 3 in figure 2). The total number of document in the filtered set is given at the top on the navigation bar. Also, the user can *export* the filtered data to CSV format.

Set Operations

There are 2 main set operations that are supported in *LDA-Explore* i.e. include and exclude. The *include* operation is performed by the *Keep* and exclusion using *Exclude* as seen in figure 2.

Generating Top Words

Once the documents are processed using LDA (from the gensim library [19]), we end up with a list of topics and the probability of a word being associated with the topic. There is a minimum probability threshold for each topic so that it can be considered in the final distribution. Topics below the minimum threshold are assigned a probability of 0. To generate the top words for each document we implement the following:

$$P(w_x, d_y) = \sum_{t=1}^{m} \sum_{i=1}^{n} P(w_x, t_i) \times P(t_i, d_y) \qquad (1)$$

where $P(w_x, d_y)$ is the probability of a word with respect to a document, $P(w_x, t_i)$ is the probability of a word with respect to a topic and $P(t_i, d_y)$ is the probability of topic with respect to a document.

Equation 1 calculates the probability of each word in every topic for a specific document and then finds the top k words which have the highest probability of being associated with the given document. This set of words generated are shown as the top n searchable words on the visualization. Also, the words are used to create the topic-word hierarchy that is used in the treemap visualization for topics.

Ranking Vs Probability

The parallel coordinates can be used to display topic-document relations based on the rank of the topic (rank 1 being highest probable topic) or based on the actual probability of the topic. The current design displays the rank as it is easier to interpret by the user. As the difference probabilities between topics in a given distribution for a document reduce, it becomes harder for the users to define which topic are more important across a set of documents than the others on the parallel coordinates scale.

Eliminate High Frequency Words

The list of top words for each document and topic, contain many words which are common to most documents such as *data* and *visualization* as seen in figure 2. To alleviate this problem, Term Frequency - Inverse Document Frequency (*TF-IDF*) [18] is used, albeit not directly in combination with Latent Dirichlit Allocation (LDA). *TF-IDF* is used to clear terms which are in high frequency across a large subset of the documents. It is executed with an arbitrary threshold of t, which describes the ratio of how many documents in the total corpus, can the word be present. Words over the threshold are from the final list of tokens before performing LDA.

PILOT STUDY

Study Details

The Pilot study is an initial study designed to gauge if the user can perform the tasks outlined, how usable the tool really is and get early feedback about the tool. It has been conducted using 5 external participants. The responses from the first 3 participants have been used to make some changes to the questions and then further feedback has been collected from the remaining 2 participants. The number of participants having prior knowledge about *Topic Modeling* and how LDA works is 2. For the purpose of the study, the users have been asked to work with the tool on a standard wide screen monitor. The study has been partitioned into 4 parts i.e. the overview, topic visualization, filtering and keyword searching.

The questions in the study are of 4 main types, Tasks execution questions (such as *How many documents are represented in this visual?*), Understanding questions (such as *Does the visual have too many things on it?*), Reasoning questions (such as *Which is the least important topic? (Important means highly ranked topic)*) and Usability questions (such as *Are the rankings on the axis visible / readable?*).

We provide details on how to access the tool, a brief overview of the features of the tool with the help of an annotated example diagram (such as figure 2). Also, the users are given information about the data set, they will be visualizing. For the purpose of our study, the data set consists of the abstracts from 322 research papers from the domain of *Data Visualization*. Each abstract is considered as a separate document. The information contains attributes like the title of the paper and the abstract of the research paper. LDA is used to identify the plausible underlying topic / topics for each research paper. The *Title* of paper that is seen in the “Documents” column and the abstract of each research paper. The next section describes each part of the study and the associated results.

Survey & Results

Overview

The overview questions are designed to understand if participants can understand any useful information immediately after loading the main page and check what they think about the overall look and feel of the tool. The main page shows all the data without any kind of filtering.

The users find the initially displayed parallel coordinate to be slightly cluttered, when it is presented without any filtering. Also, the users find it difficult to search for markings on the axis in the parallel coordinate such as axis legends. The overall information on the number of topics and the number of documents represented in the visual can be deduced easily.

Visualizing Topics

The *topics* questions are designed to see if the participants can understand what kind of topics are generated by LDA and what are the words (with their individual probabilities) associated with the topic. This section is very important because while inspecting LDA results, users are likely to investigate what topics are produced from a corpus of documents.

The users are able to understand which topics have which words and which of them are important within these topics. The results confirm that the treemap is useful in visualizing topics and that the areas of rectangles used to define which topic / word is more or less important (larger size being more important) is useful.

Filtering

Filtering questions are designed to see if users can use all the features within *LDAExplore* to effectively filter documents to find specific patterns. Users are tasked with filtering a specific topic's axis on the parallel coordinate for a range of ranks and isolate the subset of documents which find the topic to be important. Users are then asked to find topics that might be similar and documents (research papers) that can be grouped together. The study shows users are able to filter the document corpus easily. Some users find it difficult to deduce which topic is more or less important. This because the parallel coordinates design has a single line for each document and does not *bundle* lines together at every rank of each topic, thus giving the user an idea of the number of documents at each rank.

Keyword Search

The last set of questions are on the usefulness of *keyword* search. Users are able to see the effect of *filtering using keywords* on other sections (documents & search) of the tool. Searching is one of more useful features of the tool because it instantaneously reduces the number of lines in the parallel coordinate and provides immediate information about documents based on the user's search query.

Recommended Changes to the Design

Based on the information gathered from the study, we have a list of changes that can be done. To reduce the clutter in the parallel coordinates (for large corpora), edge bundling can be introduced [13]. To de-clutter it further, documents can be clustered to visualize the topics that are ranked for these groups. The topics themselves can be grouped together based upon their similarity. The navigation between topics in the treemap can be improved.

Users in the study find it diffcult to understand the "Documents" column. The column has titles of documents and acts as filter to remove a document from the parallel coordinate rather than selectively show it. This is counter intuitive to how humans use *URL's/ clicks* on an item, which is mainly to select a page or the highlighted item. Hence this filter can be removed or altered to select a specific document and filter out the rest of the corpus (or highlight the specific document with respect to others).

The treemap (topics) and parallel coordinate(topic distribution) are disconnected. This can be modified, so that the treemap can be used as additional filtering where the documents across the parallel coordinates can be filtered using topic words. To find out which topics have similar probability in the topic distribution for a given document, user can be given the option to switch between the rank mode and probabilistic modes in the visualization.

Searching using words is useful, but there is not mechanism to search the corpus for a specific document using attributes such as the title. Searching can be improved by adding the title of the document to the *keywords search* section.

Once the documents have been filtered users need a way to *save* the selection, so that it can be utilized at a later time. This *in-memory* storage of the filter applied can be useful.

CONCLUSION

Being able to understand LDA results through a visualization is a critical part of helping end users to explore documents and recognize patterns. This paper describes *LDAExplore*, a tool to visualize a document corpus. It provides the results of a pilot study that shows that the targeted audience have a better understanding of what the corpus contains & what LDA results mean.

We have worked closely with 5 individuals, 2 of whom are able to understand LDA. They are able to navigate through the visualization with minimal guidance in order to solve the tasks such as finding documents which are related to specific topics that are difficult to do without a visualization.

In the future we plan to investigate how to enable users to cluster topics if they decide that the topics selected are not distinct enough. Similarly, users will like to cluster documents whose topics look similar and see how these aggregated clusters differ. User defined clustering, changes the underlying model within LDA. It will be interesting to see how the changes in model affect other topics or documents.

ACKNOWLEDGMENTS

We would like to thank Li-Chien Lee for the help on this paper and all the participants of the study, who spent a portion of their valuable time evaluating our project.

This work was supported in part by grants from NSF IIS-1302755, IIS-1341254, DBI-1260795, DBI-1341352, and NIST 70NANB13H181. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation. Certain commercial products are identified in this paper in order to specify the experimental procedure adequately. Such identification is not intended to imply recommendation or endorsement by the National Institute of Standards and Technology, nor is it intended to imply that the products identified are necessarily the best available for the purpose.

REFERENCES

1. Alexander, E., Kohlmann, J., Valenza, R., and Gleicher, M. Serendip: Turning topics back to the text. *IEEE Visualization Poster Proceedings*.
2. Bertin, J. Semiology of graphics: diagrams, networks, maps.
3. Blei, D. M. Probabilistic topic models. *Communications of the ACM 55*, 4 (2012), 77–84.
4. Blei, D. M., Ng, A. Y., and Jordan, M. I. Latent dirichlet allocation. *the Journal of machine Learning research 3* (2003), 993–1022.

5. Bostock, M. Data-driven documents (d3. js), a visualization framework for internet browsers running javascript, 2012.
6. Brehmer, M., Ingram, S., Stray, J., and Munzner, T. Overview: The design, adoption, and analysis of a visual document mining tool for investigative journalists.
7. Chaney, A. J.-B., and Blei, D. M. Visualizing topic models. In *ICWSM* (2012).
8. Chang, J., Gerrish, S., Wang, C., Boyd-graber, J. L., and Blei, D. M. Reading tea leaves: How humans interpret topic models. In *Advances in neural information processing systems* (2009), 288–296.
9. Choo, J., Lee, C., Reddy, C. K., and Park, H. Utopian: User-driven topic modeling based on interactive nonnegative matrix factorization. *Visualization and Computer Graphics, IEEE Transactions on 19*, 12 (2013), 1992–2001.
10. Cui, W., Liu, S., Wu, Z., and Wei, H. How hierarchical topics evolve in large text corpora.
11. Dou, W., Wang, X., Chang, R., and Ribarsky, W. Paralleltopics: A probabilistic approach to exploring document collections. In *Visual Analytics Science and Technology (VAST), 2011 IEEE Conference on*, IEEE (2011), 231–240.
12. Gad, S., Javed, W., Ghani, S., Elmqvist, N., Ewing, T., Hampton, K., and Ramakrishnan, N. Themedelta: Dynamic segmentations over temporal topic models.
13. Holten, D. Hierarchical edge bundles: Visualization of adjacency relations in hierarchical data. *Visualization and Computer Graphics, IEEE Transactions on 12*, 5 (2006), 741–748.
14. Inselberg, A., and Dimsdale, B. Parallel coordinates. In *Human-Machine Interactive Systems*. Springer, 1991, 199–233.
15. Koch, S., John, M., Worner, M., Müller, A., and Ertl, T. Varifocalreader–in-depth visual analysis of large text documents.
16. Lee, H., Kihm, J., Choo, J., Stasko, J., and Park, H. ivisclustering: An interactive visual document clustering via topic modeling. In *Computer Graphics Forum*, vol. 31, Wiley Online Library (2012), 1155–1164.
17. Pan, S., Zhou, M. X., Song, Y., Qian, W., Wang, F., and Liu, S. Optimizing temporal topic segmentation for intelligent text visualization. In *Proceedings of the 2013 international conference on Intelligent user interfaces*, ACM (2013), 339–350.
18. Rajaraman, A., and Ullman, J. D. *Mining of massive datasets*. Cambridge University Press, 2011.
19. Rehrek, R., Sojka, P., et al. Gensimstatistical semantics in python.
20. Yang, Y., Pan, S., Downey, D., and Zhang, K. Active learning with constrained topic model.