DOI: 10.1111/j.1467-8659.2012.03108.x
Eurographics Conference on Visualization (EuroVis) 2012
S. Bruckner, S. Miksch, and H. Pfister
(Guest Editors)
Volume 31 (2012), Number 3

iVisClustering: An Interactive Visual Document Clustering via Topic Modeling

Hanseung Lee[1], Jaeyeon Kihm[2], Jaegul Choo[1], John Stasko[1], and Haesun Park[1]

[1]Georgia Institute of Technology, USA
hanseung.lee@gatech.edu, {jaegul.choo, stasko, hpark}@cc.gatech.edu
[2]Cornell University, USA
jk2443@cornell.edu

Abstract

Clustering plays an important role in many large-scale data analyses providing users with an overall understanding of their data. Nonetheless, clustering is not an easy task due to noisy features and outliers existing in the data, and thus the clustering results obtained from automatic algorithms often do not make clear sense. To remedy this problem, automatic clustering should be complemented with interactive visualization strategies. This paper proposes an interactive visual analytics system for document clustering, called iVisClustering, based on a widely-used topic modeling method, latent Dirichlet allocation (LDA). iVisClustering provides a summary of each cluster in terms of its most representative keywords and visualizes soft clustering results in parallel coordinates. The main view of the system provides a 2D plot that visualizes cluster similarities and the relation among data items with a graph-based representation. iVisClustering provides several other views, which contain useful interaction methods. With help of these visualization modules, we can interactively refine the clustering results in various ways. Keywords can be adjusted so that they characterize each cluster better. In addition, our system can filter out noisy data and re-cluster the data accordingly. Cluster hierarchy can be constructed using a tree structure and for this purpose, the system supports cluster-level interactions such as sub-clustering, removing unimportant clusters, merging the clusters that have similar meanings, and moving certain clusters to any other node in the tree structure. Furthermore, the system provides document-level interactions such as moving mis-clustered documents to another cluster and removing useless documents. Finally, we present how interactive clustering is performed via iVisClustering by using real-world document data sets.

Categories and Subject Descriptors (according to ACM CCS): Information Systems [H.1.2]: User/Machine Systems—Human information processing; Database Management [H.2.8]: Database Applications—Data mining

1. Introduction and Motivation

Clustering is widely used in various fields such as data mining, machine learning, pattern recognition, information retrieval, and bioinformatics. Given data with unknown class labels (i.e., an unsupervised setting), clustering discovers the natural groupings of a data set. Popular applications of clustering include gene sequence analysis [ABN*99], image segmentation [CCZ07], document summarization [RJST04], social network analysis [KY08], and recommender systems [SKKR02].

As data grow exponentially [LV03], sense-making processes for information are becoming more and more important. In a general sense-making process, once users gather data for analysis, they want to re-represent and develop insight into the data [TC05]. However, since handling massive data sets in human sense-making processes is difficult, we apply data mining techniques such as clustering, which allows a better understanding of the data in terms of grouping similar data. Despite the need to use data mining algorithms, obtaining good analytical results and making enough sense out of them based only on automatic clustering algorithms is difficult. First, it is challenging to select an appropriate clustering algorithm. Various clustering algorithms [Jai10, JMF99] exist, and each of them has a different objective function that produces different clustering structures on the data set. Jain [Jai10] states that "there is no best clustering algorithm," which indicates even though a clus-

tering algorithm may perform effectively on a specific data set, it does not guarantee its performance on another data set. Even though the effectiveness of clustering on a chosen data set highly depends on the clustering algorithm, it is difficult to obtain prior knowledge about the data and determine the most appropriate clustering algorithm for their data set. Therefore, it is needed to refine the automatic clustering results via interactive methods to obtain better results despite the clustering algorithm.

Another challenge in clustering is that even if an appropriate clustering algorithm is selected, the algorithm may not reveal the correct results and several possible results may exist depending on the task. This happens since there is a "semantic gap" between low-level features and high-level human concepts. For example, if a set of facial images need to be clustered into age groups (e.g., infants, children, teenagers, and adults), it may be difficult to map the objective function of the clustering algorithm (a low-level characteristic such as the pixel values of an image) to the semantics of the clusters in the data (a high-level characteristic such as the age of a person in a picture). Since no clustering algorithm can identify true clusters (or semantic clusters), we cannot rely only on the automatic clustering procedure.

To complement the limitations of automatic clustering algorithms for a real-world problem, various interactive visualization techniques can be combined with automatic clustering algorithms. In addition, the amount of online document data is growing rapidly due to widespread availability and distribution via the internet such as online news articles, blogs, and e-mails. Motivated by this, we propose a visual analytics system, iVisClustering, that performs interactive clustering for document data. For the interactive clustering process, our system visualizes the automatic clustering results in various perspectives and provides interaction with these results through refining the clusters and constructing a hierarchical cluster structure towards a better understanding of the data. Throughout the visual analytic process, we use latent Dirichlet allocation (LDA) [BNJ03], a popular method that uses a probabilistic document modeling method, as an automatic clustering algorithm. In contrast to other algorithms such as *k*-means [Mac67], spectral clustering [NJW01], and linkage based methods [Sib73, Def77], which simply provide cluster indices, LDA models a document as a mixture of topics (or clusters) and assigns probability values to the document so that we can better understand the clustering results and interactively refine the results. LDA also overcomes the drawbacks of other topic models such as naive Bayes model [DP97] and probabilistic latent semantic indexing (pLSI) [Hof99]. More specifically, the naive Bayes model is too limited to model a large document data set, while the pLSI topic model easily leads to overfitting since it has too many parameters. Given these advantages of LDA, we decided to use it as the base algorithm in our interactive system. However, other clustering algorithms that generate latent topics and model a document as a mixture of those latent topics (e.g., Gaussian mixture models [Bis06], nonnegative matrix factorization [KP11], and various LDA extensions) can be easily adapted as the base algorithm of our system.

Clustering is a task that requires highly intensive mental load from humans and since different users want to create a different clustering based on their intents, we want to allow users to have full control over the clustering process via interactive visualization while having the benefits of machine assistance. With the help of LDA, we can start from an initial clustering result. However, in the initial clustering result, the meaning of most clusters may be unclear. iVisClustering provides interactions such as maintaining clusters with coherent meanings, removing clusters that contain outliers, re-clustering, sub-clustering, and constructing a hierarchical structure in order to help users manage good meaningful clustering results.

2. Related Work

In this section, we offer a brief summary of recent results on interactive clustering. iVibrate [CL06] enables users to verify intermediate clustering results and refine the cluster boundaries given by the algorithm. The system maps high-dimensional data points to a form similar to star coordinates [Kan00]. Seo and Shneiderman [SS05] proposed an interactive visual system for hierarchical clustering using a dendrogram. Simunic and Novak [SN04] visualizes labels from automatic clustering along with user-selected labels to perform interactive clustering. Bekkerman et al. [BRAE07] interactively revised the selected clusters by deleting noisy features, relocating misplaced features, and adding new features in every iteration of their own multi-modal distributional clustering. Their contribution is mainly based on the sentiment analysis of the document instead of clustering on general topics or concepts underlying the data. desJardins et al. [dMF07] presented a system that displays clusters in a 2D spring-embedded graph layout and allows users to move data nodes to satisfy their goal. Even though this system uses a force-directed graph layout, which is similar to the main view in our system, only limited interaction such as moving data nodes to other positions is provided for the constrained clustering algorithm. In other words, it does not refine the clustering results but just add some additional constraints via visual interactions. Unlike this system, iVisClustering allows manipulating of the data using both the low-level features and high-level results by including visualization and adding rich interaction methods, and therefore refines the clustering results interactively.

More recently, the TIARA [SWL*10] system uses LDA to create topic models. It visualizes multiple facets from a text corpus and allows users to interactively analyze the text data. It emphasizes the temporal aspects of the text data so that users can see the trends of topic changes. ParallelTopics [DWCR11] and TextFlow [CLT*11] explore topic evolution

and document characteristics based on topic distribution. Especially, TextFlow explores clusters that are merging or splitting over time and also extracts critical events such as cluster birth and death. It uses a non-parametric extension of LDA, Hierarchical Dirichet Process (HDP) [TJBB06], which has an advantage that the number of topics is determined by the data. HDP assumes data can be divided into groups and there are clusters in each group where these clusters are shared between groups. TextFlow successfully applied HDP by dividing data by time. Unlike this system, we use LDA since we focus on data that can not be divided into groups, which have shared clusters.

Even though, TIARA, ParallelTopics, and TextFlow use LDA (or an extension of LDA) as a base topic model, they are more focused on interpreting the LDA results. Our proposed system differs in helping users perform interactive clustering not only by interpreting the result but also manipulating the data and the intermediate results. iCluster [DFB11] also performs the clustering task via classification and recommendation. iCluster can be seen as a bottom-up approach where users start with an empty set of clusters and create clusters while the system assists users with machine learning techniques. On the other hand, iVisClustering can be seen as a top-down approach where it starts with an initial clustering result and gradually approaches to coherent clustering results via interactions. Commercial products such as Lingo3G [Lin] and Clustify [Clu] also perform document clustering but provide limited interactivity.

3. Latent Dirichlet Allocation

This section briefly introduces latent Dirichlet allocation (LDA) [BNJ03], which is widely used in the topic modeling of text document data, and shows how it is used in document clustering. LDA assumes a generative probabilistic model on text documents and their underlying topics. Given a pre-defined number of topics, k, LDA models each topic as a probability distribution over words. Just as each topic is modeled as a probability distribution over words, each document is modeled as a probability distribution over topics.

3.1. Probabilistic Model

Suppose we have M documents that are represented in N words, and let the vocabulary size (i.e., the number of distinct words) be denoted as V. LDA follows a graphical model, as shown in Figure 1. In this figure, nodes represent the parameters of the distributions or the instances sampled from the distributions where shaded nodes are observed variables and unshaded nodes are latent variables. The plate represents the repeated sampling operation where the sampling distribution parameters are defined from the nodes around the plate.

Let us describe in detail how the sampling procedure is done in LDA. First, in Figure 1, the Dirichlet distribution

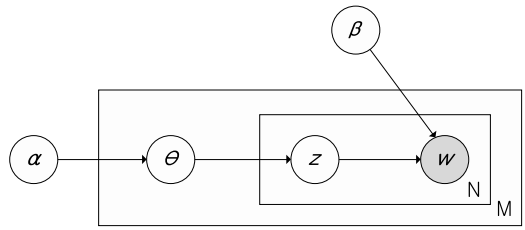


Figure 1: Graphical model representation of LDA.

parameterized by α is given as a prior from which an instance of a probability distribution over K topics are sampled for each of the M documents (the outer plate in Figure 1). Such an instance for each document i is denoted as $\theta_i = (\theta_{i1}, \theta_{i2}, \ldots, \theta_{iK}) \in \mathbb{R}^K$ $(\sum_{j=1}^{K} \theta_{ij} = 1)$, where the probability θ_{ij} represents how closely the document i is related to topic j. Afterwards, for each of the N words in document i, a single topic z is sampled from the topic distribution specified by θ_i (the inner plate in Figure 1), and then from the probability distribution over the words for the chosen topic, a word is sampled. The probability distribution over the words for each of the K topics is specified by another parameter $\beta \in \mathbb{R}^{K \times V}$, where β_{ij} represents the probability of word j in topic i and $\sum_{j=1}^{V} \beta_{ij} = 1$ for all i. For instance, when z is sampled as topic i, a word is sampled from the word probability distribution $(\beta_{i1}, \beta_{i2}, \ldots, \beta_{iV})$ associated with topic i.

3.2. Inference and Parameter Estimation

In the LDA model, the only observed variable is w, i.e., the words shown in documents. In our system, α, a parameter of the Dirichlet prior on the per-document topic distributions, is fixed as a uniform Dirichlet prior and all the other variables (i.e., β, θ, and z) are to be determined. Variable β is the (deterministic) parameter of the distribution, and θ and z are the instances sampled from such distribution. The algorithm solves these two groups alternatively by fixing one group and solving for the other each time. The inference step refers to solving for θ and z given β, and the parameter estimation is to solve for β given θ and z. By alternating the inference and the parameter estimation steps, LDA finds the topics that can best explain the given document set and the best representation of documents in terms of such topics by maximizing the overall likelihood.

In our system, either the full LDA algorithm is used to perform both of the inference step and the parameter estimation step or only the inference step can be performed to calculate θ and z given fixed β. In other words, in the inference step, the system can calculate the topic distribution for each document given the term distribution for each topic. This inference step is utilized in our system when the topic needs to be refined. By controlling the term distributions β, we can get the new topic distribution θ for each document.

3.3. LDA Computation in iVisClustering

A term-document matrix is given as an input to LDA and it outputs two matrices, the document-topic distribution matrix θ and the topic-term distribution matrix β.

The document-topic distribution matrix $\theta \in \mathbb{R}^{M\times K}$ consists of M rows, where the i-th row $\theta_i \in \mathbb{R}^K$ is the topic distribution for document i. A high probability value of θ_{ij} indicates that document i is closely related to topic j. In addition, documents with low θ_{ij} values over all the topics are noisy documents that belong to none of the topics. Therefore, by looking at the θ_{ij} values, one can understand how closely the document is related to the topic. In a similar way, the topic-term distribution matrix $\beta \in \mathbb{R}^{K\times V}$ consists of K rows, where the i-th row $\beta_i \in \mathbb{R}^V$ is the word distribution of topic i. The terms with high β_{ij} values indicate that they are the representative terms of topic i. Therefore, by looking at such terms one can grasp the meaning of each topic without looking at the individual documents in the cluster.

These two outputs are the main ingredients that are used for the interactive analysis in iVisClustering. The details on interaction will be discussed in Section 4.2.

4. System Description

4.1. Data Encoding and Preprocessing

Given a document set, iVisClustering encodes the document set as a matrix using a bag-of-words model where each document is represented as a column vector in the matrix. The dimension, which is the number of rows in the matrix, is equal to the number of distinct words in the entire document set. In general, this dimension can be easily in the hundreds of thousands. After stemming and stop-word removal, a term-document matrix is generated and it is given as an input to LDA. For details on LDA outputs, refer to Section 3.3.

4.2. Visualization Modules

iVisClustering provides various perspectives on a document set through multiple visualization modules called "views," which interact with each other. Because clustering can be performed in so many different ways, the multiple views provide flexibility for varying tasks.

4.2.1. Cluster Relation View

The Cluster Relation View, shown in Figure 2A, represents an overview of the LDA clustering results of a document set. The view uses a force-directed layout of a node-link graph, produced using parts of the Prefuse software library [HCL05]. Using this view, clusters can be visually determined to understand the overall structure of the data easily. This view shows hard-clustering results by assigning each document i to topic j with $j = \arg\max_{j\in T} \theta_{ij},$, where T is the set of all topics. Each document node is visualized as a colored circle. Document nodes with the same color belong to the same cluster. The edges between nodes represent how similar the documents are based on cosine similarity. By controlling the slider value, the edges with values larger than the slider value appear and those with values smaller disappear. We can interactively control the parameter value so that the edges are not so crowded. This is useful to find the local neighborhood structure on the data and it also speeds up the software since it is using fewer edges.

The summary of each cluster is visualized as a cluster summary node. A cluster summary node is illustrated as a color-bordered rectangle with each cluster's most representative keywords. We chose this design to investigate relations among the data without occlusions. The system also provides interactions to explore a specific cluster. First, if the mouse pointer is over the cluster summary node, it activates X-ray mode. Second, if the mouse is right-clicked on the cluster summary node, the corresponding Term-Weight View for the cluster pops up. Finally, the documents in the cluster are shown in the Document View (Figure 2G) using the mouse left-click interaction over the cluster summary node. Removing these cluster summary nodes is also allowed to investigate data-level relations more clear.

Invisible edges are connected between the cluster summary node and the nodes in its cluster. Interactions with the edge spring length can control the scatteredness of each cluster. If the edge spring length is set to a small value, all the nodes that belong to a same cluster are gathered to a single point. Eades and Huang [EH00] used a similar approach to visualize clustered graphs by creating forces between all vertices in the same cluster.

When the Cluster Relation View is cluttered and the cluster summary nodes are difficult to read, we can also explore using the Cluster Summary View, shown in Figure 2C. In the Cluster Summary View, only the cluster summary nodes are shown in an organized grid layout and provides the same interactions as in the Cluster Relation View.

X-ray Mode

The Cluster Relation View and the Cluster Summary View provides an X-ray mode, shown in Figure 3A. If the mouse pointer is over a cluster summary node, it activates the X-ray mode and the documents in the corresponding cluster are highlighted in the Parallel Coordinate View, as shown in Figure 3B. The X-ray mode displays documents in a grid layout, and each square in the grid indicates a single document that is included in the current cluster. The grid square is dark if the document is strongly related to this topic and light if it is weakly related. The X-ray mode also contains a color spectrum below the document grid squares, that shows how this cluster is related to other clusters. In this color spectrum, each color indicates a cluster, so the relatedness between the chosen cluster and other clusters can be identified by observing the ratio of the colored area.

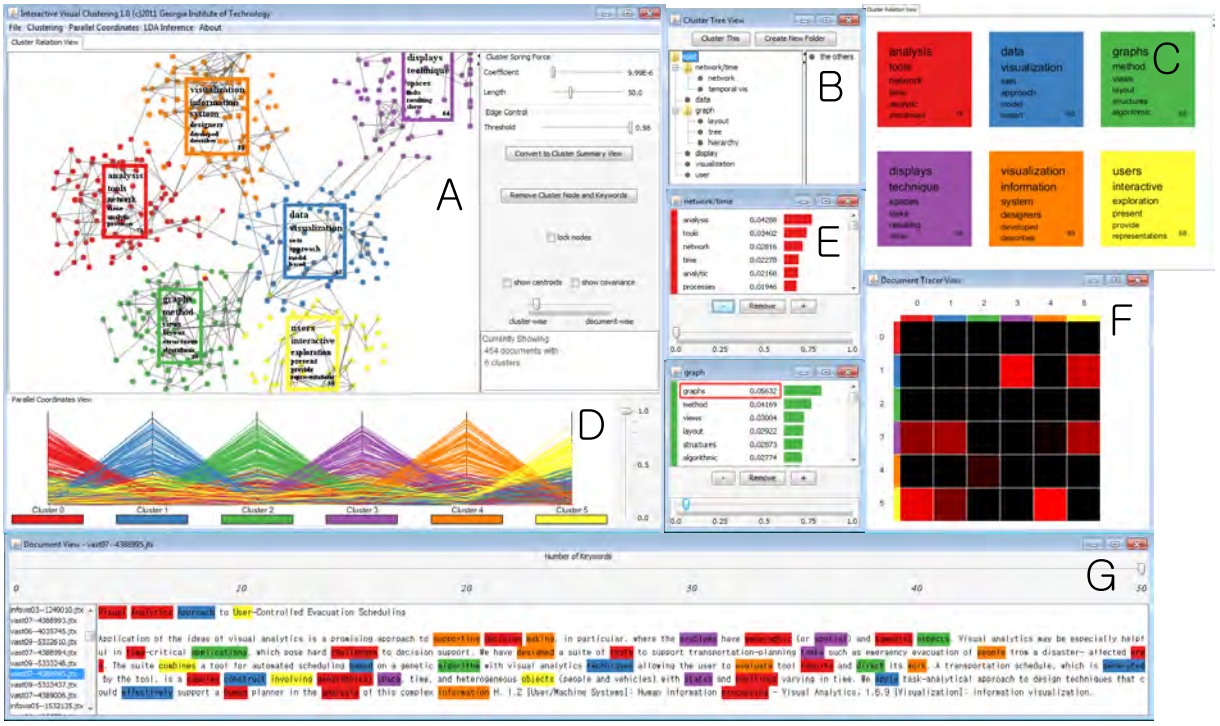


Figure 2: The overview of the system. The InfoVis and VAST papers data set is used. (A) Cluster Relation View. Visualizes clustering results in a graph-based layout. (B) Cluster Tree View. Maintains the hierarchical cluster structure with user-defined topics. (C) Cluster Summary View. A simplified version of the Cluster Relation View. (D) Parallel Coordinates View. The topic distribution of each document is visualized. (E) Term-Weight View. Visualizes term weights of each topic and can modifying its value. (F) Document Tracer View. The number of documents which changed its cluster membership is shown as a heat map and those documents are accessible. (G) Document View. The original document is shown with keywords highlighted.

If the mouse pointer is moved over a grid square, the corresponding document is highlighted in the Parallel Coordinates with a thick black line, as shown in Figure 3C. The color spectrum on the bottom of the document grid squares shows the relatedness between the chosen document and the clusters. This interaction between the X-ray and the Parallel Coordinates View allows data exploration in various ways. For example, by moving the mouse cursor over the grids in the X-ray, patterns of individual documents can be observed in the Parallel Coordinates View. As a result, we can quickly find documents that contain several topics, which will be further discussed when we explain the Parallel Coordinates View in Section 4.2.3.

Term-Weight View

The Term-Weight View in Figure 2E shows all the terms and the corresponding β_{ij} values (probability value of word j in topic i), and the weight values can be interactively modified. By controlling the weights of certain terms in each topic, one can impose their own cluster meanings. After new cluster meanings are imposed, new topic models are generated by running the LDA inference step, discussed in Section 3.2. For example, if the weight of a certain word is decreased in

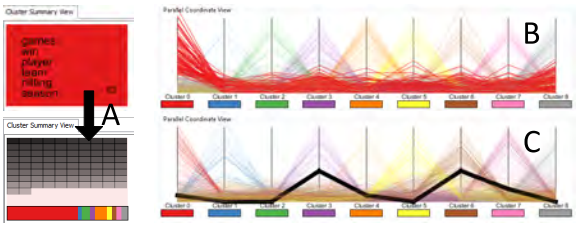


Figure 3: X-ray mode: (A) Cluster node's X-ray mode. It shows document grids with the color spectrum. (B) Parallel Coordinates' X-ray mode with the selected cluster. (C) Parallel Coordinates' X-ray mode with the selected document.

the chosen topic, then LDA optimizes the document-topic distribution so that documents containing the word have a smaller weight on this topic. If the weight of a certain word is increased, LDA optimizes the document-topic distribution so that the system collects the documents that have a high probability of containing the word. The documents that change their cluster assignment during this inference step are shown in the Document Tracer View, as explained in Section 4.2.5. We will show an usage scenario related to this interaction in Section 5.2.

4.2.2. Cluster Tree View

One of the main goals of iVisClustering is to refine vague topics while maintaining coherent topics. When LDA performs clustering, both coherent topic clusters and meaningless topic clusters may exist together. In the Cluster Tree View, users can perform several interactions to clarify existing topics or find new topics during the analytic process.

The Cluster Tree View, shown in Figure 2B, maintains the hierarchical structure of the clusters based on users' intentions. It represents the clustering results hierarchically using a traditional tree visualization. After loading the data set and performing an initial clustering, the tree has a root node that contains the entire document set, and the root node has k child nodes where k indicates the number of clusters. The Cluster Tree View enables users to interact and reproduce the clustering results based on the following five interactions: deleting, merging, moving, re-clustering, and sub-clustering.

Deleting, merging, and moving operations are performed by drag-and-drop mouse interactions. One can delete unimportant clusters by dragging the cluster to the right column of the view. Deleted documents are maintained since it can potentially be used to find new topics. In addition, one can also add the deleted documents back into the clusters by performing classification after they generate coherent clusters. Merging clusters that have similar meanings and moving clusters to other nodes in the tree are also possible.

Re-clustering is used to perform the clustering again with a different k value. Different choices of k (number of clusters) produce different topics, and so users can experiment with this value to detect coherent cluster topics. When re-clustering is performed, the Hungarian algorithm [Kuh55] finds the best pairwise matchings between the original clusters and the new clusters. Then, the cluster colors are changed so that the matching clusters have the same color. Sub-clustering is used to deeply explore the data in a specific cluster when the topic cluster is broad. For example, clustering results in the first level are often topics of broader subjects such as "sports," "medicine," and "religion." If we intuitively identify more specific sub-topics such as "basketball" or "soccer," then we can interact with the results by sub-clustering the "sports" cluster. Sub-clustering is also beneficial when some topics are more distinguishable than others. For example, let us assume a document set consists of five topics: "baseball," "basketball," "soccer," "tennis," and "physics." If we run LDA to create a topic model with five clusters, then it will easily distinguish "physics" from the others since the algorithm generally focuses on splitting the most distinguishable topic first. However, the other four topics may be mixed together. We can create a subset of the entire document set by excluding the documents related to "physics" and re-run LDA with four clusters to obtain clearer topics. Another way is running sub-clustering on each of the mixed topics and merging similar sub-clusters to create a coherent topic. We will show an usage scenario using this interaction in Section 5.3.

The Cluster Tree View can also create a new empty folder or annotate cluster names. By clicking the "Create New Folder" button at the top of the view, an additional cluster is generated and documents that represent a new topic can be moved to this new folder. We can also assign a name to the cluster by right clicking on the cluster.

4.2.3. Parallel Coordinates View

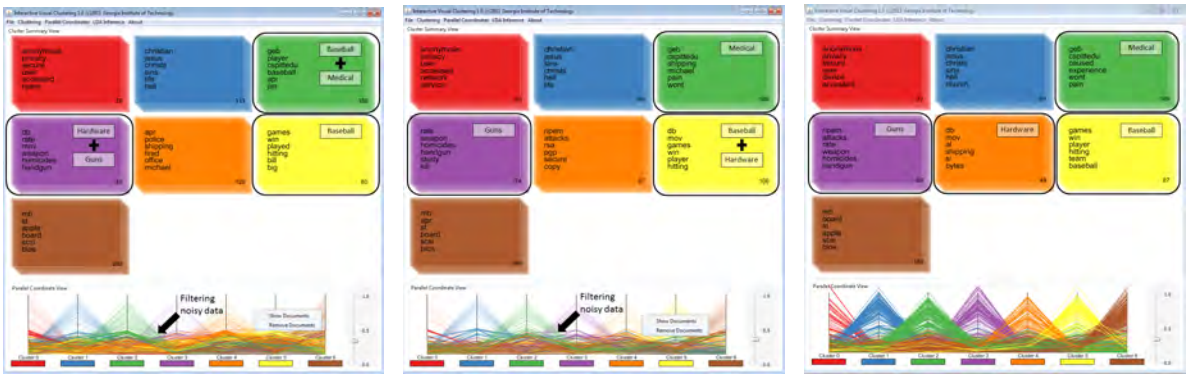
The Parallel Coordinates View, shown in Figure 2D, visualizes multiple dimensions on a 2D screen. Parallel coordinates was chosen since it is a simple way to visualize and interact with multiple variables at the same time. In this view, for each document i, we visualize the vector $\theta_i = (\theta_{i1}, \theta_{i2}, \ldots, \theta_{iK})$, which represents its soft clustering result on K topics as in ParallelTopics [DWCR11]. Each vertical axis corresponds to a topic and its value represents the probability value. Each line in the view indicates a document and is color-coded depending on its cluster membership.

The Parallel Coordinates View interacts with the Cluster Relation View or the Cluster Summary View. When the mouse cursor is placed on the cluster summary nodes, the X-ray mode is activated, and the documents of the corresponding cluster are highlighted in the Parallel Coordinates View. The purpose of this interaction is to let users understand the overview of the document-topic distribution of clusters or the characteristics of a single document. For example, a document line with several peaks in the parallel coordinates indicates that the document is a mixture of topics that are potentially related. In addition, if most of the documents in a cluster have multiple peaks in common, it indicates that the topics with the peaks are relevant to one another.

The Parallel Coordinates View also has a slider to set a threshold value. If a threshold value is set, it filters the documents with θ_{ij} values lower than the threshold value over all the topics in order to filter out noisy documents that do not clearly belong to a certain cluster. We will show an usage scenario including this interaction in Section 5.1.

4.2.4. Document View

During interaction, accessing the original text is important to understand why a document is assigned to its cluster. Our system includes the Document View, shown in Figure 2G, that displays the original documents. Along with the original text, the Document View also highlights terms in different colors according to which topic the terms belong to. We can explore the documents while investigating other views. It shows the documents in the chosen cluster if a mouse left click on the Cluster Relation View or the Cluster Summary View is performed and also shows the documents in a user-selected region from the Parallel Coordinates View. It also displays documents that change cluster assignment after the LDA inference step, explained in Section 4.2.5.



(a) The initial clustering result (b) The intermediate clustering result (c) The final clustering result

Figure 4: Interactive clustering by filtering noisy data. Filtering out noisy documents leads to a clear clustering results.

The Document View consists of three parts: a document text representation, a document list, and a slider to control the k value. In the center of the view, the actual text of the selected document is visualized with the highlighted terms. The document list on the left side shows a list of all the included documents. If the cluster summary node is clicked in the Cluster Relation View or the Cluster Summary View, the documents are sorted in a decreasing order of the probability values that the document is related to the selected cluster. The terms in the document are highlighted with different colors according to their topics, and the terms to be highlighted are determined by the top k weighted terms from each cluster where the slider at the top controls the k value. We can initially set k to a small value and see the most important words first and then gradually increase the k value to see which color dominates the document. If a certain color dominates the Document View, it implies that the selected document strongly belongs to the corresponding topic. In addition, one can change the term weight values by right clicking on the term in the Document View.

4.2.5. Document Tracer View

The Document Tracer View, shown in Figure 2F, is a heat map view visualizing the transition of documents between clusters. After manipulating the topic-term distribution in the Term-Weight View (i.e., the β_{ij} values), the LDA inference step is performed to obtain new clustering results. The Document Tracer View shows how many documents move from one topic cluster to another when the clustering result changes. The heat map has $k \times k$ elements, where k is the number of clusters and the (i, j)-th square represents how many documents changed their cluster assignment from cluster i to j after the LDA inference step. By clicking the (i, j)-th square, the documents that change their cluster assignment from i to j are loaded in the Document View. Exploring these documents, users can analyze how their user-defined topic affects the clustering results.

4.3. General Analysis Procedure

In this section we will discuss a possible analysis procedure using iVisClustering. Let us say that an analyst needs to review a large set of documents and wants to divide the documents into groups to efficiently analyze it separately. Using iVisClustering, the analyst can perform the following steps: 1) perform automatic clustering with a predefined number of clusters; 2) filter out noisy documents (i.e., documents that do not strongly belong to a single cluster) to maintain relatively clear documents; 3) perform cluster-level interactions such as combining similar clusters, dividing clusters that are a mixture of multiple topics, and removing clusters that hold no interest so that only meaningful topics remain; 4) refine the meaning of each cluster using the LDA inference algorithm so that the analyst can control the meanings of the clusters; and 5) perform data-level interactions to fine-tune the clusters such as reviewing the documents located on the boundary of the clusters using the Document View to easily capture what the topic is, and either move the documents to another cluster or remove the document if not needed. By performing these five steps, an analyst is able to maintain meaningful topic clusters for future use. For example, the analyst can classify incoming documents or easily locate old documents of interest.

5. Usage Scenarios

In this section, we present an interactive clustering analysis using a real-world data set of text documents such as publication data and news data. The InfoVis and VAST paper data set [Inf] consists of 454 titles and abstracts published in InfoVis and VAST from 1995 to 2009. The 20 newsgroups data set [AN07] is a collection of newsgroup documents composed of 20 topics. In our study, we chose eight topic subsets, each of which contains 100 documents. We use these labeled data to show analysis results based on true labels. In the following subsections, we present four usage scenarios.

5.1. Interactive Clustering by Filtering Noisy Data

When we cluster the documents based on LDA, the results may contain many noisy documents. In general, the noisy documents have low θ_{ij} values over all the topics. In our system, we can detect these noisy documents through the Parallel Coordinates View.

When we perform initial clustering with seven clusters (users can choose any number as the initial number of clusters, but here we chose an arbitrary number), we observe several mixed clusters as shown in Figure 4a. For example, one cluster represents a mixed topic of "baseball" and "medicine," and another represents a mixed topic of "hardware" and "guns." If we decrease the threshold value in the Parallel Coordinates View, it leaves only noisy data items that do not strongly belong to a specific cluster. By filtering out these noisy documents, we expect to obtain more coherent topics. After the first filtering interaction, we have better clustering results as shown in Figure 4b. However, this result still includes one mixed cluster, a mixed topic of "baseball" and "hardware." After we perform the filtering interaction again, most of the clusters represent coherent topics, as shown in Figure 4c.

5.2. Interactive Clustering by Refining Topics

We can obtain an understanding of the initial clustering results from the Cluster Relation View, where representative keywords in each cluster are visualized. The initial results do not always provide clear topics, and cluster meanings should be refined. iVisClustering guides users to perform this process using the Term-Weight View (detailed in Section 4.2.1). We will show how to explore and interact with the 20 newsgroups data set using this process.

At first, from the Cluster Relation View we acquire an initial understanding of the clustering results. In Figure 5a, there are nine clusters, each with six keywords. By observing the keywords in each cluster, we can identify the characteristics of each cluster. For example, cluster 0, in light blue, includes keywords highly related to "religion." We can see that cluster 2, in light green, contains many keywords related to "criminal," and cluster 8, in purple, contains business-related keywords. However, we can see that some clusters contain irrelevant keywords. Cluster 0 shows the term "disease," which does not seem to be related to the topic "religion." We can also find that the terms "patients" and "pain" are not associated with clusters 2 and 8 (i.e., the arrow-pointed terms in Figure 5a).

By controlling these unrelated keywords in the Term-Weight View, we can refine the topics with user-driven cluster meanings. As shown in Figure 5b, we decrease the weight of "disease" and "pain" in clusters 0 and 8, respectively. However, by exploring all the terms in the Term-Weight View of cluster 2, we can identify that this cluster contains a number of medical-related terms. Therefore, we increase the weights of medical-related terms to get medical-oriented documents in this cluster. After changing the weights of the terms, the system performs the LDA inference step and shows the new clustering results.

Figure 5c shows the Document Tracer View, from which we can observe documents moving from the original cluster to a new cluster due to the term weight changes in the previous step. As shown in this figure, many documents are re-assigned to cluster 2. We can see that many documents moved to cluster 2. By exploring such documents using the Document View, we can observe that they are mostly medical-related. This is because we increased the weights of medical-related terms of cluster 2 in Figure 5b. In addition, one technology-related document is re-assigned from cluster 3 to cluster 6. When some confusing topics are adjusted to coherent topics, we can also expect an indirect effect that other topics will be adjusted. From this observation, we can see that changing the term weights in a subset of clusters may lead to better clustering results across the entire data set.

5.3. Interactive Clustering by Sub-clustering and Merging

Here we illustrate how we can identify more coherent topics by sub-clustering mixed clusters and merging similar sub-clusters with the InfoVis and VAST paper data set. Figure 6A shows the initial clustering result and Figure 6B shows the result of sub-clustering the blue cluster in Figure 6A. In Figure 6B, we can see that red, blue, and purple cluster has a topic related to "treemap," "graphs," and "tree," respectively. However, the green cluster is not related to the other clusters. The green cluster in Figure 6B has similar topics to the yellow and red clusters in Figure 6A, which are neighborhood clusters of the blue cluster in Figure 6A. Further sub-clustering the green cluster in Figure 6B, we discover clusters related to "users" and "display" as shown in Figure 6C. Finally, we can merge similar clusters by a drag and drop interaction and obtain the final clustering results, as shown in Figure 6D. This example illustrates that even if some clusters are initially made up of mixed topics, performing sub-clustering and merging may lead to coherent and meaningful topics.

5.4. Document Exploration in the Cluster Relation View

We now show how the Cluster Relation View allows us to further explore individual documents. If a document is placed near the cluster boundary and have many edges connected to documents in other clusters, it is likely to be a document with mixed topics or a mis-clustered document. For example, since two documents on the cluster boundary, shown in Figure 6B, include many other keywords that represent other cluster topics, we may assume these documents are mis-clustered, and remove or move these documents to an appropriate cluster interactively.

(a) The initial clustering result. (b) The change of the term weights in each topic. (c) Result of the user-driven cluster meanings.

Figure 5: Interactive clustering by refining topics. Users focus on the arrow-pointed terms, and either increase or decrease the term weights in their topic. The result is shown with three documents as an example.

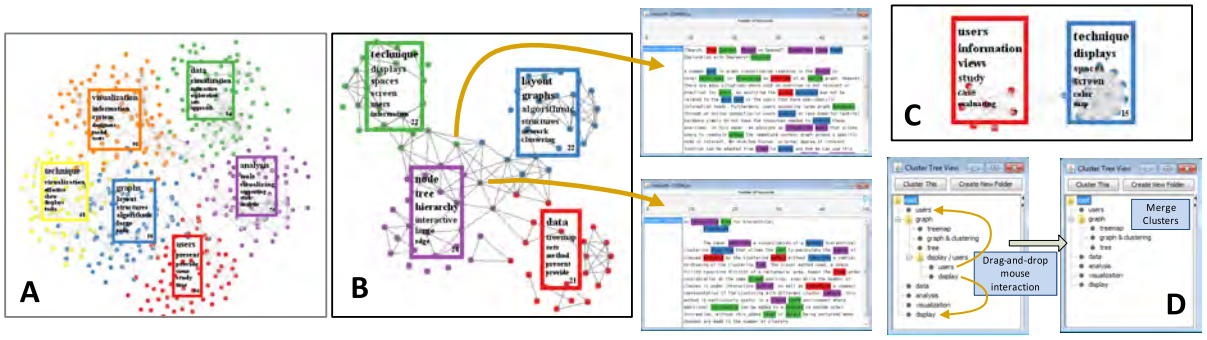


Figure 6: Interactive clustering by sub-clustering and merging. Document exploration in the Cluster Relation View.

6. Conclusion and Future Work

In this paper, we presented iVisClustering, a visual analytics system that performs document clustering interactively. We used latent Dirichlet allocation (LDA), a state-of-the-art topic modeling method for the automatic clustering procedure, and integrated it in an interactive visual analytics system that allows users to guide the clustering process. iVisClustering has multiple views that visualize the clustering results so that users can easily interact with them. We have presented four usage scenarios that show how the system can lead to better clustering results by effectively combining the automatic clustering algorithm with interactive visualization techniques. iVisClustering can work with any soft clustering algorithm such as Gaussian mixture model [Bis06] and non-negative matrix factorization [KP11], while maintaining the full interactive visual capability. Therefore, regarding scalability issues when the data size is large, computationally efficient methods such as Gaussian mixture model can easily replace LDA in our system. iVisClustering is designed especially for clustering of a document data set. Interaction in the system utilizes a special characteristics of document data which is that the low-level features (e.g., keywords) of documents have meanings that are understood by people.

We illustrated the capability of iVisClustering by showing several usage scenarios. To further evaluate the system, we plan to conduct a user study. The user study needs to be carefully designed and conducted since our system has complex interfaces with many views and interactions and the clustering task itself is not easy to evaluate. By gathering data and feedback from users, we will be able to analyze patterns of usage, usability of visualization modules and interactions, and the efficiency of people using data mining techniques in the system.

7. Acknowledgements

The work of these authors was supported in part by National Science Foundation grant CCF-0808863. Any opinions, findings, conclusions, or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

References

[ABN*99] ALON U., BARKAI N., NOTTERMAN D. A., GISH K., YBARRA S., MACK D., LEVINE A. J.: Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *Proceedings of the National Academy of Sciences of USA 96*, 12 (1999), 6745–6750. 1

[AN07] ASUNCION A., NEWMAN D.: UCI machine learning repository. University of California, Irvine, School of Information and Computer Sciences, 2007. 7

[Bis06] BISHOP C. M.: *Pattern Recognition and Machine Learning*. Springer, 2006. 2, 9

[BNJ03] BLEI D. M., NG A. Y., JORDAN M. I.: Latent dirichlet allocation. *Journal of Machine Learning Research 3* (March 2003), 993–1022. 2, 3

[BRAE07] BEKKERMAN R., RAGHAVAN H., ALLAN J., EGUCHI K.: Interactive clustering of text collections according to a user-specified criterion. In *Proceedings of IJCAI-20* (2007), pp. 684–689. 2

[CCZ07] CAI W., CHEN S., ZHANG D.: Fast and robust fuzzy c-means clustering algorithms incorporating local information for image segmentation. *Pattern Recognition 40* (March 2007), 825–838. 1

[CL06] CHEN K., LIU L.: iVIBRATE: Interactive visualization-based framework for clustering large datasets. *ACM Transactions on Information Systems 24*, 2 (2006), 245–294. 2

[CLT*11] CUI W., LIU S., TAN L., SHI C., SONG Y., GAO Z., QU H., TONG X.: TextFlow: Towards better understanding of evolving topics in text. *IEEE Transactions on Visualization and Computer Graphics 17* (Dec. 2011), 2412–2421. 2

[Clu] Clustify. http://www.cluster-text.com/. 3

[Def77] DEFAYS D.: An efficient algorithm for a complete link method. *Comput. J. 20*, 4 (1977), 364–366. 2

[DFB11] DRUCKER S. M., FISHER D., BASU S.: Helping users sort faster with adaptive machine learning recommendations. In *Proceedings of the 13th IFIP TC 13 International Conference on Human-Computer Interaction* (2011), INTERACT'11, Springer-Verlag, pp. 187–203. 3

[dMF07] DESJARDINS M., MACGLASHAN J., FERRAIOLI J.: Interactive visual clustering. In *Proceedings of the 12th international conference on Intelligent user interfaces* (New York, NY, USA, 2007), IUI '07, ACM, pp. 361–364. 2

[DP97] DOMINGOS P., PAZZANI M.: On the optimality of the simple bayesian classifier under zero-one loss. *Machine Learning 29* (November 1997), 103–130. 2

[DWCR11] DOU W., WANG X., CHANG R., RIBARSKY W.: ParallelTopics: A probabilistic approach to exploring document collections. In *Proceedings of IEEE Conference on Visual Analytics Science and Technology (VAST)* (2011). 2, 6

[EH00] EADES P., HUANG M. L.: Navigating clustered graphs using force-directed methods. *Journal of Graph Algorithms and Applications 4*, 3 (2000), 157–181. 4

[HCL05] HEER J., CARD S. K., LANDAY J. A.: prefuse: a toolkit for interactive information visualization. In *Proceedings of the SIGCHI conference on Human factors in computing systems (CHI)* (2005), ACM, pp. 421–430. 4

[Hof99] HOFMANN T.: Probabilistic latent semantic indexing. In *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval (SIGIR)* (1999), ACM, pp. 50–57. 2

[Inf] InfoVis and VAST papers. http://www.cc.gatech.edu/gvu/ii/jigsaw/datafiles.html. 7

[Jai10] JAIN A. K.: Data clustering: 50 years beyond k-means. *Pattern Recognition Letters 31*, 8 (2010), 651 – 666. 1

[JMF99] JAIN A. K., MURTY M. N., FLYNN P. J.: Data clustering: A review. *ACM Computing Surveys 31*, 3 (1999), 264–323. 1

[Kan00] KANDOGAN E.: Star coordinates: A multi-dimensional visualization technique with uniform treatment of dimensions. In *In Proceedings of the IEEE Information Visualization Symposium, Late Breaking Hot Topics* (2000), pp. 9–12. 2

[KP11] KIM J., PARK H.: Fast nonnegative matrix factorization: An active-set-like method and comparisons. *SIAM Journal on Scientific Computing 33*, 6 (2011), 3261–3281. 2, 9

[Kuh55] KUHN H. W.: The Hungarian method for the assignment problem. *Naval Research Logistic Quarterly 2* (1955), 83–97. 6

[KY08] KNOKE D., YANG S.: *Social Network Analysis*. Sage Publications, Inc. Thousand Oaks, CA, USA, 2008. 1

[Lin] Lingo3G. http://carrotsearch.com/lingo3g-overview.html. 3

[LV03] LYMAN P., VARIAN H. R.: How Much Information? SIMS, University of California at Berkeley, CA, US. http://www.sims.berkeley.edu/how-much-info-2003, 2003. 1

[Mac67] MACQUEEN J. B.: Some methods for classification and analysis of multivariate observations. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability* (1967), vol. 1, University of California Press, pp. 281–297. 2

[NJW01] NG A. Y., JORDAN M. I., WEISS Y.: On spectral clustering: Analysis and an algorithm. In *Advances in Neural Information Processing Systems (NIPS)* (2001), MIT Press, pp. 849–856. 2

[RJST04] RADEV D. R., JING H., STYŚ M., TAM D.: Centroid-based summarization of multiple documents. *Information Processing and Management 40* (November 2004), 919–938. 1

[Sib73] SIBSON R.: SLINK: An optimally efficient algorithm for the single-link cluster method. *The Computer Journal 16*, 1 (January 1973), 30–34. 2

[SKKR02] SARWAR B., KARYPIS G., KONSTAN J., RIEDL J.: Recommender systems for large-scale e-commerce: Scalable neighborhood formation using clustering. In *Proceedings of the Fifth International Conference on Computer and Information Technology* (2002). 1

[SN04] SIMUNIC K., NOVAK J.: Combining visualization and interactive clustering for exploring large document pools. In *Proceedings of the 4th IAESTED International Conference on Visualization, Imaging, and Image Processing* (2004), pp. 141–146. 2

[SS05] SEO J., SHNEIDERMAN B.: A rank-by-feature framework for interactive exploration of multidimensional data. *Information Visualization 4* (July 2005), 96–113. 2

[SWL*10] SHI L., WEI F., LIU S., TAN L., LIAN X., ZHOU M. X.: Understanding text corpora with multiple facts. In *Proceedings of IEEE Conference on Visual Analytics Science and Technology (VAST)* (2010), pp. 99–106. 2

[TC05] THOMAS J. J., COOK K. A.: *Illuminating the Path: The R&D Agenda for Visual Analytics*. National Visualization and Analytics Center, 2005. 1

[TJBB06] TEH Y. W., JORDAN M. I., BEAL M. J., BLEI D. M.: Hierarchical Dirichlet processes. *Journal of the American Statistical Association 101*, 476 (2006), 1566–1581. 3