https://doi.org/10.2352/ISSN.2470-1173.2017.1.VDA-388

Visual Interactive Creation and Validation of Text Clustering Workflows to Explore Document Collections

Tobias Ruppert[1], **Michael Staab**[2], **Andreas Bannach**[1], **Hendrik Lücke-Tieke**[1], **Jürgen Bernard**[2], **Arjan Kuijper**[1,2], **Jörn Kohlhammer**[1,2]
[1] **Fraunhofer Institute for Computer Graphics Research IGD, Darmstadt, Germany**
[2] **Technische Universität Darmstadt, Germany**

Abstract

The exploration of text document collections is a complex and cumbersome task. Clustering techniques can help to group documents based on their content for the generation of overviews. However, the underlying clustering workflows comprising preprocessing, feature selection, clustering algorithm selection and parameterization offer several degrees of freedom. Since no "best" clustering workflow exists, users have to evaluate clustering results based on the data and analysis tasks at hand. In our approach, we present an interactive system for the creation and validation of text clustering workflows with the goal to explore document collections. The system allows users to control every step of the text clustering workflow. First, users are supported in the feature selection process via feature selection metrics-based feature ranking and linguistic filtering (e.g., part-of-speech filtering). Second, users can choose between different clustering methods and their parameterizations. Third, the clustering results can be explored based on the cluster content (documents and relevant feature terms), and cluster quality measures. Fourth, the results of different clusterings can be compared, and frequent document subsets in clusters can be identified. We validate the usefulness of the system with a usage scenario describing how users can explore document collections in a visual and interactive way.

Introduction

The volume of digitally available textual data is continuously increasing. Examples for document collections include newspaper articles, scientific papers, technical reports, patents, legislative documents or social media entries like tweets, blog posts or customer reviews. These documents are highly relevant for many types of stakeholders like journalists, researchers, political decision makers, and online-shop customers. Methods from information retrieval are the means of choice, if stakeholders can specify their information need precisely, e.g., by formulating a *search* query. However, these fact retrieval or known-item search techniques often become ineffective, if document collections are large, complex, or unknown. In such scenarios, the goal to gain an overview of the document collection can be achieved via the *exploration* of structural information within the collection.

The mechanisms needed to enable the exploration of document collections strongly differ from classical search methods [39]. Among others, data aggregation methods support the generation of content-based overviews, by condensing large numbers of documents into a small set of representatives. One of the most prominent classes of aggregation methods is data clustering with its plethora of techniques and its ability to solve various real-world problems. However, analysis approaches based on clustering are confronted with a variety of challenges.

First, clustering algorithms require numerical feature vectors as input, they cannot process unstructured text documents. The definition of an appropriate feature vector representing text documents is a non-trivial task. As a common practice in text analysis, a feature represents a term that occurs in a document and the feature value describes the relevance of the term to the document (cf. vector space model). Using the entire vocabulary of the document collection would result in large feature vectors that are sensitive to noise and inefficient to process. Thus, the size of the feature vector needs to be reduced by selecting a content-preserving feature subset as representatives of the documents. The effective selection of relevant features can be supported by several metrics. However, different metrics may produce different feature rankings. Moreover, the definition of appropriate thresholds for selecting or deselecting features is challenging.

A second challenge is the choice of an suitable clustering algorithm. Different clustering algorithms produce different groupings of objects owed to the fact that they are designed for different problems. Additionally, the results depend on the parameterization of the clustering algorithm. Multiple cluster quality measures exist that allow to quantify the internal quality of a clustering result (e.g. compactness and separation of clusters). However, these measures focus and different characteristics and some of them may even contradict each other. Since, there is no ground truth to measure against, a "best" clustering method does not exist [18]. It is up to the user to evaluate the quality of a clustering depending on the document collection and the analytical task at hand.

This leads to the third core challenge, the comparison of different clustering results. Since no best clustering method exists, users need to be supported in the choice of the most appropriate among several clustering results. Comparing the clustering results based on internal quality metrics is reasonable. However, the comparison of document-cluster affiliations of several clustering results is a non-trivial task, since the analysts are confronted with interesting document subsets distributed over different clusters over different clusterings.

The rationale of this research approach is to formalize the design space for text document clustering processes. The resulting framework builds the basis for the design of text clustering workflows to be applied on document collections in strong accordance to the involved users, data, and tasks. In particular, we aim at enabling analysts without prior knowledge on text analysis to create analytical document clustering workflows. Visualization and interaction techniques from information visualization and visual analytics have proven to ease the access to complex data spaces and analytical models, respectively. Visual comparison and guid-

ance concepts can be applied to make meaningful decisions in the choice of algorithms and parameters. The contributions of our approach are as follows:

1. Feature selection: we present a visual interface for the selection of textual features (terms) from a document collection with the goal to reduce the size of the feature space. Ranking based on the feature selection metrics, and filtering based on the feature types support the users during the selection process. Intermediate feedback is provided to the users by directly displaying selected and deselected features.
2. Cluster analysis: we present a visual interface for the analysis of document clustering results. The clustering can be analyzed based on the cluster contents and cluster quality measures. The content-based perspective is defined by the documents assigned to a cluster and the prevalent features and terms in a cluster. In addition, cluster quality measures support the user in evaluating the cluster's compactness and separation.
3. Cluster comparison: we present a visual interface for the comparison of clustering results. A content-based and quality-metrics-based perspective is provided. Users can identify intersecting subsets that appear throughout several clusterings and inspect the documents and terms contained in theses subsets. Additionally, the cluster quality measures of different clusterings can be compared and the F-measures of the clusterings to a manually defined reference clustering can be inspected.

Related Work

We review visual analytics and information visualization approaches related to document clustering. First, we discuss visual and interactive approaches that support the feature selection process, mostly executed prior to clustering. Second, we review visualization systems that apply clustering or other aggregation techniques to derive structural information from document collections to generate overviews. Third, we provide a short summary on related work about the comparison of clustering results. And finally, we review uncertainty visualization techniques that raise the users' awareness of projection errors.

Feature Selection. Several approaches exist, that address the visual and interactive selection of features. Examples are the work by Guo [15], SmartStripes [28], INFUSE [21], or the Rank-by-Feature framework [36]. We share the idea of defining ranking criteria to enable the reduction of the multi-dimensional feature space. However, these systems differ from our approach, since all of them work on numerical data, while we are focusing on textual features. Most textual features selection approaches only allow users to define thresholds for the feature selection metrics. Features with values beyond these thresholds are excluded from further analysis steps in the clustering workflow (e.g. [8]). Other text clustering approaches use the entire feature space without applying any feature selection mechanism. To the best of our knowledge, no system exists that allows user to visually select textual features.

Document Collection Overviews. Several approaches from the field of visual analytics target the exploration and/or analysis of document collections. Some approaches use meta-information like author, publication year, citations, etc. to group documents. Examples include SurVis [4], CiteRivers [16], PolicyLine [33] and an approach by Oelke et al. [29]. We do not use any meta-information except the document title (if available) but focus on the data content. A class of content-based approaches use a vector space model (consisting of feature term weights) to represent documents. From these vectors, topic models can be extracted in order to structure the document collection. Among others, Latent Dirichlet Allocation (LDA) is a prominent topic modeling approach [5]. Each document is represented by a mixture of topics. The topics are represented by a weighted set of terms. Examples of interactive visualization systems that apply LDA to provide overviews of document collections are ParallelTopics [11], TIARA [38], and TextFlow [10]. All of these approaches are limited to one single clustering algorithm. The experimentation with different clustering techniques and the comparison of differing results is not provided. Moreover, none of the approaches support the visual and interactive refinement of workflow steps, which was one of the goals of our system.

The iVisClustering approach [22] and the UTOPIAN system [7] allow the refinement of topic models via user interaction and visual feedback. In addition, both approaches project documents to the display space. Cluster affiliations are represented by categorical color maps, and weighted topic keywords can be adjusted by the user. While iVisClustering incorporates an enhanced LDA model, the UTOPIAN system introduces an alternative approach for the interactive refinement of topic models, non-negative matrix factorization. Although both approaches support the interactive refinement of the underlying models, they are restricted to only one model. They do not address the comparison of results coming from different models.

We highlight two approaches that provide content-based overviews of document collections via clustering: IN-SPIRE [40], and Overview [6]. IN-SPIRE generates thematic document landscapes by combining document clustering, projection, and keyword extraction. The Overview system organizes document collections in a tree structure based on the results of a hierarchical clustering. While IN-SPIRE has limited interaction and refinement capabilities, the Overview system allows users to document findings by manual tagging. However, both systems rely on a single clustering method, the comparison of clustering results is not addressed.

The Jigsaw visual analytics system supports the exploration of a document collection by extracting entities in documents and analyzing their co-occurrences [13]. In addition, document clustering and document summarization techniques are incorporated. However, the approach differs from ours since it neither addresses the feature selection process nor the comparison of different clustering algorithms.

Visual Cluster Comparison. Our work is related to techniques supporting the visual comparison of multiple clustering results. In our approach, we selected the parallel set visualization to compare document affiliations to clusters from different clustering results [20]. Further visualization techniques that support the comparison of sets are presented in a survey by Alsallakh et al. [2]. The clustering comparison component in our approach is also inspired by XCluSim, a visual analytics tool applied in the area of bioinformatics [26]. The tool allows the comparison of clustering results coming from different clustering algorithms. It uses an enhanced parallel sets visualization that incorporates a tree color

map to allow the identification of related clusters coming from different clustering results. The clusters are colored according to their similarity. Documents are depicted via gray bands between the clusters. Since the main target in our clustering comparison is to identify stable subsets, we follow the parallel set visualization approach (see above).

The paper most related to our work was presented by Choo et al. [8]. It introduces an interactive visual testbed system that allows the definition of dimension reduction and clustering workflows. While the paper primarily focus on the integration of different data types, our system targets textual data. We provide content overviews, summarizing most relevant features in clusters, and cluster subsets. In addition, we also support users in the feature selection phase.

Uncertainty Visualization. Finally, we draw a connection to uncertainty visualization. In our approach documents are projected on the display space and represented as circles to analyze their similarities. In a recent publication by Sacha et al. the role of uncertainty, awareness, and trust in visual analytics is discussed [34]. A comprehensive overview about visualizing geospatial uncertainty is provided by MacEachren et al. [27]. The work summarizes sources of uncertainty and possible techniques for visualizing them. Among others, saturation can be used to depict the uncertainty of objects. This technique is also called pseudo-coloring in a survey about depicting uncertainty in scientific visualization approaches by Pang et al. [31]. We adopt this concept to represent projection errors using a sequential colormap.

Design Considerations

In our approach, we introduce a visual interface for the creation of text clustering workflows with the goal to structure and explore document collections. The targeted *user* group are data analysts. The approach aims at opening up the design space for text clustering workflows, and making them accessible for data analysts. The resulting system should also be applied by users without a specific expertise in data mining, NLP, or statistics. However, prior knowledge about the applied methods is beneficial for the selection of algorithms and the interpretation of results. In a realistic scenario a data analyst will use the system to design an optimal text clustering workflow for a stakeholder with a specific interest in a text document collection. The resulting clustering workflows should answer several questions, that the stakeholder might have specified prior to the design. Examples include: What is the collection about? Which groups of documents emerge? What are the groups about? Are there alternative groupings? How do these groupings differ? Which documents are similar? Why are the documents similar? What is a document about? Keeping these questions in mind, we defined some concrete requirements that should be considered during the design of a visual text clustering system.

First, the desired clustering workflow should heavily rely on the underlying data and task. Therefore, users have to get *access* to the entire clustering workflow. That way, domain knowledge can be incorporated into the analysis process. Workflow steps include text preprocessing, feature selection, clustering specification, analysis of a single clustering result, and comparison of multiple clustering results.

Second, since there is no "best" clustering workflow, users should *analyze* the quality of a clustering result, depending on the underlying data and task at hand. The quality assessment can be supported in two ways: (i) by providing the user overviews on the clusters' content (showing prevalent documents and features/terms), and (ii) by incorporating cluster quality measures in the overviews.

Third, users should be enabled to *compare* several clustering results. This requirement is of key importance to allow the user to decide upon the most appropriate clustering result for the task and data at hand. Similar to the analysis of a single clustering result, the comparison of several clustering results should be based on (i) the cluster content and (ii) the cluster quality measures. To simplify the comparison, the users should be supported in identifying subsets of documents that are constantly grouped together in a cluster across many clustering workflows.

Fourth, to allow the iterative refinement and comparison of clustering results, intermediate results in the workflow should be stored and made accessible for the user in a *history*. This allows to recall and/or refine previous results for comparisons and/or improvements, respectively. The benefits and purposes for data provenance have been presented recently by Ragan et al. [32].

The design considerations can be summarized as follows:

- **DC**$_1$ Access: each workflow step needs to be made accessible to the user. The parameterization of workflow steps should be controlled by the user. Interim results of each workflow step should be presented.
- **DC**$_2$ Analysis: quality of a clustering result should be evaluated by the user. The cluster assignment of documents, the prevalence of feature terms, and the cluster quality documented by cluster measures should help users in their judgment
- **DC**$_3$ Comparison: users should be able to compare different clustering results based on the resulting clusters, and the respective quality measures
- **DC**$_4$ History: to enable an iterative workflow, intermediate analysis results should be stored in a workflow history.

Text Analysis & Clustering Methods

For the realization of our approach, we incorporated techniques from the field of data mining, natural language processing (NLP), and statistics. An overview of the applied techniques is given in Table 1.

We use a vector space model, representing each document with a weight vector [24]. The dimensions in the vector represent unique terms (features), the weights are calculated based on the term frequency-inverse document frequency (tf-idf) in the underlying document. The resulting vectors are used in the text clustering workflow, e.g., to calculate the similarity between documents.

Preprocessing. To generate the vector space model for a document collection the originally unstructured texts is preprocessed. Preprocessing includes (a) optional stop word removal, (b) optional punctuation removal, (c) optional stemming, (d) the extraction of single terms, 2-grams, and 3-grams, (e) part-of-speech tagging (POS), and (f) named entity recognition.

Feature selection. Since the vector space model contains the entire vocabulary of the document collection, the resulting feature vector might be very large. Therefore, feature selection is applied to reduce the dimensionality of the model. The features are ranked based on feature selection metrics to support the user in the selection. We included three commonly applied met-

feature selection	clustering specification	cluster representation	document projection
feature selection metrics:	**clustering methods:**	**content-based cluster**	**projection method:**
document frequency (df)	k-means++	**representation:**	MDS
term frequency-inverse	hierarchical clustering	document affiliation	Sammon mapping
document frequency (tf-idf)	power iteration clustering	most frequent (tf) or correlated (χ^2)	**projection error:**
term contribution (tc)	bisecting k-means	cluster terms and features	neighborhood preservation
feature type-filters:		feature type-filters:	trustworthiness
Part-Of-Speech (POS) tagging		POS, named entities	
named-entity recognition		**cluster quality measures:**	
token, 2-gram, 3-gram extraction		compactness, separation	
		Dunn & Davies-Bouldin index	

Table 1. *Integrated methods and metrics: NLP, clustering, and projection methods are incorporated in the text clustering workflow. Additional feature selection, cluster quality, and projection error metrics support the user in the creation and validation of the workflow.*

rics: term frequency-inverse document frequency (tf-idf), document frequency (df), and term contribution (tc) [25]. In addition to the ranking, filters can be applied on the features. In our approach, we incorporated (a) POS filtering, (b) named entity filtering, and (c) token, 2-gram, and 3-gram filtering based on respective extraction techniques. After the feature selection step, a document is represented by the reduced feature vector with weights defined by the tf-idf.

Clustering. Clustering algorithms require the definition of document similarity. We apply the cosine distance between the documents' feature vectors. We incorporated four clustering methods which are often applied for clustering documents: k-means++ [3], hierarchical clustering, bisecting k-means [37], and power iteration clustering (PIC) [23]. For the evaluation of the clustering results, we apply four cluster quality measures: compactness, separation, Dunn- and Davies-Bouldin-index [17].

Feature extraction. To represent the content of a cluster we extract the most relevant and the most frequent cluster terms (or features). The cluster-wide term frequency (tf) is applied to extract the most frequent terms in a cluster. However, terms that are frequent in several clusters are not discriminative. Therefore, we include a second measure to extract terms that highly correlate with a cluster, the χ^2 statistics. To ensure that terms are extracted that occur in the cluster, we only take the positively correlated terms into account. The two measures (tf and χ^2) can be applied on both the reduced feature space (derived from the feature selection step), and the full document vocabulary.

Projection. Finally, we incorporated two projection and layout techniques to provide a visual overview of the documents space: multi-dimensional scaling (MDS) [9] and Sammon Mapping [35]. Due to the curse of dimensionality, the projection error might induce a misinterpretation of the vector space similarities between documents. To make the users aware of these effects, we included two measures representing the projection error: trustworthiness and neighborhood preservation [19].

Visualization System

Our visualization system supports the creation and validation of text clustering workflows to explore document collections. The system was designed based on the requirements presented in the previous sections. The standard text clustering workflow comprising the stages preprocessing, feature selection, clustering method selection and parameterization, and cluster analysis was expanded by an additional stage, which allows users to compare several clustering results (see Figure 1). In our approach, the text clustering workflow is grouped into three stages of which each is presented in a separate view: the Feature Selection View, the Cluster Analysis View, and the Clustering Comparison View. Details about these interfaces will be provided in the following sections.

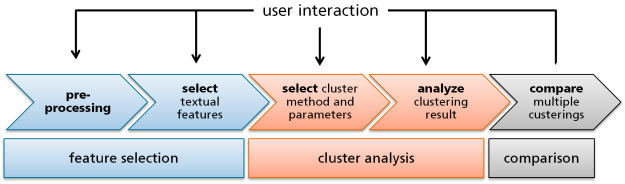

Figure 1. *Text clustering workflow: the standard workflow comprising preprocessing, feature selection, clustering selection, and cluster analysis is expanded by the clustering comparison step. The complete workflow is covered by three views: the Feature Selection View, Cluster Analysis view, and Clustering Comparison View.*

Feature Selection View

The Feature Selection View (see Figure 2) supports the user in the visual selection of features ($\mathbf{DC}_1$). To prepare the feature selection, preprocessing on the original documents needs to be executed. Therefore, the user has to generate a new workflow in the *workflow generation panel*. Users can choose whether the preprocessing should include stop word removal, punctuation removal, and stemming. After the preprocessing the new workflow is shown in the *history panel* (see an example in Figure 3 (bottom right)). Here, each workflow is represented by a quadruplet: the workflow ID, the latest workflow step that was successfully executed in this workflow, a copy button that allows users to duplicate existing processes, and the delete button to remove a workflow from the history. The *history panel* is shown in each view, and allows users to switch between different workflows. Moreover, it allows users to define several workflows with differing parameterizations, which can be compared in the Cluster Comparison View ($\mathbf{DC}_4$). In the *feature selection metrics panel* range bar charts represent the feature metrics extracted in the preprocessing step. Three metrics are incorporated: document frequency (df), term frequency-inverse document frequency (tf-idf; here, the tf in the entire document collection is used), and term contribution (tc). To enable the visualization of large vocabularies, the features are grouped into buckets and mapped on the vertical axis.

Figure 2. *Feature Selection View:* ***Workflow generation panel*** *(top right): preprocessing is executed to generate a new workflow; optional preprocessing steps include stop word removal, punctuation removal, and stemming.* ***Feature selection metrics panel*** *(top): three metrics are shown (df, tf-idf, tc), features are grouped into buckets on the vertical axis, for each bucket the value range of the underlying features is shown on the horizontal axis; users can adjust bucket size and sorting of features (each chart sorted individually, or all charts sorted based on single metric); intersection or union of selected features can be applied.* ***Feature panel*** *(bottom): shows samples of selected and not selected features.* ***Filter panel*** *(middle right): user can filter features based on length (token, 2-gram, 3-gram), named entity types, or parts-of-speech.*

The number of buckets (resolution) can be selected by the user. Each bar represents a bucket by depicting the value range of the feature selection metric on the horizontal axis, from the minimal to the maximal feature value in the bucket. The features on the horizontal axis can be sorted individually, or based on one of the feature selection metrics. Users can select buckets in the chart via a rectangular rubber-band. Intermediate feedback is provided to the user by showing samples of the selected and (unselected) features in the *feature panel* below the respective chart ($\mathbf{DC}_1$). Users may select features from any of the three charts, and decide whether the actual feature subset is defined as a union or intersection. The resulting feature space is shown in the *feature panel*, represented by "selected" and "not selected" features ($\mathbf{DC}_1$). The *filter panel* (below the *workflow generation panel*) allows users to apply additional filters on the features. The filtering of tokens, 2-grams, 3-grams, named entities (e.g. locations, persons, etc.), and parts-of-speech (e.g. nouns, adjectives, punctuations, etc.) are supported. The selected features are used for the representation of documents, and the calculation of document similarities in the subsequent workflow steps.

Cluster Analysis View

In the Cluster Analysis View (see Figure 3), users can define and validate a clustering based on the feature representation selected in the previous view ($\mathbf{DC}_2$). The view is divided into three panels. The *clustering specification panel* allows users to select one of four clustering methods and set its parameterization. After the clustering is executed, the clustering results are shown in the *cluster panel* and the *document projection panel*. In the *cluster panel* the clusters are represented by distinct colors from a categorical color map. The user has several options to explore the content of a cluster by showing (a) the documents in the cluster, (b) the most frequent terms or features in the cluster (tf) (again POS and named entity filters can be applied), (c) the terms or features most correlated to the cluster (χ^2), and (d) the cluster quality, represented by the compactness and separation of the cluster. Users can switch between these options on top of the *cluster panel*. The overall quality of the clustering can be derived from the average compactness and separation, and the overall Dunn and Davies-Bouldin indexes shown in the *clustering statistics panel*. To help the user to examine the clustering result with respect to the similarity of documents, we incorporated an additional visualization. The *document projection panel* in the middle of the Cluster Analysis View shows documents represented as circles projected onto a 2D plane. The projection is derived by executing an MDS on the document's feature vectors. The colors of the dots reflect the cluster affiliation. The projection attempts to minimize the error between the feature vector distances and the Euclidean distances in the 2D panel. Therefore, similar documents are shown close to each other in the visualization. Due to the reduction of a high-dimensional vector to a 2D vector a projection error is introduced to the visualization. Neighborhood preservation and trustworthiness measures shown at the top of the *document projection panel* make the users aware of this fact. By selecting one of the measures, the individual scores are mapped on the color of the document dots via a sequential grayscale color map. The *document projection panel* and the *cluster panel* are linked, documents highlighted or selected in one view, are also highlighted in the other

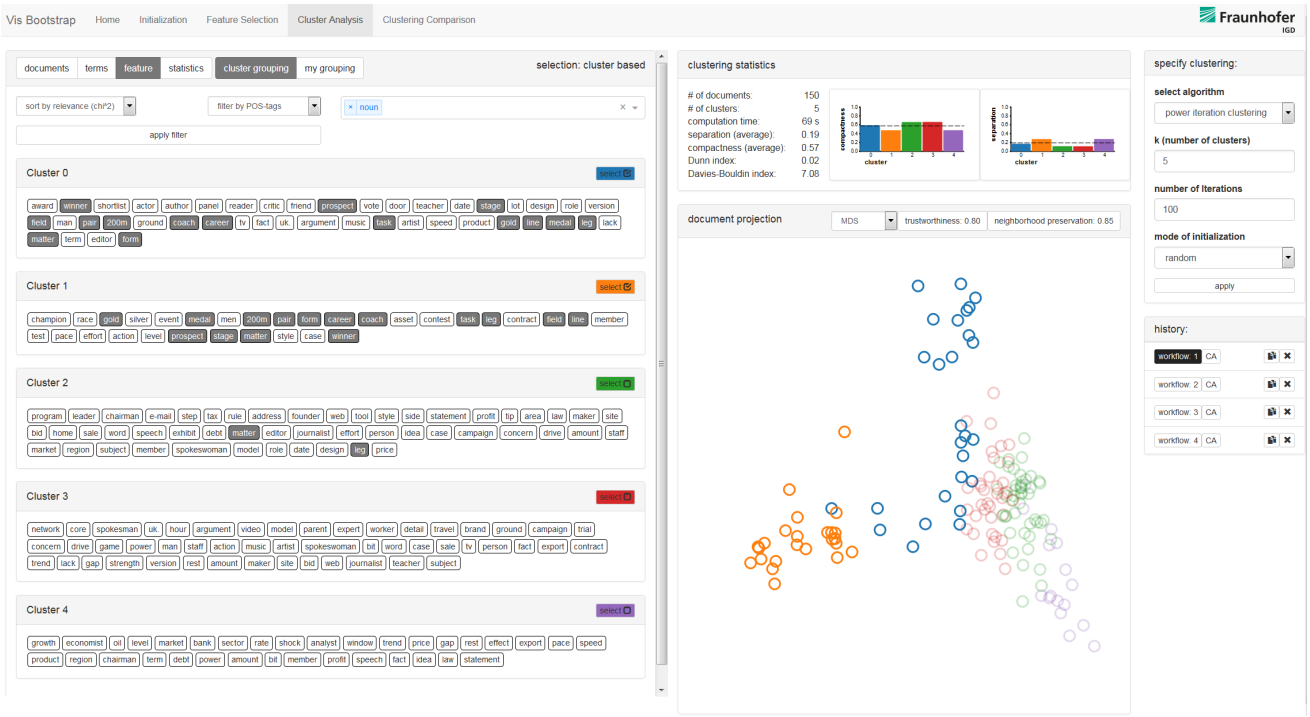

Figure 3. *Cluster Analysis View:* ***Clustering specification panel*** *(right): clustering algorithm and parameters can be defined.* ***Clustering statistics panel*** *(top): cluster quality measures, computation time, and number of documents and clusters are shown; two bar charts show the separation, and compactness per cluster.* ***Document projection panel*** *(bottom): documents are projected on display space based on selected projection method; colors represent cluster affiliation; trustworthiness and neighborhood preservation values are shown.* ***Cluster panel*** *(left): clusters can be represented by affiliated documents, by most frequent (or most correlated) terms, or by most frequent (or most correlated) features; POS- and named entity-filters can be applied on terms and features. Document and term highlighting supports the comparison of clusters. Here, blue and orange clusters are highlighted, most correlated features are shown.*

view. In Figure 3, the "blue" and the "orange" clusters are selected. The respective document dots are highlighted in the *document projection panel*. Moreover, feature terms that occur in both clusters are highlighted. The Cluster Analysis View allows user to evaluate the quality of the clustering result from two perspectives, with respect to the clusters' content and based on the shown cluster quality measures ($\mathbf{DC}_2$).

As an additional method to evaluate the quality of a clustering result, users can manually group documents into clusters. An example is given in on the left side of Figure 4. Here, all documents on "sports" are grouped into Cluster 0. The remaining documents are grouped into an "unassigned" cluster. Users can add further clusters by clicking on the plus button, and add documents to these clusters. The manual document grouping ("my grouping" tab) is used as a ground truth representing the mental model of a user. The quality of a clustering can also be evaluated by measuring how adequate the clustering represents the manual grouping of the user. Therefore, an F-measure is calculated and presented in the Clustering Comparison View, discussed in the following section.

Clustering Comparison View

Finally, the Clustering Comparison View (see Figure 4 and Figure 6) allows the user to compare several clustering results generated by different clustering workflows ($\mathbf{DC}_3$). In Figure 4 the clustering workflow results are compared to the reference clustering, in Figure 6 only the clustering workflow results are shown. The view is divided into three panels. In the *workflow statistics panel*, users can compare the quantitative cluster quality scores of different clustering results. Six bar charts show (i) the the F-measure between the respective clustering and the user-defined "my grouping", (ii) the computation time, (iii) the average compactness, (iv) the average separation, (v) the Dunn index, and (vi) the Davies Bouldin index of the different clustering results. The *workflow comparison panel* in the middle of the view shows a parallel sets visualization [20] that allows users to analyze document-cluster affiliations over several clusterings. In the visualization, each row represents a clustering. The rows are divided into sections of differing lengths, representing the clusters and their respective sizes (number of associated documents). The order of the clusterings can be adapted via drag-and-drop. The clusters of different clusterings are connected via colored bands. The colors of the bands are defined by the clusters of the workflow shown on top of the view (in Figure 6, workflow 1). The colored bands between two rows represent a subset of documents that appear in both clusterings in a single cluster. This helps the user to see overlapping sub-clusters coming from different clusterings. The analysis of stable document subsets that appear in one single cluster in each clustering is supported. Details about these subsets are shown in the *cluster intersection panel*. Subsets can be selected by clicking on the respective band. The user can choose whether the documents or the most frequent (or correlated) terms

Figure 4. *Clustering Comparison View:* ***Workflow statistics panel*** *(right): F-measure (if reference clustering is available), computation time, average separation, average compactness, Dunn and Davies-Bouldin indexes of different clusterings can be compared.* ***Cluster statistics panel*** *(top): clustering statistics of workflow on top of cluster comparison panel are shown.* ***Workflow comparison panel*** *(middle): clustering results (rows) are separated into clusters according to underlying document counts; order of clusterings can be adapted via drag-and-drop; color of bands represent clusters of clustering on top; band width represents number of documents in band.* ***Cluster intersection panel*** *(left): shows documents within a band selected in the cluster comparison panel; alternatively, most frequent or correlated terms can be shown.* ***History panel*** *(bottom right): created workflows are represented by ID, latest workflow step, copy and delete buttons; current workflow is highlighted, via the check boxes users can select workflows to be compared.*

in this document subset are shown.

Discussion on Design Decisions

In the previous section, we described our visual interactive system including visualization techniques to support the exploration of document collections. In this section, we want to briefly discuss the design decisions and chosen visualization techniques. In the Feature Selection View, we needed visualization techniques to (a) show the quantitative feature selection metrics, and (b) the resulting features. We chose a range bar chart and a word list to address these issues, respectively.

Range bar chart (to visualize feature selection metrics): The purpose of the feature selection is to reduce the dimensionality of the vector space model while retaining the quality of the representation. Feature selection metrics can support users in removing non-informative features from the feature space. A visualization technique for the visualization of features selection metrics needs to fulfill the following requirements: (a) features should be sortable based on their feature selection metrics, (b) features should easily be selected or deselected, (c) the visualization should be scalable, since the vocabulary of the entire document collection needs to be represented, (d) the combination of different metrics should be supported. A simple sortable table that shows the features and their metrics is not scalable due to the high dimensionality of the vocabulary. Still, we want to keep the metaphor of a list. Hence, an aggregation of the features is needed. We aggregate the features into buckets of equal size. To visualize the scores, we decided to show the user the range (min-max) of the prevalent features selection metrics within a bucket. We also discussed alternative representations like box plots, or dot plots. However, for box plots and dot plots overplotting becomes an issue if the user increases the resolution to a large numbers of buckets. Dot plots with only one dot representing the average score in a bucket are a further alternative. However, if the resolution is high, it is difficult to spot them in the chart. Moreover, outliers are not covered by the average. For example, users cannot grasp the minimal score within a bucket, which could be important for defining score thresholds. We also discussed alternative aggregation methods. For example, by grouping features based on their feature selection metrics, histograms of the metrics' distributions could be shown. However, this would impede users to estimate the ratio of selected features, while our aggregation method explicitly shows the proportions on the vertical axis. Moreover, our aggregation method allows the combination of feature selection metrics, e.g., by sorting a metrics chart based on another metric. Other aggregation methods would not allow comparisons due to varying bucket sizes. The results of the feature selection in the range bar charts are presented in the word list.

Word list (to visualize features/terms): In several views, we needed a visualization technique to represent relevant features or terms: selected and not selected features in the Feature Selection View; most frequent terms and most correlated features in the Cluster Analysis View, and named entities within cluster intersections in the Cluster Comparison View. We needed a compact representation, since the available space in the view was limited. Moreover, the most relevant terms should be identified quickly. Although, word clouds are a popular tool to visualize text, research has shown that simple tables are the better choice for identifying (a) the presence or absence of terms, and (b) most and least relevant terms (e.g., [30]). Due to the matter of space, we attempt to combine the benefits of both tables and word clouds. Therefore, we sort the terms based on their relevance, and show all terms with the same size to keep the structure of a list, but display them like a space-filling word cloud to reduce white space.

Bar chart (to visualize cluster quality metrics): In the Cluster Analysis View and the Clustering Comparison View, we needed a visualization technique to compare the cluster quality scores of different clusters and clusterings. Research has proven that bar charts are most appropriate for the comparison of quantitative data (e.g. [12]).

Scatterplot projection (to visualize similarity and cluster affiliation of documents): In the Cluster Analysis View, we needed a visualization that intuitively represents the similarity and the cluster affiliation of documents. For this purpose, a similarity matrix or a projection-based scatterplot visualization are appropriate choices. We decided to apply a projection-based scatterplot that shows the similarity between documents by their spatial distance, and the cluster affiliation via a color coding. While the precision of a matrix visualization is higher, a projection-based scatterplot offers a global perspective on the distances. Patterns in a matrix visualization are more difficult to interpret. In a projection view, the user can directly inspect the similarity between documents via the spatial distance. Moreover, it is easier to spot outliers in a cluster.

Parallel sets visualization (to visualize document intersections between clusters from different clusterings): Finally, for comparing different clusterings, we needed a visualization that was capable of showing stable cluster subsets that always appear in the same cluster, independently of the chosen clustering method. Related research can be found in the visualization and comparison of sets. A comprehensive overview of set visualization techniques has been recently published [2]. Out of the proposed techniques, we selected the parallel sets visualization, introduced by Kosara et al. [20], since it fits best to our purpose and is still easy to comprehend. We also discussed the alternative of a heatmap matrix representing clusters of one clustering as rows, and clusters of a second clustering as columns. The cells could represent the number of documents in the intersection via a color map [1]. However, this view would only be suitable for comparing two clusterings. Finally, as discussed in the related work section Lyi et al. present an enhanced parallel sets visualization coloring the clusters across all clusterings and depicting stable subsets via edges between the clusterings [26]. However, we prefer to color the bands instead of the rows in order to better comprehend where specific cluster subsets can be found in the different clusterings.

Usage Scenario

In the following, we demonstrate the usefulness of our visualization system with a real-world dataset. The BBC dataset containing news articles in five topics from 2004-2005 serves as a testcase [14]. We selected a random sample of 150 documents (30 per topic) from the collection. The topic labels are business, entertainment, politics, sport, and tech. Each document title consist of the underlying topic label and a document ID (e.g., tech22). Our system only uses the document content for clustering. In this usage scenario, the document titles and labels will help us to validate the clustering results. In an unlabeled dataset, the user will use his previous knowledge to assess the topics of the documents via their titles or content.

We will structure our analysis process as follows. First, we will select a small subset of the features extracted from the document collection in the preprocessing step. Second, we will run several clustering algorithms on the resulting feature space and analyze the derived clusterings in the document projection panel, observing whether similar documents are associated to the same cluster. We will choose the most promising clustering result and analyze the clusters via most frequent cluster terms and documents in the clusters. Third, we will compare the performance of our feature vector with alternative feature vectors. Therefore, the respective clustering results will be compared based on their internal cluster quality, and towards a manually defined reference clustering. The usage scenario just illustrates one possible way to use the presented system. The order of the analysis steps might be adapted and the process may have more or less iterations.

Feature Selection. As a first step in our usage scenario, we create a new workflow in the Feature Selection View as shown in Figure 2. We keep the preprocessing routines stemming, stop word removal, and punctuation removal activated, since they already help to decrease the feature space. To further reduce the size of the feature vector, we only use nouns as features by applying a part-of-speech filter on the features. The resulting feature quality metrics are shown in Figure 2. Instead of sorting all metric charts individually, we sort them based on the features' document frequencies (df). It can be seen that the document frequency and the term contribution charts show similar shapes. Still, the upper bucket in the term contribution chart also contains small values (as in the tf-idf chart). To focus on features with high scores throughout all metrics, we select the upper buckets in the tf-idf chart excluding the top bucket that also contains low scores. Here, the combination of metrics helps to remove features with low scores from the feature vector. The selection results in 238 selected features which is less than 10% of the total vocabulary with 2482 features. We will evaluate in a later analysis step whether this relatively small feature vector sufficiently represents the document collection. In the *feature panel*, we can get a first notion about the content of the dataset. Features ranked highest are "profit", "revenue", "advertisement", "analyst", "custom", which gives us the notion of a business dataset. However, it is difficult to estimate the coverage of a document collection by looking at a small excerpt of the vocabulary. To get a better overview of the dataset by grouping documents with similar content, we proceed to the cluster analysis step.

Cluster Analysis. We switch to the Cluster Analysis View (Figure 3), and execute several clustering algorithms, since we cannot say yet, which clustering algorithm will perform well with

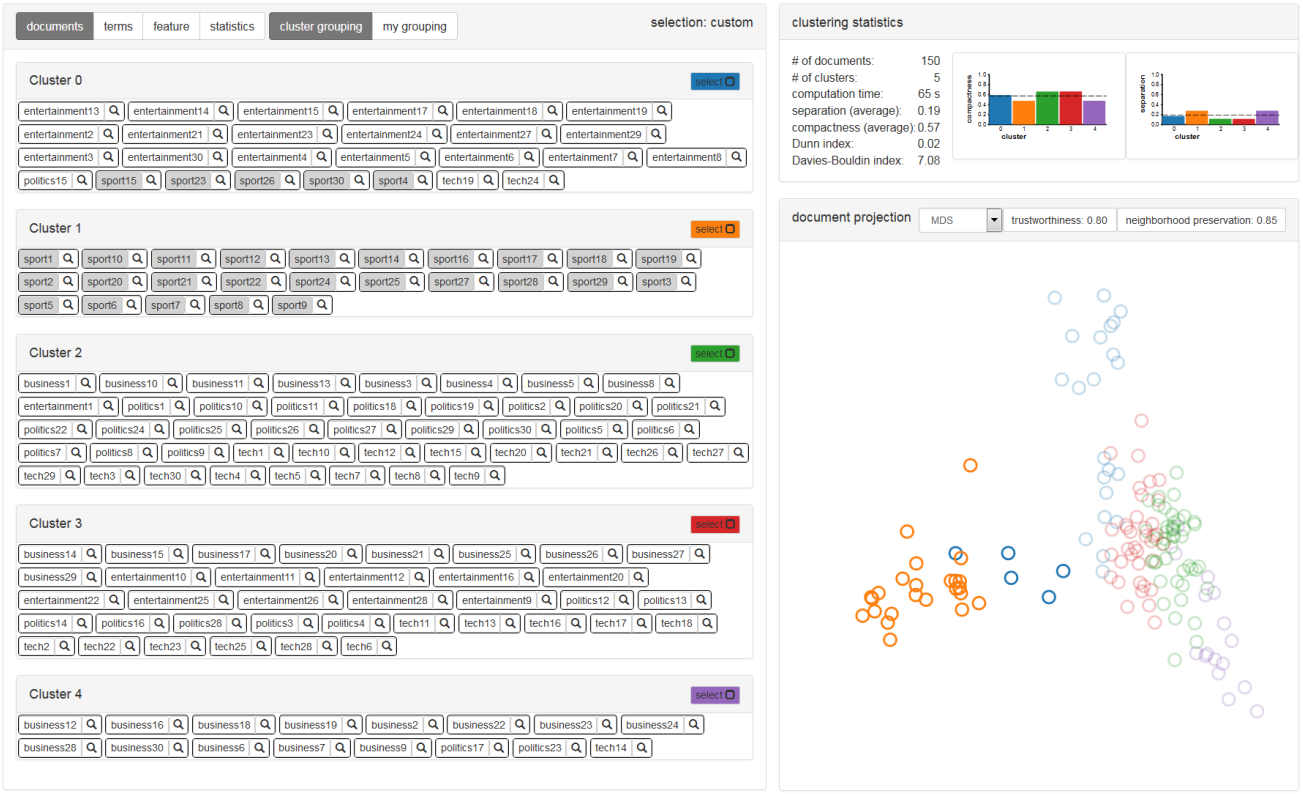

Figure 5. *Cluster Analysis View: in the cluster panel (left), the documents per cluster are represented by their title (in this dataset a combination of genre and an ID). In the document projection panel, documents horizontally separated from the others have been selected. These document are highlighted in the document projection panel (right) and the cluster panel (left). All selected documents are about sports. The "orange" cluster only contains documents on sport. The "blue" is a mixture of topics. The other clusters do not contain documents on sports.*

the selected features. In the *clustering statistics panel* and in the *document projection panel*, we can assess the separation and compactness of the resulting clusters. In this specific scenario, we prefer the clustering results achieved with the power iteration clustering, as shown in Figure 3. In the *document projection panel* the five clusters are depicted via five colors. The orange cluster seems to be well separated from the other clusters. Only a few "blue" documents are close to the "orange" documents. In the *cluster panel*, we inspect the most relevant features per cluster (derived from the χ^2 measure). By selecting the orange and the blue cluster, the underlying documents are highlighted in the *document projection panel*. The relevant features shared by both clusters are highlighted in the *cluster panel*. The orange and the blue cluster have several features in common. Relevant feature terms include "gold", "medal", "200m", and "coach". Further features occurring in the orange cluster are "champion", "race", and "contest". Without knowing anything about the dataset in advance, this gives us the notion that the orange cluster mainly contains documents about "sports". The "blue" cluster contains sportive terms, too, but also terms like from arts and entertainment like "award", "actor", "author", and "tv". From this, we learn that the blue cluster represents more than one theme. The "green" cluster contains political and business terms like "program", "leader", "chairman", "tax", "law". The "red" cluster contains a mixture of terms like "network", "model", "worker", "artist", "music", while the "purple" cluster includes several business terms like "growth", "economist","oil", "market", "bank". We conclude that the orange and the purple clusters provide coherent topics, while the other clusters contain a mixture of themes. This can also be validated by looking at the separation scores of the respective clusters in the *clustering statistics panel*.

As a next step, to confirm our analysis results, we want to inspect the documents per cluster. Therefore, we switch from features to documents in the *cluster panel* (see Figure 5). All documents contained in the clusters are shown represented by their titles (in this scenario including topic labels). Our analysis results presented in the previous paragraph can be confirmed. The "orange" cluster only contains documents about sports. The purple cluster mainly contains documents about business. The blue cluster contains a mixture of sports and entertainment documents.

To further experiment with the dataset, we want to create a "sports" cluster that can be used as a ground truth for alternative clustering workflows. We use the interface depicted in Figure 4 to create a manual grouping with all sports documents in one group and the remaining documents in another group.

Clustering Comparison. Finally, we want to compare the performance of our selected vector space model (cf. Figure 2) with alternative models. Therefore, we create additional workflows by selecting different feature subsets in the Feature Selection View. Table 2 shows an overview of the selected features. In addition to the initial vector space model (workflow 1), we created alternative models using all nouns (2), all single token features

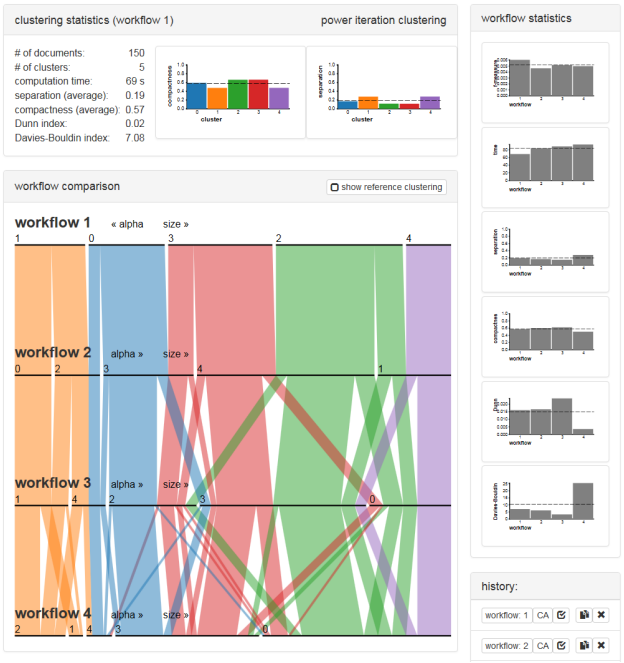

Figure 6. *Clustering Comparison View: The results of four different clustering workflows based on four vector space models are shown. The clustering results have several overlappings in the different clusters. We can assume that the clusterings are already rather stable. The documents contained in the first two orange and the first blue cluster bands can be inspected in Figures 7, 8, and 9. All of them only include documents about "sports".*

(3), and an intersection of the noun features with the best values per feature selection metric (4).

For all of these vector space models we run a power iteration clustering with five clusters, which allows us to analyze the variation induced by different vector space models. In Figure 6, we compare the individual cluster groupings. The clustering results are rather stable. Hence, our initial feature selection including only 238 features (Workflow 1) produces similar results like the full vector space model including all single term tokens (Workflow 3). Moreover, our feature vector consumes less computation time than the full feature vector (see second bar char on the right). By clicking on the first two orange and the first blue bands, we can inspect the documents contained (see Figures 7, 8, and 9, respectively). All of these documents are about "sports".

In Figure 4, we compare the clustering results to the reference clustering that we manually created in the previous analysis step. The orange band represents the documents of the reference cluster - documents on "sports". We can see that in our initial cluster workflow these documents are distributed over two clusters, while the other workflows dispense them into three clusters. By comparing the F-measures in the *workflow statistics panel* (see

ID	description	size
1	noun features with best df scores w/o 1st bucket	238
2	all nouns in vocabulary	841
3	all features in vocabulary	2482
4	intersection of best noun features per metric	243

Table 2. *Usage Scenario: Defined workflows.*

Figure 7. *First cluster intersection band (orange)*

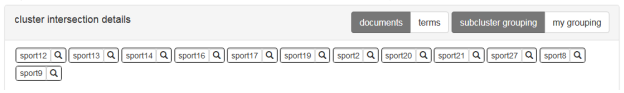

Figure 8. *Second cluster intersection band (orange)*

Figure 9. *Third cluster intersection band (blue)*

Figure 4 (top right)), we can conclude that Workflow 1 represents our reference cluster best. Therefore, we conclude that our selected feature vector represents the reference clustering best, while consuming less computation effort and memory space. To finalize our usage scenario, we report the following findings:

1. By combining POS-filtering and multiple feature selection metrics, we were able to select a relatively small feature vector that already provides satisfying clustering results.
2. Inspecting a sample of the selected features provided us a high-level view on the dataset. Deeper insights could only be gained by exploring the clusters' content.
3. The document projection panel helped us to get a first overview on the quality of the clusters. Overlapping clusters and well separated clusters were found immediately.
4. The workflow comparison panel supported us in identifying stable document subsets throughout several clusterings. Moreover, the computed clustering could be effectively compared with the reference clusterings.

Discussion

We identify three main challenges with potential for future improvements of our work.

Usability. The presented visualization system was designed for analysts that have some experience in NLP, data mining, and statistics. The main goal was to open up the design space for text clustering workflows by providing visual access to crucial steps in the process. However, it would be desirable to reduce the system's complexity to allow a larger user group the exploration of document collections. As a possible solution to address this challenge, we could incorporate workflow patterns into the system that users can select and adapt to their specific scenario. The patterns could be extracted by replicating best practices in the configuration of workflow steps from related research. Additional meta-information should explain the specific characteristics and targets of the underlying configuration to the users.

Scalability. First, so far we tested our system mainly on document collections containing around two hundred documents. For the exploration of larger collections, the scalability of the *document projection panel* needs to be addressed. As an option, the documents could be represented by their cluster centroids. By zooming into the view, the documents contained in the cluster could be shown. Second, if the document collections become larger, the computation time will also increase. To improve the efficiency, and realize user interaction more calculations could be

shifted into the preprocessing phase (if possible). This could also be achieved by calculating the workflow patterns mentioned in the previous paragraph in the preprocessing step.

Expandability. Finally, the visualization system could be expanded by further feature selection metrics, clustering and projection algorithms, cluster quality metrics, projection error measures, and visualization techniques. Moreover, an interface for integrating external algorithms into the system could be envisioned. This would allow researchers to test and evaluate new algorithms with the system.

Conclusion

In this work, we presented a visualization system for the visual interactive creation and validation of text clustering workflows with the purpose to explore document collections. First, we discussed current challenges in the exploration of document collections via document clustering. Second, we presented requirements on the design of a visualization system that addresses these challenges. The requirements were: (1) open up the design space for text clustering workflows by providing visual and interactive access to all workflow steps; (2) support the users' evaluation of the clustering results based on the clusters' content (relevant terms and documents) and cluster quality metrics; (3) also support the comparison of different clustering results on these levels; (4) provide a workflow history that allows to switch back and forth between different clustering workflows. Based on these design considerations, we introduced a visualization system structured into three views to support each step of the text clustering workflow: (a) preprocessing and features selection, (b) cluster definition and analysis, and (c) clustering comparison. We also provided our design rationale discussing the choice of the visualization techniques integrated in the system. To underline the usefulness of the system, we showed its applicability in a usage scenario that targeted the exploration of BBC news articles. Finally, we discussed limitations of our approach and highlighted further challenges to be addressed as future work.

References

[1] Bilal Alsallakh, Wolfgang Aigner, Silvia Miksch, and Helwig Hauser. Radial sets: Interactive visual analysis of large overlapping sets. *IEEE Transactions on Visualization and Computer Graphics (TVCG)*, 19(12):2496–2505, 2013.

[2] Bilal Alsallakh, Luana Micallef, Wolfgang Aigner, Helwig Hauser, Silvia Miksch, and Peter Rodgers. The state-of-the-art of set visualization. *Computer Graphics Forum*, 35(1):234–260, 2016.

[3] David Arthur and Sergei Vassilvitskii. k-means++: The advantages of careful seeding. In *ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pages 1027–1035. Society for Industrial and Applied Mathematics, ACM, 2007.

[4] Fabian Beck, Sebastian Koch, and Daniel Weiskopf. Visual analysis and dissemination of scientific literature collections with survis. *IEEE Transactions on Visualization and Computer Graphics (TVCG)*, 22(1):180–189, 2016.

[5] David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022, 2003.

[6] Matthew Brehmer, Stephen Ingram, Jonathan Stray, and Tamara Munzner. Overview: The design, adoption, and analysis of a visual document mining tool for investigative journalists. *IEEE Transactions on Visualization and Computer Graphics (TVCG)*, 20(12):2271–2280, 2014.

[7] Jaegul Choo, Changhyun Lee, Chandan K. Reddy, and Haesun Park. UTOPIAN: User-driven topic modeling based on interactive nonnegative matrix factorization. *IEEE Transactions on Visualization and Computer Graphics (TVCG)*, 19(12):1992–2001, 2013.

[8] Jaegul Choo, Hanseung Lee, Zhicheng Liu, John Stasko, and Haesun Park. An interactive visual testbed system for dimension reduction and clustering of large-scale high-dimensional data. In *Proceedings of SPIE, Visualization and Data Analysis*, volume 8654. SPIE, 2013.

[9] Trevor Cox and Michael Cox. *Multidimensional scaling*. CRC press, 2000.

[10] Weiwei Cui, Shixia Liu, Li Tan, Conglei Shi, Yangqiu Song, Zekai Gao, Huamin Qu, and Xin Tong. Textflow: Towards better understanding of evolving topics in text. *IEEE Transactions on Visualization and Computer Graphics (TVCG)*, 17(12):2412–2421, 2011.

[11] Wenwen Dou, Xiaoyu Wang, Remco Chang, and William Ribarsky. Paralleltopics: A probabilistic approach to exploring document collections. In *IEEE Conference on Visual Analytics Science and Technology (VAST)*, pages 231–240. IEEE Computer Society, 2011.

[12] S. Few. *Now You See it: Simple Visualization Techniques for Quantitative Analysis*. Analytics Press, 2009.

[13] Carsten Görg, Zhicheng Liu, Jaeyeon Kihm, Jaegul Choo, Haesun Park, and John Stasko. Combining computational analyses and interactive visualization for document exploration and sensemaking in jigsaw. *IEEE Transactions on Visualization and Computer Graphics (TVCG)*, 19(10):1646–1663, 2013.

[14] Derek Greene and Pádraig Cunningham. Practical solutions to the problem of diagonal dominance in kernel document clustering. In *International Conference on Machine learning (ICML)*, pages 377–384, 2006.

[15] Diansheng Guo. Coordinating computational and visual approaches for interactive feature selection and multivariate clustering. *Information Visualization*, 2(4):232–246, 2003.

[16] Florian Heimerl, Qi Han, Steffen Koch, and Thomas Ertl. CiteRivers: visual analytics of citation patterns. *IEEE Transactions on Visualization and Computer Graphics (TVCG)*, 22(1):190–199, 2016.

[17] Diego Ingaramo, David Pinto, Paolo Rosso, and Marcelo Errecalde. *Evaluation of Internal Validity Measures in Short-Text Corpora*, pages 555–567. Springer, 2008.

[18] Anil K Jain. Data clustering: 50 years beyond k-means. *Pattern recognition letters*, 31(8):651–666, 2010.

[19] Samuel Kaski, Janne Nikkilä, Merja Oja, Jarkko Venna, Petri Törönen, and Eero Castrén. Trustworthiness and metrics in visualizing similarity of gene expression. *BMC Bioinformatics*, 4(1):1–13, 2003.

[20] Robert Kosara, Fabian Bendix, and Helwig Hauser. Parallel sets: Interactive exploration and visual analysis of categorical data. *IEEE Transactions on Visualization and Computer Graphics (TVCG)*, 12(4):558–568, 2006.

[21] J. Krause, A. Perer, and E. Bertini. Infuse: Interactive feature selection for predictive modeling of high dimensional

data. *IEEE Transactions on Visualization and Computer Graphics (TVCG)*, 20(12):1614–1623, 2014.

[22] Hanseung Lee, Jaeyeon Kihm, Jaegul Choo, John Stasko, and Haesun Park. iVisClustering : An interactive visual document clustering via topic modeling. *Computer Graphics Forum*, 31(3):1155–1164, 2012.

[23] Frank Lin and William W Cohen. Power iteration clustering. In *International Conference on Machine Learning (ICML)*, pages 655–662, 2010.

[24] Bing Liu. *Web data mining: exploring hyperlinks, contents, and usage data*. Springer, 2007.

[25] Tao Liu, Shengping Liu, Zheng Chen, and Wei-Ying Ma. An evaluation on feature selection for text clustering. In *International Conference on Machine Learning (ICML)*, volume 3, pages 488–495, 2003.

[26] Sehi L'Yi, Bongkyung Ko, DongHwa Shin, Young-Joon Cho, Jaeyong Lee, Bohyoung Kim, and Jinwook Seo. XCluSim: a visual analytics tool for interactively comparing multiple clustering results of bioinformatics data. *BMC bioinformatics*, 16(11):1, 2015.

[27] Alan M. MacEachren, Anthony Robinson, Steven Gardner, Robert Murray, Mark Gahegan, and Elisabeth Hetzler. Visualizing geospatial information uncertainty: What we know and what we need to know. cartography and geographic. *Cartography and Geographic Information Science*, 32(3):139–160, 2005.

[28] T. May, A. Bannach, J. Davey, T. Ruppert, and J. Kohlhammer. Guiding feature subset selection with an interactive visualization. In *IEEE Conference on Visual Analytics Science and Technology (VAST)*, pages 111–120. IEEE Computer Society, 2011.

[29] D. Oelke, H. Strobelt, C. Rohrdantz, I. Gurevych, and O. Deussen. Comparative exploration of document collections: A visual analytics approach. *Computer Graphics Forum*, 33(3):201–210, 2014.

[30] Josh Oosterman and Andy Cockburn. An empirical comparison of tag clouds and tables. In *Conference of the Computer-Human Interaction Special Interest Group of Australia on Computer-Human Interaction (OZCHI)*, pages 288–295. ACM, 2010.

[31] Alex T. Pang, Craig M. Wittenbrink, and Suresh K. Lodha. Approaches to uncertainty visualization. *The Visual Computer*, 13(8):370–390, 1997.

[32] Eric D Ragan, Alex Endert, Jibonananda Sanyal, and Jian Chen. Characterizing provenance in visualization and data analysis: an organizational framework of provenance types and purposes. *IEEE Transactions on Visualization and Computer Graphics (TVCG)*, 22(1):31–40, 2016.

[33] Tobias Ruppert, Andreas Bannach, Jürgen Bernard, Hendrik Lücke-Tieke, Andreas Ulmer, and Jörn Kohlhammer. Supporting collaborative political decision making - an interactive policy process visualization system. In *International Symposium on Visual Information Communication and Interaction (VINCI)*. ACM, 2016.

[34] Dominik Sacha, Hansi Senaratne, Bum Chul Kwon, Geoffrey Ellis, and Daniel A Keim. The role of uncertainty, awareness, and trust in visual analytics. *IEEE Transactions on Visualization and Computer Graphics (TVCG)*, 22(1):240–249, 2016.

[35] J. W. Sammon. A nonlinear mapping for data structure analysis. *IEEE Transactions on Computers*, 18(5):401–409, 1969.

[36] Jinwook Seo and Ben Shneiderman. A rank-by-feature framework for interactive exploration of multidimensional data. *Information Visualization*, 4(2):96–113, 2005.

[37] Michael Steinbach, George Karypis, and Vipin Kumar. A comparison of document clustering techniques. In *KDD workshop on text mining*, pages 525–526. ACM, 2000.

[38] Furu Wei, Shixia Liu, Yangqiu Song, Shimei Pan, Michelle X. Zhou, Weihong Qian, Lei Shi, Li Tan, and Qiang Zhang. TIARA: A visual exploratory text analytic system. In *SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*. ACM, 2010.

[39] Ryen W. White and Resa A. Roth. Exploratory search: Beyond the query-response paradigm. *Synthesis Lectures on Information Concepts, Retrieval, and Services*, 1(1):1–98, 2009.

[40] James A Wise. The ecological approach to text visualization. *Journal of the Association for Information Science and Technology*, 50(13), 1999.