An Interactive Visual Analytics System for Incremental Classification Based on Semi-supervised Topic Modeling

Yuyu Yan* Yubo Tao† Sichen Jin‡ Jin Xu§ Hai Lin¶

State Key Lab of CAD&CG, Zhejiang University

Abstract

Text labeling for classification is a time-consuming and unintuitive process. Given an unannotated text collection, it is difficult for users to determine what label to create and how to label the initial training set for classification. Thus, we present an interactive visual analytics system for incremental text classification based on a semi-supervised topic modeling method, modified Gibbs sampling maximum entropy discrimination latent Dirichlet allocation (Gibbs MedLDA). Given a text collection, Gibbs MedLDA generates topics as a summary of the text collection. We design a scatter plot to display documents and topics simultaneously to show the topic information, and this helps users explore the text collection structurally and find labels for creating. After labeling documents, Gibbs MedLDA is applied to the text collection with labels again, and it generates both the topic and classification information. We also provide a scatter plot with the classifier boundary and a matrix view to present weights of classifiers. Users can iteratively label documents to refine each classifier. We evaluate our system via a user study with a benchmark corpus for text classification and case studies with two unannotated text collections.

Index Terms: Human-centered computing—Visualization—Visualization application domains—Visual analytics;

1 Introduction

As an increasing number of texts are being produced and archived, organizing these documents has become an essential task in text analysis. Text classification is a widely used method for organizing an extensive collection of documents and has many applications, such as text retrieval and filtering. Generally, text classification is a supervised or semi-supervised method and requires a sufficient number of annotated documents to train a high-quality model. Different applications may need differently annotated documents in various domains, and thus, many documents are labeled manually by domain experts. However, text labeling is usually a time-consuming and unintuitive process. Thus, obtaining high-quality annotated documents with which to train a strong classifier is a challenging task in text classification.

Active learning is a machine learning method widely used to reduce labeling cost. An active learning algorithm iteratively selects a sample that is most deserved to be labeled based on selection strategies. However, users have little control over the sample selection. To overcome this flaw, visual-interactive labeling with visual analytics has been proposed. Seifert and Granitzer [27] provided an interactive visualization of a classifiers a-posteriori output probabilities to help users to select a sample to label. Heimerl et al. [17] used the search results of the Apache Lucene framework as an initial annotated training set and proposed three methods for labeling documents for classifiers. However, when users are not familiar with the text collection, previous works do not help them determine what label to create.

To address the above shortcomings, we introduce topic information to the practice of visual-interactive labeling. Topic information provides a global overview of what labels to create. Furthermore, visual-interactive labeling requires classification information to guide users in finding document candidates for labeling. Thus, we require both topic information and classification information for users to create labels and to label documents. We select Gibbs MedLDA [33] to analyze the text collection, because Gibbs MedLDA integrates a topic model (e.g., latent Dirichlet allocation (LDA)) with a max-margin prediction model (e.g., support vector machines (SVMs)), and provides both topic information and classification information. However, Gibbs MedLDA may not be suitable for visual-interactive labeling, as labels are added to a text collection gradually. Therefore, we modify Gibbs MedLDA to produce a multi-label semi-supervised topic model with an active learning algorithm.

To allow users to label documents intuitively, we propose an interactive visual analytics system for incremental classification. It contains three parts: topic visualization, classification visualization, and document visualization. Topic visualization helps users understand the text collection to create labels. Classification visualization helps users understand and refine classifiers. Document visualization displays the meta information and labels of documents.

The main contributions of this paper are as follows:

- We change Gibbs MedLDA to a multi-label semi-supervised topic model and integrate a margin-based active learning algorithm with Gibbs MedLDA for visual-interactive labeling.
- We present an interactive visual analytics system for incremental classification. Our system helps users create labels and find document candidates for labeling.
- We evaluate the usability of our system through two case studies and a user study.

2 Related Work

Because our system is based on the supervised topic model, we introduce the supervised topic model, topic model visualization, and interactive visual classification.

2.1 Supervised Topic Model

LDA, proposed by Blei et al. [6], stratifies an extensive collection of documents by projecting every text into a low-dimensional space spanned by a set of bases that capture the semantic aspects, also known as topics, of the text collection. Although we can easily obtain useful information other than text content, such as rating scores of reviews and tags in documents, this information cannot be directly utilized in the original LDA to generate a better topic model. Blei et al. [5] further proposed a supervised topic model that captures other information as a regression response and yields latent

*e-mail: yanyuyu001@gmail.com
†e-mail: taoyubo@cad.zju.edu.cn (corresponding author)
‡e-mail: 3150104297@zju.edu.cn
§e-mail: jinxu.zju@gmail.com
¶e-mail: lin@cad.zju.edu.cn

2165-8773/19/$31.00 ©2019 IEEE
DOI 10.1109/PacificVis.2019.00025

topical representations that are more discriminative and more suitable for prediction tasks. Supervised topic models predominantly employ likelihood-driven objective functions, which may become complicated for learning and inference, especially the exponential family response. Zhu et al. [32] introduced MedLDA, which integrates the mechanism behind the max-margin classification models with hierarchical Bayesian topic models and made the learning and inference process much simpler. Recently, Zhu et al. [33] employed a fast and straightforward Gibbs sampling algorithm to infer the MedLDA model. We use Gibbs MedLDA in our paper, and adapt it as a multi-label semi-supervised model due to the characteristics of visual-interactive labeling.

2.2 Topic Model Visualization

Topic models have commonly been used to understand text collections. Chaney and Blei [8] used the form of a list of words to show the latent semantic structure produced by a topic model and to help users understand topic meanings. However, this method does not display the correlation between topics. Chuang et al. [12] proposed Termite, which uses a tabular layout to promote the comparison of terms both within and across latent topics. Some studies not only visualize the topic model but also refine the topic model via user interactions. Choo et al. [10] introduced UTOPIAN, which employs a modified t-SNE as the visual layout to display documents and topics in a scatter plot. Because topics are obtained by a semi-supervised nonnegative matrix decomposition, users can add constraints to refine the topics. El-Assady et al. [15] presented a modular visual analytics framework, tackling the understandability and adaptability of topic models through a user-driven reinforcement learning process.

Moreover, many studies have analyzed the change in topics over time. ThemeRiver [16] visualizes thematic variations over time within an extensive collection of documents. Liu et al. [31] combined the river metaphor with a word cloud and proposed TIARA to better show the topic meaning of theme rivers. To better display the hierarchy of topics, Dou et al. [13] employed a Bayesian rose tree (BRT) to organize topics into a hierarchical structure and then used a tree map to show the topics. They also applied the hierarchical ThemeRiver to show topics over time. TopicOnTiles [9] reveals the social media information relevant to an anomalous event in a multi-level, tile-based map interface. It adopts the STExNMF topic modeling technique to extract spatiotemporally exclusive topics corresponding to a particular region and time point.

In this paper, we visualize topics to help users visually explore text collections. Documents and topics are displayed in the topic scatter plot via a dimensionality reduction method. We also present label information in the topic scatter plot, unlike UTOPIAN [10].

2.3 Interactive Visual Classification

Several research studies have had users iteratively refine a classification model by labeling new instances or modifying previous classification decisions. Eaton et al. [14] showed a regression model using a 2D scatter plot. The horizontal axis of the scatter plot represents the diversity of documents, and the vertical axis represents the prediction value of the regression function. Each repositioned data instance acts as a control point for altering the learned model, using the geometry underlying the data.

In addition, many studies have integrated visualization with active learning. Berger et al. [2] proposed a 2D scatterplot interface rather than a list-based interface for efficient and effective data annotation. They also proposed a semi-supervised NEC approach to learn custom embeddings for the entities being classified. Seifert and Granitzer [27] integrated visualization with active learning, which can result in a better judgment of whether sample points are outliers or misclassified. Users can use an interactive visualization of the classifiers a-posteriori output probabilities to select a sample to label. Settles [29] proposed a new interactive annotation interface with a novel semi-supervised learning algorithm, DUALIST, which can pose queries on both features (e.g., words) and instances (e.g., documents).

It is difficult for users to label a large-scale text collection. To solve this problem, Seifert et al. [28] displayed clusters of documents and labels by using Information Landscape. A text collection is explored through the cluster hierarchy, and document candidates are found for classification. Similarly, Heimerl et al. [17] obtained an initial annotated training set for classification using the search results of the Apache Lucene framework and discussed three methods for labeling data. Poursabzi-Sangdeh et al. [25] developed an interactive system similar to ours to help users annotate documents: topic models provide a global overview of what labels to create, and active learning guides users to the appropriate documents to label. Moehrmann and Heidemann [23] designed an interface to show clustering results for images and enabled users to quickly and efficiently label a large-scale dataset. Paiva et al. [24] displayed the similarity between images by using neighbor-joining (NJ) trees. Other images can then be labeled according to the NJ trees. To illustrate the impact of visual-interactive labeling approaches, Bernard et al. [3] experimented to compare active learning approaches with visual-interactive labeling approaches. Moreover, Bernard et al. [4] contributed a systematic quantitative analysis of different user strategies when selecting instances for labeling with visual-interactive interfaces.

With the growing adoption of machine learning techniques, there has been a surge of research interest in making machine learning systems more transparent and easier to interpret. Choo et al. [11] employed linear discriminant analysis for classification and dimensionality reduction, and displayed the results through a parallel coordinate plot, a scatter diagram, and a heat map. The display helps users to better understand the meaning of each reduced dimension. Brooks et al. [7] investigated approaches for supporting feature ideation and proposed FeatureInsight, an interactive visual analytics tool for building new dictionary features for text classification problems. To help users assess model performance quickly and accurately, Ren et al. [26] presented Squares, a performance visualization tool for multi-class classification problems. Moreover, Krause et al. [19] proposed a visual analytics workflow to help data scientists and domain experts explore, diagnose, and understand the decisions made by binary classifiers. Ming et al. [22] extracted standardized rule-based knowledge representation from the model's input-output behavior and designed RuleMatrix, a matrix-based rule visualization, to help users navigate and verify the rules and the model.

In our paper, we employ a topic model instead of the hierarchical clustering [28] to help users explore the text collection to create labels and find document candidates for labeling. In contrast to the active learning selection used with ALTO [25], We combine topic overview with visual-interactive labeling. In visual-interactive labeling, users have more control over the sample selection.

3 Task Abstraction and System Overview

This section introduces task abstraction and the pipeline of our system.

3.1 Task Abstraction

To help users explore text collections structurally and find document candidates for classification, we carefully analyzed user requirements. The users of our system are people who need to train classifiers for an unannotated text collection or retrieve documents of interest. Users may be not familiar with the text collection and may not know its categories. Our system should be able to display a summary to help users quickly analyze the text collection. Furthermore, the original text of documents should be shown for users

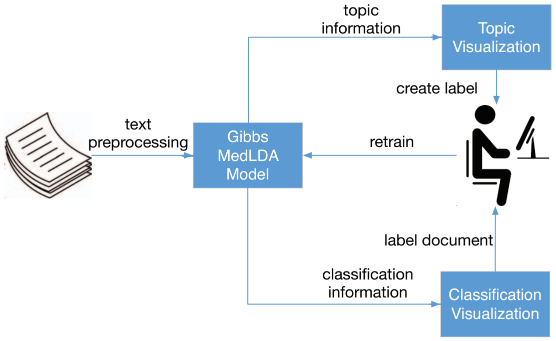


Figure 1: The pipeline of our system. Our system loads the text collection and trains the Gibbs MedLDA to generate topics. The user creates new labels according to the topic visualization and retrains Gibbs MedLDA to refine the classifiers. With the updated topic and classification information, the user continues to label documents to refine the classifiers until a satisfactory result is obtained.

to ensure what labels should be added to documents.

Moreover, it is time-consuming to label all the documents in a text collection. We expect users to label only some document candidates, which significantly improve the performance of classifiers. Additionally, we need to present classification information to help users who have machine learning knowledge to understand, diagnose, and refine the classifiers. The analysis tasks are summarized as follows:

T1 Provide a summary of the text collections so that users can quickly understand its content.

T2 Display information about the classifiers to help advanced users to understand, diagnose, and refine them.

T3 Assist users in finding document candidates to improve the performance of the classifiers.

T4 Show the original text of the documents for users to ensure the accuracy of the document labels.

3.2 System Overview

According to the tasks above, Fig. 1 shows the pipeline of our system, including two parts: the semi-supervised topic model (a modified Gibbs MedLDA), and the visualization component. We first preprocess the text collection by tokenizing the text, removing the stop words, and performing lemmatization. Then, the text collection is represented with the bag-of-words model as the input to the topic model. After the text collection is processed using the topic model, the topic information and classification information are fed to the visualization component. Fig. 2 shows the interface of our system. The visualization component contains three parts: topic visualization, classification visualization, and document visualization. For topic visualization (T1), we design a topic scatter plot to visualize the topic and document distributions, and use a word cloud to show the meaning of the topics. The topic visualization provides a global overview of the text collection. According to the topic visualization, users can create new labels with the corresponding initial training documents and retrain the model. After the model is retrained, the visualization component will update. Users can then find and label the document candidates from the classification visualization and document visualization to refine the Gibbs MedLDA until they are satisfied. For classification visualization (T2, T3), we design a classification scatter plot to display the classification result (Fig. 6) and a topic weight view to help users understand the classifiers (Fig. 2(f)). Users can acquire the classification information from this view, and then refine the classifiers according to their domain knowledge. The label list displays the basic classification result of the classifiers. For document visualization (T3, T4), we provide a text list and a plain text view to help users verify the document labels. The text list shows the document meta information. Moreover, the text list can show uncertain documents that are near the classifier boundary for users to label (Fig. 2(d)). By clicking a document, users can see the original text of a document in the plain text view.

4 Semi-Supervised Topic Model

In this section, we first introduce Gibbs MedLDA [33] and describe how to adapt it as a multi-label semi-supervised model. We also employ the margin-based active learning concept in our model. To better account for topics, we adopt a method to extract topic-related phrases.

4.1 Gibbs MedLDA

Supervised topic models integrate a topic model with a text classification model to generate more discriminative, more suitable topics for text classification. Recent studies show that text classification based on a supervised topic model performs better than that based on a combination of LDA and SVM. Thus, we use the supervised topic model Gibbs MedLDA instead of an unsupervised topic model in our system. Gibbs MedLDA integrates a max-margin classification model with a topic model rather than likelihood-driven objective functions. It employs the Gibbs sampling algorithm, which makes training fast and iterative.

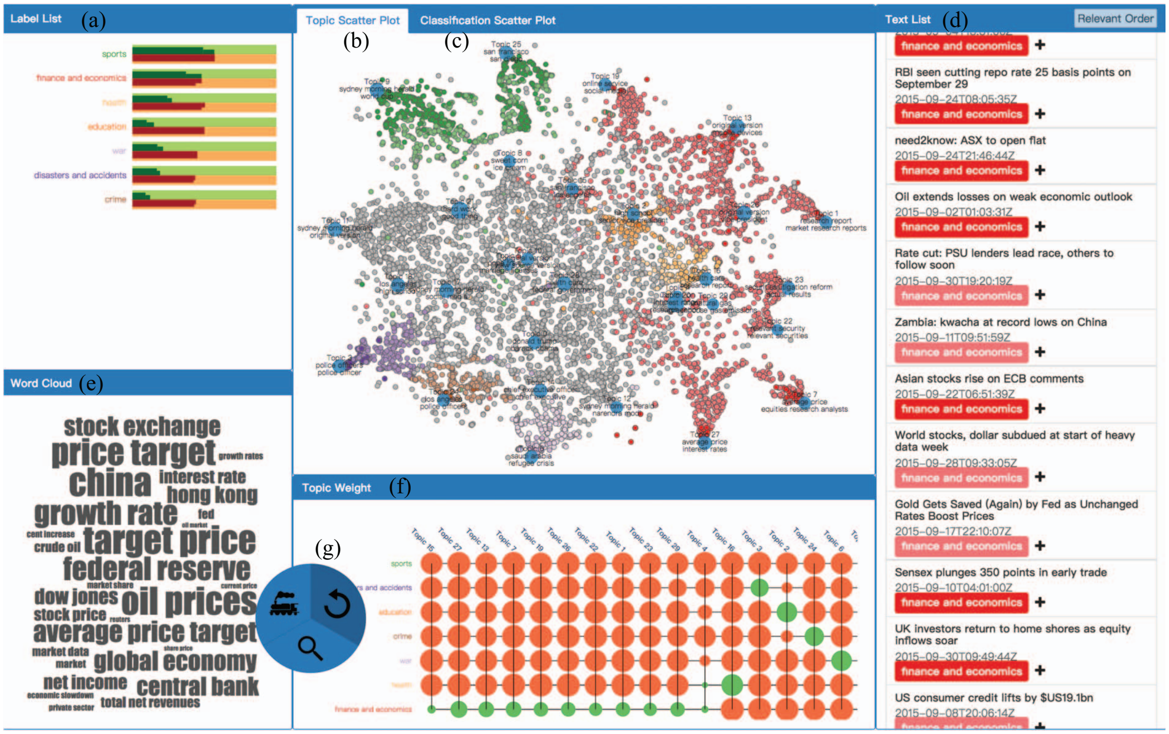
For Gibbs MedLDA, the text collection can be represented as $C = (w_d, y_d)_{d=1}^{D}$, where $w_d = (w_{dn})_{n=1}^{N_d}$ denotes the words appearing in document d, and y_d denotes the label for document d (-1 or $+1$). D represents the document number in the text collection. N_d represents the number of words in document d. Gibbs MedLDA first generates topics from documents, and then a label is generated for each document according to the topics.

For our system, the original Gibbs MedLDA is not suitable for two reasons: a document may have multiple labels, while the original Gibbs MedLDA only classifies with a single label; there are a large number of documents without labels in our system, while the original Gibbs MedLDA model is a supervised learning algorithm.

To address the first point, we extend the Gibbs MedLDA to use multiple labels with the one-vs.-rest strategy. For each label, we apply a two-class classifier. Thus, y_d is a vector instead of a scalar, and $y_d = (y_{dl})_{l=1}^{L}$ denotes the labels appearing in document d. L represents the number of labels in the text collection. To address the second point, we generate topics according to all documents and calculate the classification loss based on the annotated documents. For the documents that are not labeled, we add a new label value 0 ($y_{dl} = -1, 0, +1$). Thus, the topics are generated from the documents. Each document label is generated according to the topics if the label value is not 0.

4.2 Active Learning with Gibbs MedLDA

Users can assign labels to document candidates and retrain the model. Compared with the number of documents, few documents are labeled by users. Documents with a high absolute classifier predicted value are likely to be predicted correctly, especially in the case of negative samples. Moreover, the classification part of Gibbs MedLDA is based on max-margin classification models. The label margin is the prediction value multiplied by the true label value $y_{dl}(\eta_l^T \cdot \bar{z}_d)$. η_l are the parameters of the classifiers in Gibbs MedLDA, and $\bar{z}_d$ is the topic proportion of document d. Topic proportion is the average topic assignment of a document $\bar{z}_d$ or a text collection θ. We calculate classification loss according to the hinge loss $max(0, \ell - y_{dl}(\eta_l^T \cdot \bar{z}_d))$. The ℓ is the margin of the classifiers, which is four in this paper. Therefore, only documents with a label margin less than ℓ affect the classifiers. Thus, documents with a



RBI seen cutting repo rate 25 basis points on September 29
2015–09–24T08:05:35Z
finance and economics +

need2know: ASX to open flat
2015–09–24T21:46:44Z
finance and economics +

Oil extends losses on weak economic outlook
2015–09–02T01:03:31Z
finance and economics +

Rate cut: PSU lenders lead race, others to follow soon
2015–09–30T19:20:19Z
finance and economics +

Zambia: kwacha at record lows on China
2015–09–11T09:51:59Z
finance and economics +

Asian stocks rise on ECB comments
2015–09–22T06:51:39Z
finance and economics +

World stocks, dollar subdued at start of heavy data week
2015–09–28T09:33:05Z
finance and economics +

Gold Gets Saved (Again) by Fed as Unchanged Rates Boost Prices
2015–09–17T22:10:07Z
finance and economics +

Sensex plunges 350 points in early trade
2015–09–10T04:01:00Z
finance and economics +

UK investors return to home shores as equity inflows soar
2015–09–30T09:49:44Z
finance and economics +

US consumer credit lifts by $US19.1bn
2015–09–08T20:06:14Z

Figure 2: Interface of our system using a result from Signal Media news with seven labels. (a) The label list shows the labels in the text collection, and the glyph presents the basic classification results of the labels. (b) The topic scatter plot shows the topic and document distributions. (c) The classification scatter plot displays the classification results of a classifier (Fig. 6). (d) The text list contains documents sorted by uncertainty and supports document labeling. (e) The word cloud shows the keywords of a topic to help users understand the meaning of the topics. (f) The topic weight view helps users understand the classifiers. (g) The right-click menu provides training, keyword search, and undo operations.

high absolute classifier predicted value have little influence on the classifiers if the labels for the documents are predicted correctly. We intend to automatically assign labels to these documents and allow users to pay close attention to documents near the classifier boundary.

Algorithm 1 Margin-Based Active Learning Adapted to Gibbs MedLDA

```
Input: text collection $C = (w_d, y_d)_{d=1}^{D}$.
for t=1 to T do
    Assign labels to documents.
    Train the text collection $C$ via Gibbs MedLDA.
    for each label $l$ in $L$ do
        for each document $d$ in $D$ do
            Predict its label $y_{dl}$, the score is the absolute prediction value $score(y_{dl}) = |\eta_l^T \cdot \bar{z}_d|$.
        end for
        calculate the threshold value $b_l$ according to the score values of annotated documents, $b_l = max(w_1\ell, w_2 mean(score(y_{dl}), \forall d, y_{dl} \neq 0))$, where $w_1$, $w_2$ are the weights to balance values. In this paper, we set $w_1$ as 2, and $w_2$ as 0.8.
        Select a set of documents $C_s = \{dl|score(y_{dl}) > b_l, y_{dl} = 0\}$, and update the labels $y_{dl} = sign(\eta_l^T \cdot \bar{z}_d)$.
    end for
end for
```

We integrate the margin-based active learning algorithm [1] with the Gibbs MedLDA, as shown in Algorithm 1. We average the absolute classifier predicted value of the documents labeled by users, and use this mean value multiplied by a weight as threshold b_l to filter out the unannotated documents with a high absolute classifier predicted value. We then add the classifier predicted labels for these documents to the model and retrain the model. The mean value may be too small. Therefore, we set a minimum threshold b_l according to the margin of classifiers ℓ.

4.3 Topical Phrase Mining

Each generated topic is a distribution over words and is usually represented by the top k words. It would be difficult for users to understand the meaning of topics if only the top single terms were shown. Single terms are often part of indicative phrases, which are lost in a simple unigram representation. Thus, we select noun phrases to interpret the topics. We extract topic-related phrases using a simplified version of an automatic labeling algorithm [21]. We generate candidate phrases by extracting noun phrases chunked by TextBlob and filter out the noun phrases that appear only one time. TextBlob is a Python library for processing textual data. We define the semantic relevance score $Score(P,t)$ of a candidate phrase $P = w_0 w_1 ... w_m$ (w_i is a word) for topic t as follows:

$$Score(P,t) = log\frac{p(P|t)}{p(P)} = \sum_{0 \leq i \leq m} log\frac{p(w_i|t)}{p(w_i)} \qquad (1)$$

where the independence of w_i is assumed, $p(w_i|t)$ is the distribution of words in topic t, and $p(w_i)$ is the distribution of words. $p(w_i|t)$ is the topic result of Gibbs MedLDA. In equation (1), we sum the distribution $p(w_i|t)$ of the words in the phrase. The larger the sum, the more topic-relevant the phrase. $p(w_i)$ is used to correct the bias toward favoring short phrases. Moreover, we expect a good label to have high semantic relevance to the target topic and low relevance to other topics. Therefore, we use the following modified scoring function:

$$\begin{aligned} & Score'(P,t_i) \\ = \; & Score(P,t_i) - \mu Score(P,t_{1,\ldots,i-1,i+1,k}) \\ = \; & (1+\frac{\mu}{k-1})Score(P,t_i) \\ & -\frac{\mu}{k-1}\sum_{j=1,\ldots,k} Score(P,t_j) \end{aligned} \tag{2}$$

We use $Score'(P,t_i)$ to rank the candidate phrases and select several top candidate phrases to represent the topic.

5 Visual Design

In this section, we introduce three major parts of our visualization: topic visualization, classification visualization, and document visualization.

5.1 Topic Visualization

Gibbs MedLDA generates topic distributions over words, the topic proportions of documents, and classifier predicted labels for documents. As shown in Fig. 2, we design a topic scatter plot to show both documents and topics based on the topic distribution.

We show documents and topics simultaneously in the topic scatter plot. The topics are encoded as a $T \times T$ matrix M via one hot encoding. T is the number of topics. Thus, matrix M is a diagonal matrix. Each row of matrix M represents a topic. The topic proportions of the documents $\bar{z}$ is a $D \times T$ matrix. We concatenate these two matrices together as a $(D+T) \times T$ matrix. We then employ t-SNE [30] to reduce the size of the matrix from $(D+T) \times T$ to $(D+T) \times 2$. Thus, the positions of the documents and topics are calculated simultaneously through t-SNE, and the positions are initialized by PCA. We choose t-SNE rather than other dimensionality reduction methods because t-SNE reveals data that lie in multiple different manifolds or clusters.

We use a circle to represent a document, and each sector of the circle represents the labels for the document. Different colors are used to distinguish different labels. Moreover, users may be interested in whether a document is labeled or not, and in the classifier predicted labels of the documents. Thus, we use two similar colors to represent a label: the lighter color for the classifier predicted positive labels, and the darker one for the user annotated labels. The gray circles represent the documents that do not contain any labels. In addition to the documents, the topics are also shown as larger blue circles with several keywords in the topic scatter plot. The size of a topic circle is based on the proportion of that topic in the text collection. It may help users to judge the topic of a cluster of documents by showing the documents and the topics in the topic scatter plot at the same time. Topic information can help users understand the topic distribution in the text collection. According to the topics, users can find corresponding documents in the topic scatter plot.

The topic distribution is updated after each training. To ensure the continuity of the t-SNE result and reduce the time cost, we use the previous t-SNE result as the initial value, and run t-SNE for ten iterations, as ten iterations are adequate to reach convergence.

Because the topic scatter plot shows only a small number of keywords, users may not be able to understand the topic meaning accurately. Thus, we provide a word cloud view in our system. The size of a word represents the score value of each phrase, for which the calculation is shown in Section 4.3. Users can browse the keywords for every topic by clicking the topic circle in the topic scatter plot.

5.2 Classification Visualization

The classification result of Gibbs MedLDA is the classifier predicted value of documents y_d. The classification scatter plot

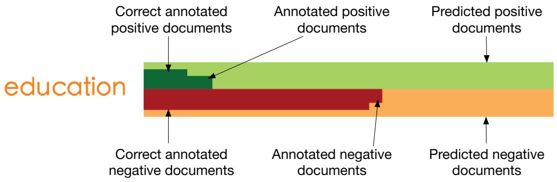

Correct annotated positive documents
Annotated positive documents
Predicted positive documents
education
Correct annotated negative documents
Annotated negative documents
Predicted negative documents

Figure 3: Glyph of the education classifier. The green and red color represent positive and negative documents, respectively. The proportion of user annotated positive or negative documents (the dark color) are compared with the proportion of classifier predicted positive or negative documents (the lighter color) via the length of the bars. The bars with greater height show the proportions of user annotated documents that are correctly predicted by classifiers.

presents the classification result of a label, as shown in Fig. 6. The classification scatter plot view and topic scatter plot view can be switched using the navigation tabs, as shown in Fig. 2.

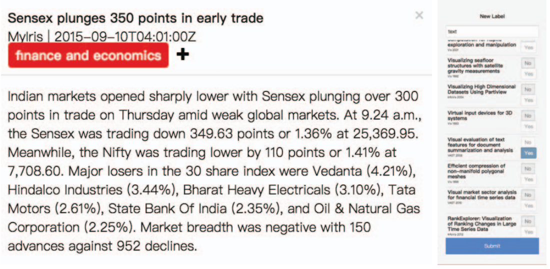
The classification scatter plot view is divided into two parts. The green area represents the classifier predicted values that are larger than zero, and the red area represents the classifier predicted values that are less than zero. Therefore, the vertical axis in the view presents the classifier predicted value. The horizontal axis represents the diversity of documents, which is projected by t-SNE based on the topic proportions of documents. The color of the circle is the same as in the topic scatter plot. The darker grey color represents user annotated negative labels.

Because the classification scatter plot can only show the classification result of a label, we add a label list to display the basic classification information for the labels. The label list shows the proportions of user annotated positive or negative documents compared with the proportions of predicated positive or negative documents, and the proportions of user annotated documents that are correctly predicted. We use a simple glyph to display the classification result, as shown in Fig. 3. The green color represents positive label information. The red color represents negative label information. The lighter color represents classifier predicted positive or negative documents, and the darker color represents the user annotated documents. The part with a darker color area of greater height represents user annotated documents that correctly predicted by classifiers, which means that the classifier predicted label is in accordance with the user annotated label. The width of the area represents the proportion of the corresponding documents.

Moreover, we also provide a topic weight view to show the classifier-topic relationship to help users understand the classifiers. As shown in Fig. 5, we display the weights of the classifiers in the form of a matrix view [12]. Each row represents a classifier, and each column represents a topic. Each circle represents the topic weight of a classifier, and the green and red colors respectively represent positive and negative values. The size of the circle represents the classifier predicted value. We reorder the rows and columns according to their similarity scores by applying the Bond Energy Algorithm [20]. As a result, similar classifiers and topics are shown in closer proximity.

5.3 Document Visualization

To help users to explore the text collection, we provide a text list and a plain text view. The text list (Fig. 2(d)) displays a list of documents, including titles and other meta information. Users can click a label in the label list or a topic in the topic scatter plot to explore the corresponding documents. For topics, we sort the documents according to the topic correlation calculated by $\bar{z}_{dk}/\theta_k$. The $\bar{z}_{dk}$ is the proportion of topic k in document d, and we use the proportion of topic k in the text collection θ_k to correct the bias toward favoring topics of low proportion in the text collection. For labels, users can select relevant documents, irrelevant documents, and uncertain



Sensex plunges 350 points in early trade

Mylris | 2015–09–10T04:01:00Z

finance and economics +

Indian markets opened sharply lower with Sensex plunging over 300 points in trade on Thursday amid weak global markets. At 9.24 a.m., the Sensex was trading down 349.63 points or 1.36% at 25,369.95. Meanwhile, the Nifty was trading lower by 110 points or 1.41% at 7,708.60. Major losers in the 30 share index were Vedanta (4.21%), Hindalco Industries (3.44%), Bharat Heavy Electricals (3.10%), Tata Motors (2.61%), State Bank Of India (2.35%), and Oil & Natural Gas Corporation (2.25%). Market breadth was negative with 150 advances against 952 declines.

New Label

text

exploration and manipulation — Vis 2001 — Yes

Visualizing seafloor structures with satellite gravity measurements — Vis 1992 — No Yes

Visualizing High Dimensional Datasets Using Partiview — InfoVis 2004 — No Yes

Virtual input devices for 3D systems — Vis 1993 — No Yes

Visual evaluation of text features for document summarization and analysis — VAST 2008 — No Yes

Efficient compression of non-manifold polygonal meshes — Vis 1999 — No Yes

Visual market sector analysis for financial time series data — VAST 2010 — No Yes

RankExplorer: Visualization of Ranking Changes in Large Time Series Data — InfoVis 2012 — No Yes

Submit

(a) (b)

Figure 4: (a) Plain text view showing the original information for a document. (b) New label view containing a new label with an initial training set.

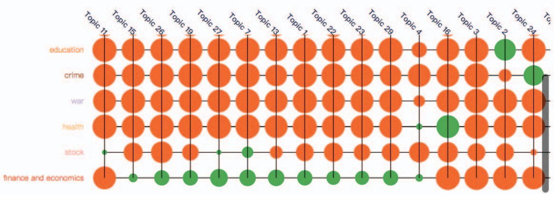


Figure 5: Topic weights for each classifier, which can be used to suggest documents in which topics that need to be labeled by users.

documents, which are calculated by the classifier predicted value. Users can browse the text list and click an item in the list to see the full text in the plain text view (Fig. 4(a)). When users find a document candidate, they can click the add button to label the document or move over the existing labels to remove labels associated with it. If there are no suitable labels for the document, users can create a new label. Our system intelligently suggests a series of documents from different topics, in addition to the selected document, and allows users to judge whether they are related, as shown in Fig. 4(b). Thus, we create an initial training set with positive and negative documents for the newly-built label.

5.4 Interactions

In addition to the views and interactions mentioned above, we further provide a set of interactions to help users explore the text collection for text classification. The user can obtain document information for a document circle in the scatter plots by moving over or clicking the document circle. When the user browses the text list, our system will highlight the corresponding document circle in the scatter plot. We provide zoom and pan interaction for the scatter plot. Our system also supports training, keyword search, and undo operations in the right-click menu, as shown in Fig. 2(g).

We provide two ways to modify the document labels. One is by directly modifying the labels in the text views, and the other is by dragging the document circle in the classification scatter plot to another area. When the user labels documents, she/he can click the training button to retrain the Gibbs MedLDA topic model.

In the beginning, our system loads the text collection and displays its topic overview. Users then create labels according to the topic scatter plot and word cloud. After retraining, our system visualizes the classification information via the classification scatter plot and topic weight view. Users can check the classifier boundary in the classification scatter plot and the topic weights of the classifiers in the topic weight view, and then label the document candidates. Moreover, users can select a label from the label list and label uncertain documents in the text list.

6 Case Studies

We use two unannotated text collections to demonstrate the usability of our system. Our first text collection is the visualization publication data collection [18], which contains IEEE Visualization/VIS publications from 1990 to 2014. This text collection has 2,592 documents, 6,310 words, and 214,917 tokens. The second text collection is a subset of the Signal Media dataset of one million news articles. Most of the articles are English, but non-English and multilingual articles are also included. The sources of these articles include major publishers, such as Reuters, in addition to local news sources and blogs. We randomly select a sample of English articles for our case study, which has 7,033 documents, 3,0158 words, and 1,434,270 tokens.

The number of topics is difficult to determine for an unknown text collection. Users usually select a large number. The larger the number of topics, the less text information is lost. However, a large number of topics is difficult to visualize, because it may easily prevent users from perceiving useful information. We tested several values for the number of topics and found that 30 was the suitable number for Gibbs MedLDA. First, we run Gibbs MedLDA for 100 iterations to train the topics. At each retraining, we run Gibbs MedLDA for ten iterations, as the likelihood function shows almost no change with a higher number of iterations.

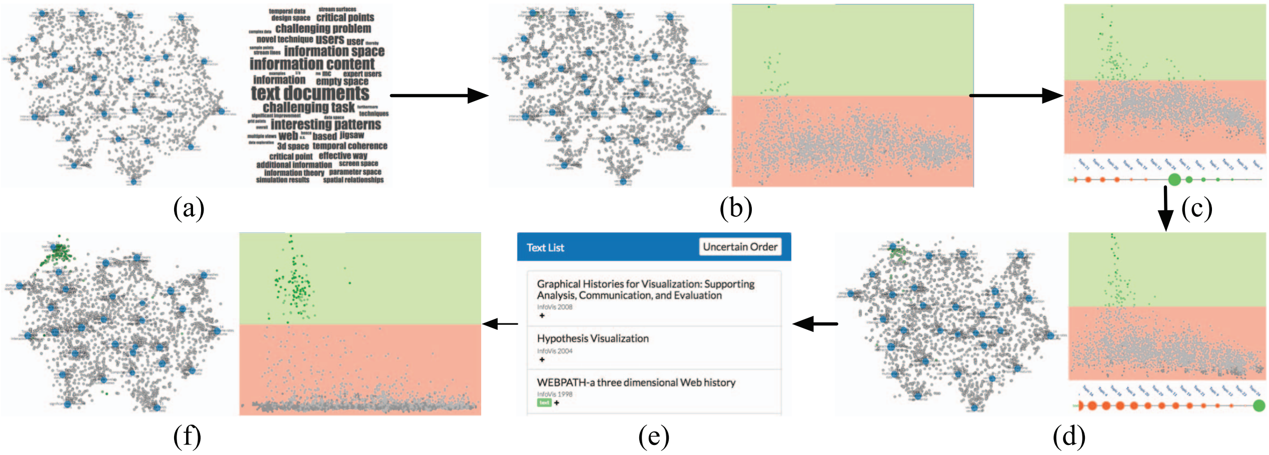
Our case studies were performed on a MacBook Pro with an Intel Core i5 CPU and 8 GB memory. For IEEE visualization publications, it takes 5.34 s for the retraining, and 17.03 s for dimensionality reduction. For Signal Media news, it takes 32.17 s for retraining, and 61.48 s for dimensionality reduction.

6.1 IEEE Visualization Publications

We use the visualization publication data collection to test whether our system can retrieve related papers according to a few annotated papers. Firstly, we load the visualization publication data and train the Gibbs MedLDA to generate 30 topics. Fig. 6(a) shows the summary of the text collection, where two keywords for each topic are displayed in the topic scatter plot. We browse keywords in the topic scatter plot and click the topic circle to verify the topic meaning in the word cloud view. We find that topic 24 is about text visualization. We select this topic and check the word cloud and the text list to see whether they indicate that the topic is about text visualization. We then select a paper about text visualization from the text list and create a new label (text data). After that, our system automatically recommends papers from different topics for us to label, as shown in Fig. 4(b). After we label these papers, our system shows the classification information in Fig. 6(b). In this figure, there is only one user annotated positive paper. We check the boundary of the classifier and label some positive papers. The result after retraining is shown in Fig. 6(c). The number of classifier predicted positive papers is increased. In addition, the diversity of positive papers is increased. Therefore, we select the topics with a positive weight, except topic 24. We find some papers show prediction errors, such as visualizations about graphs, time series data, and volume data. We then fix the labels of these papers and retrain the model. The result is shown in Fig. 6(d). The diversity of positive papers is decreased, and only topic 24 has a positive weight. Then, we select the ”text” label in the label list and order the papers by uncertainty in the text list in Fig. 6(e). We view the papers in the text list and label those papers. After several iterations, we find that all classifier predicted positive papers are about text visualization. We check the classifier boundary in the classification scatter plot. The classifier boundary shows good separation between the positive and negative papers, as shown in Fig. 6(f). Thus, the classifier can correctly distinguish papers about text visualization from other papers.

6.2 Signal Media News

For Signal Media news, we do not know the categories of the news collection. The topic summary of the text collection is shown in Fig. 7. We create labels while browsing topics. The first created label is sports. The classification result obtained after we create



Graphical Histories for Visualization: Supporting Analysis, Communication, and Evaluation
InfoVis 2008
+

Hypothesis Visualization
InfoVis 2004
+

WEBPATH-a three dimensional Web history
InfoVis 1998
+

Figure 6: Visual exploration process of IEEE visualization publication.

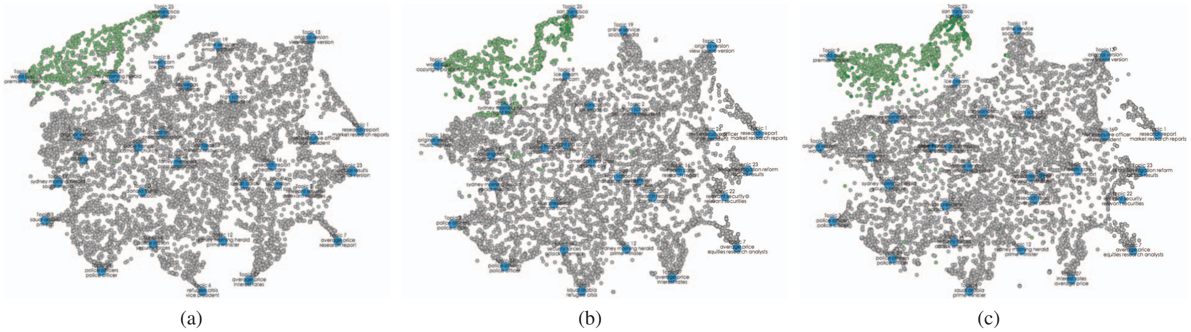


Figure 7: Classification result of sports labels in Signal Media news. (a) Creating sports labels. (b) Labeling documents around the classifier boundary. (c) Unrelated documents labeled as negative for topic 21.

the sports label is shown in Fig. 7(a). We find that some document circles in the cluster are not predicted as positive. Therefore, we click the circles to view the text and label the documents. We then retrain the model, and the result is shown in Fig. 7(b). We find that the classifier predicts some documents positive for topic 21. Therefore, we click topic 21 to check whether the topic is about sports. We find that some documents are about sports, and some documents are personal stories. We add the correct labels to the classifier mispredicted documents and retrain the model. The updated result is shown in Fig. 7(c).

We then add six labels to the text collection, as shown in Fig. 2. Documents may have multiple labels. For example, a label (finance and economics, for example) contains many topics. Thus, we want to add more detailed labels to further subclassify the documents. We find that the keywords of topic 7 and topic 27 are average price, equities research analysts, and Sydney Morning Herald. Therefore, we guess that these two topics are relevant to stocks. Additionally, we analyze some documents near topic 7 and topic 27 by viewing the detailed content. It verifies our guess. We create a new label (stock) and use it to label the relevant documents. Good results are obtained after the initial training dataset is labeled. We then check the topic weight view and find three topics (7, 11, 27) that have positive weights, as shown in Fig. 5. We view the phrases and documents for topic 11 and find that topic 11 is not related to stocks. We then click topic 11 and add correct labels to the classifier mispredicted documents related to that topic. We then retrain the model, and the weight of topic 11 becomes negative. Finally, we refine the classifier by labeling some documents with high uncertainty. The final result is shown in Fig. 8.

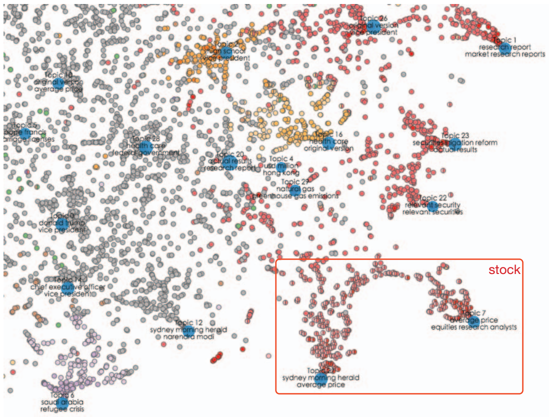


Figure 8: Topic scatter plot of Signal Media news. We add the label stock to further subclassify documents with the label finance and economics.

7 User Study

In this section, we evaluate the usability of our system through a user study. The number of topics and iterations of our system in the user study were the same as in the case studies.

7.1 Evaluation Setup and Procedure

For our user study dataset, we use a sample of Reuters RCV1-V2, which is well-known and widely used as a benchmark corpus with gold labels for text classification. The problem with using a bench-

Table 1: F1 score and label number (the number of correct labels/the total number of labels) of RCV1-V2 classification.

	train			test		
	Sport	Weather	Health	Sport	Weather	Health
linear SVM with all labels	97.76%	76.60%	2.01%	95.72%	61.85%	1.14%
LDA + linear SVM with all labels	95.89%	74.31%	0%	95.13%	68.36%	0%
Gibbs MedLDA with all labels	99.88%	96.43%	93.14%	96.99%	81.26%	68.59%
Active learning linear SVM	93.38%(100/100)	83.27%(100/100)	41.09%(100/100)	91.37%	74.80%	36.65%
Active learning LDA + linear SVM	75.00% (100/100)	60.25%(100/100)	4.31%(100/100)	74.84%	55.40%	3.14%
Active learning Gibbs MedLDA	88.97%(100/100)	60.10%(100/100)	12.15%(100/100)	87.87%	37.82%	4.41%
User 1	95.39%(64/65)	81.00%(71/75)	59.15%(54/58)	94.75%	78.47%	48.05%
User 2	73.94%(107/107)	74.39%(72/77)	56.36%(74/74)	84.50%	74.39%	61.19%
User 3	87.91%(37/37)	59.61%(32/32)	60.38%(39/39)	90.83%	57.37%	55.63%
User 4	94.27%(92/97)	80.68%(163/169)	59.50%(100/111)	95.50%	76.89%	53.74%
User 5	86.55%(37/41)	80.14%(37/40)	62.42%(51/54)	86.56%	80.62%	51.36%

mark corpus is that we need to provide the gold labels and ask users to label the documents according to labels with proven accuracy, rather than allowing them to label the documents according to their judgment. Moreover, we can not ensure that the annotated documents are in accordance with the labels with proven accuracy. The advantage is that we can compare the performance of each participant with that of automatic algorithms.

Reuters RCV1-V2 contains approximately 810,000 Reuters English language news stories from 1996-08-20 to 1997-08-19, which are organized by three different category sets: topics, industries, and regions. Each document is assigned at least one label in each category set. The dataset is divided into four parts: one training dataset, and three testing datasets. The training set contains 23,149 news articles. To reduce the workload for the user study, we use a sample from Reuters RCV1-V2 containing 6,235 documents, with documents related to the topics of government and society. We require users to label three categories of news: sports, weather, and health. There are 197 news articles about health, 135 news articles about weather, and 913 news articles about sports in our training dataset. In addition, to test the precision of the classifiers, we select a sample from the testing dataset that contains 54,631 documents. There are 1,568 news articles about health, 886 about weather, and 8,259 about sports in our testing dataset.

We invited five users to take part in our user study. Most of them are from our visualization group. User 1 and User 2 are studying text visualization. User 3, User 4, and User 5 have no background in visualization or machine learning. We demonstrated the way to use our system and the meaning of each view to each participant. We also showed them how to use our system with the 20 Newsgroups dataset, and we then asked them to label documents themselves to train the classifiers.

7.2 Results and Discussion

Table 1 shows the results of our user study. We show the classification result of each label with the F1 score. We also show the number of correct labels and the total number of labels annotated by users in brackets behind the F1 scores. We add the linear SVM with all labels, LDA + linear SVM with all labels, Gibbs MedLDA with all labels, active learning linear SVM, active learning LDA + linear SVM, and active learning Gibbs MedLDA for comparison. The SVM adopts tf-idf (term frequency-inverse document frequency) features. The penalty parameter C of the SVMs is 1.0. The number of topics in LDA is 30, the same as Gibbs MedLDA. The selection strategy for active learning depends on the predicted value of the classifiers. The size of the initial training set is thirty documents for each classifier, and we add ten documents with the minimum absolute prediction values to the classifier at each iteration. Because for most of the labels, the number of labels annotated by the five users for each classifier is less than 100, we add 100 labels for each classifier through active learning. We run all algorithms 100 times, and we use the mean values as the algorithm result.

From Table 1, we can see that Gibbs MedLDA with all labels achieves the best F1 score. Thus, Gibbs MedLDA obtains better classification performance than the linear SVM and LDA + linear SVM. Moreover, the SVM with all labels does not work well with labels that only have a few related documents in the text collection, especially for the health label. This may be because the positive and negative documents are quite imbalanced. However, the active learning linear SVM shows a high score for the weather and health labels. The active learning selects uncertain documents to learn. The positive and negative documents are not imbalanced in the selected documents. The active learning Gibbs MedLDA and active learning LDA + linear SVM both show poor classification performance. The topic models reduce the number of features from the number of terms to the number of topics. Thus, the model does not easily overfit. However, the classifiers may underfit, and not be able to find suitable features to distinguish the documents. When there are enough label constraints, Gibbs MedLDA yields more discriminative topics for classification.

Although the active learning Gibbs MedLDA has a low score for weather and health labels with one hundred annotated documents, all users show a good classification result with fewer labels. This suggests that our system can help users find document candidates for labeling, which can significantly improve the classification result. Users can use not only the prediction value but also other views and tools to find related documents in our system. The sports category can easily be classified with a few documents and generate a high score for each classifier. For health labels, users perform better than the active learning methods although they labeled fewer labels. However, they do not achieve a higher score than the active learning linear SVM for the weather label. We looked into their annotated documents, and found that some news articles about hurricanes did not belong to weather, such as USA: Hurricane Dolly rakes Yucatan, and MEXICO: Hurricane Dolly bashes into Mexican Gulf coast. The users labeled this news as weather. Thus, the mislabeling rate for the weather label is high, as shown in Table 1. We think this may result in poor classification by users for the weather label.

Fig. 9 shows the F1 score curved lines over the number of labels. From the figure, we see that the curved lines of the training dataset are similar to the testing dataset. This illustrates that the classifiers are not overfitting. Compared with the curved line of the active learning algorithms, the classifiers work better with the help of users than the active learning for the same number of labels. This shows that our method can greatly help users find essential documents to refine the classifiers. The curved lines of User 1, User 2,

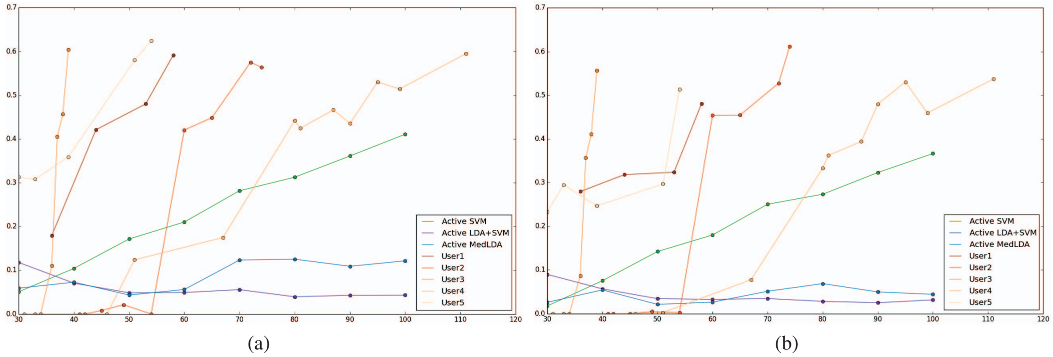


Figure 9: F1 score curved lines over the label number of RCV1-V2 classification. (a) The classification result of the training set. (b) The classification result of the testing set.

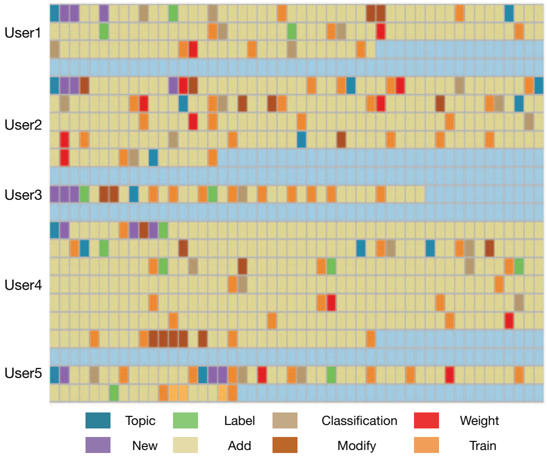


Figure 10: Operation records for five users from the user study. The color of the rectangles represents different operations. It is easy to find new operations and add or modify operations according to the views.

User 4, and User 3 are steep. They labeled a set of documents and trained the classifiers. Then, they labeled some key documents that significantly improved the performance of the classifiers.

To prove the impact of labels on topics, we display the top 10 phrases for topic 20, which is related to the health label, in Table 2. Without any labels, Gibbs MedLDA mixes phrases about health with phrases about countries in topic 20. After the labels are added, the phrases for topic 20 become more consistent. The phrases that are not about health receive a low ranking. Thus, document labels can slightly refine the topics and produce better grouping and separation of the documents.

To determine which interactions provide support for users in making decisions, we recorded user interactions during the user study, as shown in Fig. 10. The color of the rectangles represents different operations. The add, new, and modify operation after which view operation means users make decisions according to that view. We want to know which view the users used most. From Fig. 10, we can see that the users are likely to select a label, then directly label the documents in the text list. Users often use the classification scatter plot. We find User 1, User 2, User 4, and User 5 also use the topic weight view. They use fewer labels and achieve better classification results than User 4. Thus, we think that the topic weight view is useful for refining the classifiers. User 4 labeled many documents, and also modified some document labels at the end of the trial. We did not provide any reset or undo operation. If users label documents wrongly at the beginning, it is difficult for them to refine the classifiers from an error state. User 3 labeled the fewest documents, but the labels were all correct. Therefore, he labeled a few documents but still achieved an acceptable classification result. After the study, users told us that when they misannotated documents at the beginning, it was difficult to find related documents and refine the classifiers. Thus, we added an undo operation to our system. In addition, the topic weight view is useful for refining the classifiers. They browsed topics according to topic weight view and labeled documents to refine the topic weight.

From the above, we propose that the Gibbs MedLDA model used in our system can produce a better result than the linear SVM or LDA + linear SVM. With active learning, Gibbs MedLDA does not work better than the linear SVM, but our system compensates for the problem to some extent. In addition, when users continue to label documents, the Gibbs MedLDA provides more suitable topics for the classifiers. Compared with active learning, our system not only provides uncertain documents in the text list, which is usually provided in active learning, but also provides the diversity of the documents used in the classification scatter plot, and classifier weights in the topic weight view, etc. We provide more information for users to refine classifiers, and they indeed used this information to train the text classifiers. Thus, our system can help users make better decisions than standard active learning. Moreover, we find that users create labels according to the topic scatter plot. In general, our system can help users to find document candidates for labeling and refine the classifiers with fewer labels.

8 Conclusion and Future Work

In this paper, we present an interactive visual analytics system for incremental classification based on Gibbs MedLDA. We change Gibbs MedLDA to a multi-label semi-supervised topic model to meet the requirement of incremental text classification. Moreover, we integrate a margin-based active learning algorithm with Gibbs MedLDA to automatically label some highly scored documents at each iteration. Based on Gibbs MedLDA, we design several views to help users explore the text collection and refine the classifiers. We evaluate our system via two case studies and a user study.

Our system has some limitations. The number of topics must be determined by users at the beginning of the classification task. Moreover, our topic scatter plot is not suitable for showing many topics at the same time. Gibbs MedLDA still does not meet the requirements for real-time interaction. In the future, we will attempt to organize topics into a hierarchy and design new visual encoding to show hierarchical topics.

Table 2: Key phrases for topics related to health. The phrases not related to health are in bold.

	Topic 20
Gibbs MedLDA without any labels	mad cow disease, **european commission**, bovine spongiform encephalopathy, **eu**, **germany**, **britain**, **russia**, chronic fatigue syndrome, **romania**, mad cow crisis
User 1	mad cow disease, mad cow crisis, bovine spongiform encephalopathy, **european commission**, **european union**, blood products, chronic fatigue syndrome, **eu**, **romania**, **britain**
User 2	mad cow disease, mad cow crisis, bovine spongiform encephalopathy, **human rights**, **human rights groups**, **human rights activists**, blood products, heart disease, **european union**, chronic fatigue syndrome
User 3	mad cow disease, mad cow crisis, bovine spongiform encephalopathy, **comprehensive test ban**, **european commission**, **european union**, heart disease, **britain**, chronic fatigue syndrome, blood products
User 4	mad cow disease, mad cow crisis, bovine spongiform encephalopathy, blood products, heart disease, lung cancer, chronic fatigue syndrome, heart attacks, **judith curren**, health risks
User 5	mad cow disease, mad cow crisis, bovine spongiform encephalopathy, **european commission**, heart disease, chronic fatigue syndrome, blood products, **eu**, **britain**, **judith curren**

Acknowledgments

This work was supported by the National Key Research & Development Program of China (2017YFB0202203), National Natural Science Foundation of China (61472354 and 61672452), and NSFC-Guangdong Joint Fund (U1611263).

References

[1] M. Balcan, A. Z. Broder, and T. Zhang. Margin based active learning. In *Proceedings of COLT 2007*, pages 35–50, 2007.

[2] M. Berger, A. Nagesh, J. Levine, M. Surdeanu, and H. Zhang. Visual supervision in bootstrapped information extraction. In *Proceedings of EMNLP 2018*, pages 2043–2053. Association for Computational Linguistics, 2018.

[3] J. Bernard, M. Hutter, M. Zeppelzauer, D. Fellner, and M. Sedlmair. Comparing visual-interactive labeling with active learning: An experimental study. *IEEE TVCG*, 24(1):298–308, Jan 2018.

[4] J. Bernard, M. Zeppelzauer, M. Lehmann, M. Mller, and M. Sedlmair. Towards User-Centered Active Learning Algorithms. *CGF*, 2018.

[5] D. M. Blei and J. D. McAuliffe. Supervised topic models. In *Proceedings of NIPS*, NIPS'07, pages 121–128, USA, 2007. Curran Associates Inc.

[6] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *JMLR*, 3:993–1022, 2003.

[7] M. Brooks, S. Amershi, B. Lee, S. M. Drucker, A. Kapoor, and P. Simard. Featureinsight: Visual support for error-driven feature ideation in text classification. In *2015 IEEE VAST*, pages 105–112, Oct 2015.

[8] A. J. Chaney and D. M. Blei. Visualizing topic models. In *Proceedings of ICWSM*, 2012.

[9] M. Choi, S. Shin, J. Choi, S. Langevin, C. Bethune, P. Horne, N. Kronenfeld, R. Kannan, B. Drake, H. Park, and J. Choo. Topicontiles: Tile-based spatio-temporal event analytics via exclusive topic modeling on social media. In *Proceedings of CHI 18*, pages 583:1–583:11. ACM, 2018.

[10] J. Choo, C. Lee, C. K. Reddy, and H. Park. UTOPIAN: user-driven topic modeling based on interactive nonnegative matrix factorization. *IEEE TVCG*, 19(12):1992–2001, 2013.

[11] J. Choo, H. Lee, J. Kihm, and H. Park. ivisclassifier: An interactive visual analytics system for classification based on supervised dimension reduction. In *Proceedings of IEEE VAST*, pages 27–34, 2010.

[12] J. Chuang, C. D. Manning, and J. Heer. Termite: Visualization techniques for assessing textual topic models. In *Proceedings of AVI*, pages 74–77. ACM, 2012.

[13] W. Dou, L. Yu, X. Wang, Z. Ma, and W. Ribarsky. Hierarchicaltopics: Visually exploring large text collections using topic hierarchies. *IEEE TVCG*, 19(12):2002–2011, 2013.

[14] E. Eaton, G. Holness, and D. McFarlane. Interactive learning using manifold geometry. In *Proceedings of AAAI*, 2010.

[15] M. El-Assady, R. Sevastjanova, F. Sperrle, D. Keim, and C. Collins. Progressive learning of topic modeling parameters: A visual analytics framework. *IEEE TVCG*, 24(1):382–391, Jan 2018.

[16] S. Havre, E. Hetzler, P. Whitney, and L. Nowell. Themeriver: Visualizing thematic changes in large document collections. *IEEE TVCG*, 8(1):9–20, Jan. 2002.

[17] F. Heimerl, S. Koch, H. Bosch, and T. Ertl. Visual classifier training for text document retrieval. *IEEE TVCG*, 18(12):2839–2848, 2012.

[18] P. Isenberg, F. Heimerl, S. Koch, T. Isenberg, P. Xu, C. Stolper, M. Sedlmair, J. Chen, T. Möller, and J. Stasko. Visualization publication dataset. Dataset: http://vispubdata.org/, 2015. Published Jun. 2015.

[19] J. Krause, A. Dasgupta, J. Swartz, Y. Aphinyanaphongs, and E. Bertini. A workflow for visual diagnostics of binary classifiers using instance-level explanations. *CoRR*, abs/1705.01968, 2017.

[20] W. T. McCormick Jr, P. J. Schweitzer, and T. W. White. Problem decomposition and data reorganization by a clustering technique. *Oper. Res.*, 20(5):993–1009, 1972.

[21] Q. Mei, X. Shen, and C. Zhai. Automatic labeling of multinomial topic models. In *Proceedings of ACM SIGKDD*, pages 490–499, 2007.

[22] Y. Ming, H. Qu, and E. Bertini. Rulematrix: Visualizing and understanding classifiers with rules. *CoRR*, abs/1807.06228, 2018.

[23] J. Moehrmann and G. Heidemann. Efficient annotation of image data sets for computer vision applications. In *Proceedings of the 1st International Workshop on VIGTA*, page 2, 2012.

[24] J. G. Paiva, W. R. Schwartz, H. Pedrini, and R. Minghim. An approach to supporting incremental visual data classification. *IEEE TVCG*, 21(1):4–17, 2015.

[25] F. Poursabzi-Sangdeh, J. L. Boyd-Graber, L. Findlater, and K. D. Seppi. ALTO: active learning with topic overviews for speeding label induction and document labeling. In *Proceedings of ACL 2016, Volume 1: Long Papers*, 2016.

[26] D. Ren, S. Amershi, B. Lee, J. Suh, and J. D. Williams. Squares: Supporting interactive performance analysis for multiclass classifiers. *IEEE TVCG*, 23(1):61–70, Jan 2017.

[27] C. Seifert and M. Granitzer. User-based active learning. In *IEEE ICDMW*, pages 418–425, 2010.

[28] C. Seifert, V. Sabol, and M. Granitzer. Classifier hypothesis generation using visual analysis methods. In F. Zavoral, J. Yaghob, P. Pichappan, and E. El-Qawasmeh, editors, *Proceedings of Networked Digital Technologies, Part I*, pages 98–111. Springer Berlin Heidelberg, 2010.

[29] B. Settles. Closing the loop: Fast, interactive semi-supervised annotation with queries on features and instances. In *Proceedings of EMNLP*, pages 1467–1478, 2011.

[30] L. van der Maaten and G. Hinton. Visualizing data using t-SNE. *JMLR*, 9:2579–2605, 2008.

[31] F. Wei, S. Liu, Y. Song, S. Pan, M. X. Zhou, W. Qian, L. Shi, L. Tan, and Q. Zhang. TIARA: a visual exploratory text analytic system. In *Proceedings of ACM SIGKDD*, pages 153–162, 2010.

[32] J. Zhu, A. Ahmed, and E. P. Xing. Medlda: maximum margin supervised topic models. *JMLR*, 13:2237–2278, 2012.

[33] J. Zhu, N. Chen, H. Perkins, and B. Zhang. Gibbs max-margin topic models with fast sampling algorithms. In *Proceedings of ICML*, pages 124–132, 2013.