Eurographics Conference on Visualization (EuroVis) 2014
H. Carr, P. Rheingans, and H. Schumann
(Guest Editors)

*Volume 33 (2014), Number 3*

# Comparative Exploration of Document Collections: a Visual Analytics Approach

D. Oelke[1], H. Strobelt[2], C. Rohrdantz[3], I. Gurevych[1,4] and O. Deussen[5]

[1]UKP Lab, German Institute for Educational Research and Educational Information (DIPF), Frankfurt, Germany
[2]Polytechnic Institute of New York University, New York, USA
[3]Data Analysis & Visualization Group, University of Konstanz, Germany
[4]UKP Lab, Technische Universität Darmstadt, Germany
[5]Computer Graphics and Media Design Lab, University of Konstanz, Germany

**Abstract**

*We present an analysis and visualization method for computing what distinguishes a given document collection from others. We determine topics that discriminate a subset of collections from the remaining ones by applying probabilistic topic modeling and subsequently approximating the two relevant criteria distinctiveness and characteristicness algorithmically through a set of heuristics. Furthermore, we suggest a novel visualization method called DiTop-View, in which topics are represented by glyphs (topic coins) that are arranged on a 2D plane. Topic coins are designed to encode all information necessary for performing comparative analyses such as the class membership of a topic, its most probable terms and the discriminative relations. We evaluate our topic analysis using statistical measures and a small user experiment and present an expert case study with researchers from political sciences analyzing two real-world datasets.*

Categories and Subject Descriptors (according to ACM CCS): H.5.m [Information Systems]: Information Interfaces and Presentation—Miscellaneous

## 1. Introduction

In recent years, probabilistic topic modeling has become a standard analysis technique for the exploration of large document collections. In probabilistic topic modeling, topics can be automatically extracted from a document corpus and a topic is defined as a probability distribution over words, i.e., a topic consists of a set of weighted descriptive words. The descriptive words give insight into the thematic structure of a document collection and provide a semantic facet for the analysis. In previous research topic modeling is almost exclusively applied for the analysis of one single document collection. In contrast, we aim to extend the analysis to several collections or classes of documents. In this paper we will prefer the term *class*. According to our terminology a class of documents is a set of documents that can be subsumed under a common label. For example, all papers published at a certain conference (e.g., the IEEE VAST) can be considered as a class in contrast to the papers published at another conference (e.g., the IEEE InfoVis). The definition of a class is generic and thus our methodology is widely applicable. Note that the goal of our suggested approach is *not* to provide insight into one class of documents per se, but to enable a comparative analysis of different classes of documents. The main analysis tasks that our approach aims to support by automatic and visual means can be described through three key questions:

1. Which topics discriminate one class against the remaining classes, i.e., what is the content that is exclusive to one of the classes?
2. Which topics discriminate a subset of all classes against the remaining classes, i.e., what is the content that several classes share and that is not contained in the rest of the overall corpus?
3. Which topics do all classes have in common, i.e., what is the content that is strongly represented across all classes under investigation?

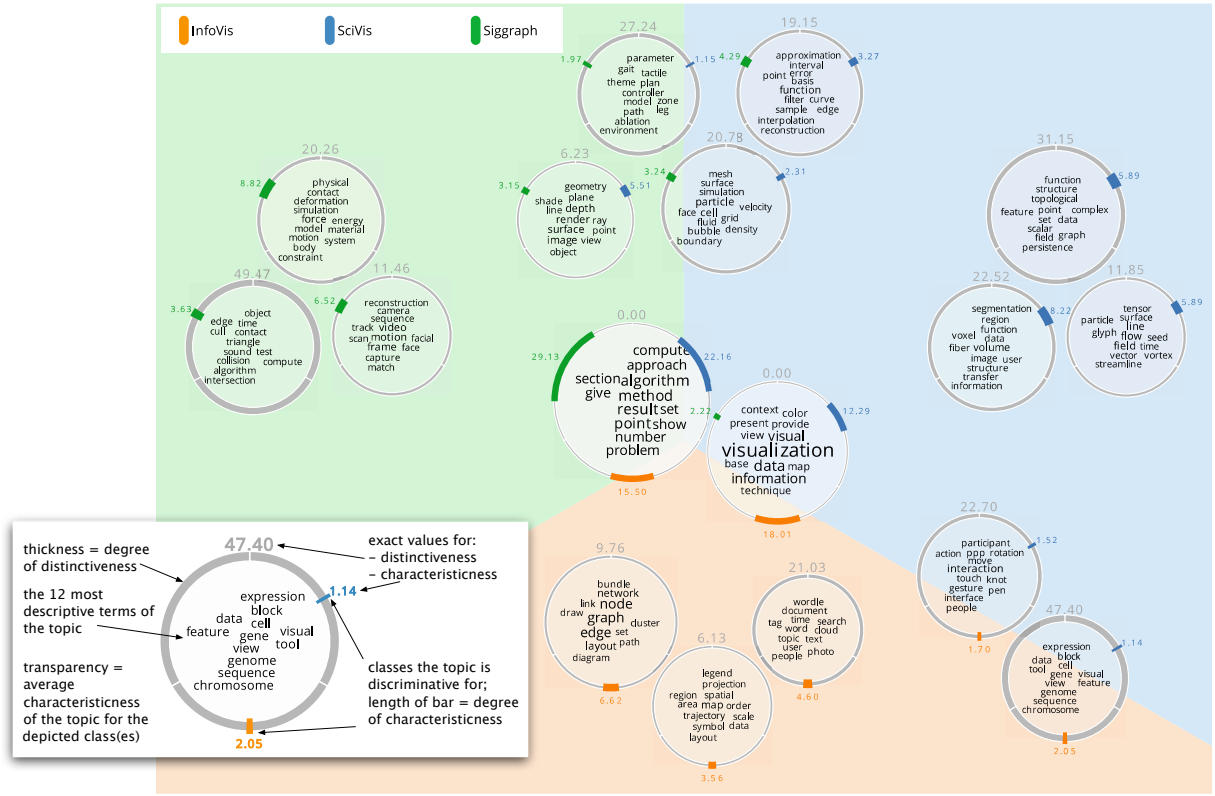


Figure 1: Comparison of 495 papers of InfoVis, SciVis, and Siggraph (discrimination threshold = 6, number of topics = 30)

Figure 1 shows the visual output when comparing proceedings of 3 visualization and computer graphics conferences. The data set comprises 495 papers, 165 of each of the three conferences (2009 - 2012 for InfoVis and SciVis, and 2011-2012 for Siggraph). The inlay of Fig. 1 illustrates how to read the glyphs called topic coins. The example coin shows a topic that is shared by SciVis and InfoVis (as can be seen by the blue and orange bar as well as its position in the diagram along the border between the blue and orange area). It discriminates the two conferences against the third one, Siggraph. The thickness of the borderline of the topic coin shows that the discriminative strength is high for this topic (metaphor of a protection wall). At the same time the topic is not a key topic of the two conferences but slightly more important for InfoVis than for SciVis (as can be seen by the rather short lengths of the colored bars that illustrate the characteristicness of the topic).

In the following we will detail our approach and our design decisions. Our contribution is twofold: First, we suggest novel automatic methods that extract discriminative and common topics for the comparative analysis of different classes of documents. Second, we suggest a visual design that enables users to explore the results in an intuitive way.

The rest of this paper is structured as follows: First, in Section 2, we describe related work. Next, in Section 3, we discuss our choice for probabilistic topic modeling and provide the definitions and formulas we use in order to automatically determine if topics are discriminative or common. We evaluate our approach both statistically and through a brief user study. Section 4 details the design of the interactive visual interface that we suggest in order to support analysts in the exploration of the automatically determined topics. The applicability and usefulness of our approach are empirically demonstrated through an expert case study in Section 5, before we conclude the paper in Section 6.

## 2. Related Work in Visual Analytics

In the following related visual analysis approaches are reviewed. Note that techniques that directly influenced our design decisions are discussed in subsequent sections.

**Exploration and Browsing of Document Collections** Many approaches exist whose goal is to support making sense of a document collection. IN-SPIRE$^{TM}$ [Ins], the topology-based approach of Oesterling et al. [OST*10], HiPP [PM08] or WebSOM [LKK04] are examples for techniques that represent document clusters by projecting them

to the two-dimensional space. Treemap-based approaches such as [New] put the focus more on the hierarchical structure of the document collection. For more details on visual text analysis in general see [AdOP12].

**Visual Analytics Approaches using Topic Modeling** Probabilistic topic modeling techniques such as Latent Dirichlet Allocation (LDA) [BNJ03] have gained in popularity in the visual analytics community in recent years.

iVisClustering [LKC*12] directly uses the clustering property of LDA to build a system that supports interactive clustering of documents. Another cluster visualization (though not related to topic modeling) is DICON [CGSQ11]. Their work is related ours because of their usage of icons to embed statistical information about the clusters. However, the suggested technique cannot deal with textual information. The Parallel Topics approach introduced in [DWCR11] employs parallel coordinates and complements this with additional visualization techniques to support users in the exploration of large text corpora. Effectively navigating a document collection is the main purpose of the system introduced in [CB12]. [DYW*13] organizes topics hierarchically. The temporal evolution of a document collection is the main perspective that some other techniques such as [LZP*12] or [CLT*11] provide.

Common to the approaches mentioned so far is that they aim at providing an overview over a collection and enabling an exploration of its content. In contrast to this, our focus is more on a comparison of the different collections. Instead of only putting the visual summaries of the different collections next to each other, we employ automatic means for pre-calculating differences and commonalities and directly present those for an interactive visual exploration.

**Word Cloud based Approaches** Besides previous work that specifically aims at improving the word cloud technique as such (cf. [KLKS10,CWL*10,PTT*12]), there are also papers that - as we do - employ word clouds as a means to summarize textual content in a visual analytics tool. Related examples include Parallel Tag Clouds [CVW09] that vertically arranges the words of a document collection with font size mapped to a significance score or the Word Storm technique [CS13] that displays one word cloud per document collection. Both approaches support a comparison between the different clouds by means like edge stubs or connecting lines that help track words (Parallel Tag Clouds) or by harmonizing the location and color of common words (Word Storm) but do not determine and display the differences directly as we do.

The TagClusters approach by Chen et al. [CSBT09] groups semantically similar tags and demarcates groups by semi-transparent background colors. Thereby, the groups may overlap each other showing a hierarchical relationship but no discriminative relations.

The WordBridge technique [KKEE11] enriches a social network that is displayed as a node-link diagram by word clouds that show documents related to specific entities and their relations. Opposed to this, we aim at showing disjunctions (distinctive topics) instead of conjunctions between sets which requires a different approach.

**Techniques for topic detection and discrimination** Different topic detection algorithms exist that perform topic identification based on a document corpus. Among the most popular ones are Latent Semantic Analysis (LSA) [DDL*90], Latent Dirichlet Analysis (LDA) [BNJ03], and simple clustering of keywords.

In recent years many variants of topic modeling algorithms have been developed. Two approaches taking class labels into account are Labeled LDA [RHNM09] and Partially Labeled Dirichlet Allocation (PLDA) [RMD11]. In contrast to Labeled LDA, PLDA can extract multiple topics per label. However, none of the approaches addresses the overlap between the classes.

Comparative Text Mining (CTM) approaches such as [GTZ*12, ZVY04] discover topics that are common across all classes and characterize for each of those topics what is unique to the different classes. In contrast to this, our approach focuses on the discriminative strength of topics extracted from the whole collection.

## 3. Automatic Detection of Discriminative and Common Topics

In this section we present and detail on our automatic processing pipeline for the extraction of *discriminative topics* from document corpora.

### 3.1. Definition of Discriminative Topics

The basic goal of this approach is to extract *topics*. A topic is defined through a list of descriptive terms. There are different topic detection algorithms that perform topic identification based on a document corpus (cf. Section 2). Our approach, however, goes beyond the state-of-the-art in that it aims at identifying special kinds of topics, namely *common* and *discriminative topics*. Discriminative topics shall support analysts in a special common analysis scenario when dealing with different classes (labeled sets) of documents. The goal is to provide answers to the analysis questions: Which common topics do all the different classes contained in a document corpus have in common? Which topics discriminate one class from the other classes? Which topics discriminate a subset of all classes from the remaining classes?

In particular, there are three main criteria we are interested in and model computationally. We look for topics which are:

1. Characteristic, i.e., describe the class(es) they are assigned to well, are important for many documents of the class, and cover the content of the documents well.

2. Distinctive, i.e., are significantly more characteristic for this class (these classes) then for the rest of the corpus, discriminate this class from the rest, and thus are its unique characteristic.
3. Interpretable, i.e., a collection of terms of which humans would say that they belong semantically together and that can readily be made sense of.

When a topic is both characteristic and distinctive for a single class or a subset of all classes, we define that this topic is considered to be *discriminative* for the given class or subset. In case that a topic is characteristic for the whole set of classes and does not discriminate subsets, we consider it to be *common*.

### 3.2. Processing pipeline

Before applying the topic modeling, we perform standard text preprocessing. Many documents, like for example scientific publications, are only available in PDF format. Consequently, we apply a PDF converter with structure recognition [SSKK10] in order to access the plain text content. Next, the text is cleaned and normalized performing stop word, noise, and number elimination and lemmatizing all words.

After preprocessing, the documents are ready for topic modeling. We apply standard latent Dirichlet allocation technology [Mal] in order to extract topics from the document corpus. Next, the discriminating and common topics are determined using the approach described in Sec. 3.3. Finally, the results are visualized and displayed to the user with the technique described in Sec. 4.

### 3.3. Determining Discriminative Topics

The heuristics we developed for determining which of the LDA topics are common or discriminative were inspired by [KOR10] that determines discriminating *terms*. The informal definitions given in Section 3.1 are implemented in the following measures:

**Distinctive Topics** For each topic $t_j$ the *average probability* per class $c_i$ is calculated by

$$\bar{p}(c_i,t_j) = \frac{1}{|\{d : d \in c_i\}|} \sum_{d \in c_i} lda\_prob(t_j|d), \qquad (1)$$

with $|\{d : d \in c_i\}|$ the number of documents that $c_i$ contains and $lda\_prob(t_j|d)$ the probability that document $d$ belongs to topic $t_j$.

A topic $t_j$ is considered as distinctive for a class $c_i$ against the remaining classes $\{c_1, c_2, \cdots, c_n\} \setminus c_i$ if and only if its average probability $\bar{p}(c_i,t_j)$ is at least $x$ times higher than the highest average probability of one of the remaining classes. We name *x discrimination threshold*.

In analogy to that a topic is defined to be distinctive for a *subset* of all classes if the lowest average probability of the classes in the subset is at least $x$ times higher than the highest average probability of the remaining classes. Thus, a topic $t_j$ is considered as distinctive for a set of classes $\{c_i, \cdots, c_k\}$ if and only if the following condition holds:

$$min(\{\bar{p}(c_i,t_j), \ldots, \bar{p}(c_k,t_j)\}) \geq x \cdot max(\{\bar{p}(c_1,t_j), \cdots, \bar{p}(c_n,t_j)\} \setminus \{\bar{p}(c_i,t_j), \cdots, \bar{p}(c_k,t_j)\}). \qquad (2)$$

It is possible that one topic fulfills the property of distinctiveness for several subsets of different sizes at the same time. In this case the one with the highest distinctiveness is chosen, i.e., the subset that would be distinctive for the highest initialization of the discrimination threshold $x$. We name this highest possible discrimination threshold the *discrimination factor* of a topic. If the discrimination factor is the same for several different subsets of all classes, which is rarely the case, we stick with the smallest subset.

**Characteristic Topics** In some cases a certain topic $t_j$ may be considered to be distinctive for a class (as part of a set of classes), for which it is rather unimportant. In order to prevent such cases we require discriminative topics not only to be distinctive, but also to be characteristic (for the classes they discriminate). A topic $t_j$ is considered as characteristic for a class $c_i$, if its average probability per document for this topic ($\bar{p}(c_i,t_j)$) is at least as high or above its total average probability over the set of all documents $\{d\}$:

$$\bar{p}(c_i,t_j) \geq \bar{p}(t_j) \qquad with \qquad \bar{p}(t_j) = \frac{1}{|\{d\}|} \sum_d lda\_prob(t_j|d).$$

It is important to consider that there are different kinds of topics generated by LDA and that not all topics are equally strongly represented in a document corpus. Some topics are important for only few documents and some are important for almost all documents of the corpus.

### 3.4. Determining Common Topics

In order to determine whether a certain topic $t_j$ is descriptive for the whole document corpus, we make use of a measure from information theory, namely the *entropy*. The entropy is high for topics that have similar occurrence probabilities across all of the documents. In particular, we use the *normed entropy* which is calculated by dividing the entropy of a topic through the maximal possible entropy, which in turn depends on the number of documents $|\{d\}|$ contained in the corpus. Equation 3 details how the normed entropy can be calculated based on the LDA output probabilities ($lda\_prob$).

$$Normed\_Entropy(t) = -\sum_{i=1}^{|\{d\}|} p(d_i|t) \cdot log_{|\{d\}|} p(d_i|t) \qquad (3)$$

with

$$p(d_i|t) = \frac{lda\_prob(t|d_i)}{\sum_{j=1}^{|\{d\}|} lda\_prob(t|d_j)}$$

The normed entropy produces values in the interval ]0,1[. The closer the normed entropy value of a topic approaches 1, the more equally distributed is this topic across all documents. As common topics shall be characterized by such an equal distribution, we empirically determined a threshold of 0.9. That is, topics with a normed entropy above 0.9 are considered to be common topics in our application.

### 3.5. Evaluation

Evaluating a method that is designed to extract semantic aspects from natural language texts is a quite difficult task. There is no standard evaluation methodology that would yield hard unquestionable facts and the results of any applied evaluation will always be arguable at least to a certain extent. In order to address this issue, we combine two complementary evaluation strategies, namely a quantitative statistical evaluation and a user study.

#### 3.5.1. Statistical analysis

If the technique works well, then we can expect that no discriminative topics are found if there is nothing to discriminate in the document corpus. To evaluate how our algorithm can deal with such a situation, we partitioned the papers of two conferences (IEEE InfoVis and ACM Siggraph) into four classes (two random classes per conference with 82 papers each). Next, we conducted a leave-one-out test by building all four possible triples of classes, each time leaving a different class out. We then analyzed the distribution of the discrimination factors and the amount of topics assigned to each class or pair of classes.

Table 1 shows the results. Each column in the table represents one class or combination of classes. Thereby, A1 and A2 refer to the classes that contain papers from the same conference and B to the one with papers from another conference. Each row represents one of the four trials (each time leaving a different class out). The depicted values show the average discriminative factor of the assigned topics and (after the slash) the number of topics assigned to the class(es).

If the method performs well, the first two as well as the fourth and fifth column should contain rather low values (i.e., few and weakly discriminative topics) and the third and sixth column large ones (i.e., many strongly discriminating topics), which is indeed the case.

The fact that there are even topics that discriminate A1 from A2 (though with low discrimination factors) reflects topic biases within the classes of a single conference. To further investigate on this effect, we conducted a second test in which we only discriminated the InfoVis subset A1 with the subset A2. Because the documents of the two classes report on the same topic only subtle differences can be observed and consequently, the discrimination factors are low, ranging from 1.0 to 3.5 with an average of 1.78. When we experimentally replaced 4 papers of one of the classes (about 5%) with Siggraph papers on light ray techniques, the overall distribution of discrimination factors remained similar but the infiltration was reflected by one strongly discriminative topic (77.0) covering the light ray vocabulary.

| A1 | A2 | B | A1,B | A2,B | A1,A2 |
|---|---|---|---|---|---|
| 3.5/1 | 2.0/2 | 23.1/13 | -/0 | -/0 | 16.1/14 |
| 1.9/2 | 2.2/2 | 21.2/13 | -/0 | -/0 | 33.1/13 |
| 1.9/1 | 2.2/2 | 38.3/10 | -/0 | -/0 | 25.6/17 |
| -/0 | 1.5/2 | 8.0/11 | -/0 | -/0 | 32.2/17 |

Table 1: Leave-one-out Test. Numbers: Discrimination factor / Number of topics.

#### 3.5.2. User Experiment

Unfortunately, no ground-truth exists saying how much a topic discriminates one class from another. Therefore, we conducted a small user experiment to understand how well our algorithmic measure reflects the users' notion of characteristicness and distinctivity. In the study the participants were shown a visual representation of the results of the algorithm and were asked to rate each topic with respect to how characteristic / distinctive it is for the specific class(es) [scales in the questionnaire ranged from "very characteristic / distinctive" to "not at all characteristic / distinctive" with two intermediate stages + an "I don't know" choice].

For preparation, participants were handed out an information sheet with informal definitions of the terms "characteristic", "distinctive" and "discriminative" as we use them in our project (see Sec. 3.1). Also, they were given an example for a topic that discriminates two classes from the third one and one that discriminates one class from the remaining ones as an explanation of how to read the visualization. No further training was provided. The 10 participants were all PhD students and PostDocs of different computer science labs (either specialized in visualization or computational linguistics). To ensure that they all are experts for the data and can indeed assess it, two different data sets were generated (one with the 90 most recent papers of three Computational Linguistics researchers working in related areas and another one with 495 papers of three visualization conferences). The participants were also asked to self-assess their background knowledge in the respective domain in the questionnaire.

In each dataset three randomly chosen topics of the corpus (two discriminating one class from the rest and one discriminating two classes together from the third one) were exchanged with each other. This allows us to compare the results for the randomly assigned topics with the ones that the

| | algorithmic assignm. | random assignm. |
|---|---|---|
| characteristic | 0.82 (1.24) | -0.57 (1.33) |
| distinctive | 0.3 (1.54) | -1.24 (1.28) |

Table 2: Result of user study (automatic analysis), standard deviation in brackets

algorithm classified as discriminative and at the same time ensures that good results are not merely due to a suggestive power of the visualization.

Evaluation: Our hypothesis was that the randomly assigned topics get significantly lower scores than the ones that our algorithm assigned. We therefore calculated the average weighted score of characteristicness / distinctivity for both the randomly assigned topics and the ones of the algorithm. "Weighted" in this case means that a classification as "very characteristic/distinctive" was counted twice and "not at all characteristic / distinctive" was recorded with -2 (intermediate steps were weighted with 1 for rating "somehow..." and -1 for "rather not...", there was no 0 option).

Table 2 shows the result of the evaluation. In both cases the average scores for the topics assigned by the algorithm are positive, whereas the scores for the randomly assigned topics are negative. This indicates that the human notion of discrimination is approximated by our algorithm.

## 4. Visualization Method

The set of discriminative and common topics are input to DiTop-View, the visualization technique described in this section. To decide on an effective visual mapping and to motivate our design rationales, we first list tasks that are important when comparing document collections with respect to the topics they address:

- **T1:** Understanding the concept a topic represents.
- **T2:** Identifying topics that discriminate a class (or combinations of classes) against the remaining classes.
- **T3:** Identifying single classes or combinations of classes that have no topic that discriminates them against the remaining classes.
- **T4:** Determining discrimination properties of a topic such as its degree of distinctiveness and characteristicness.
- **T5:** Identifying outliers such as topics which are significantly more distinctive or characteristic for a class than the remaining discriminative topics of this class.
- **T6:** Reasoning about the data in terms of discriminative topics which requires comparing different classes (or class overlaps) and their topics and setting them in relation to each other (getting the big picture).

### 4.1. Visual Design

**Showing the affiliation of terms to topics**

The automatic analysis results in sets of topics that are either discriminative or common for the collections. Thereby, topics are defined as distributions over words which have different probabilities to occur within the specific topic. Task T1 requires to give sufficient details about a topic to allow the user to derive the underlying concept. Therefore, we summarize a topic by displaying its most probable terms in a word cloud which is a common representation to display term clusters. We map the occurrence probability of each term within the given topic to its font size which is globally normalized to permit a comparison between the clouds. When placed on the canvas with sufficient distance between each other, the visual encoding of term-to-topic affiliation as clouds underlies the Gestalt law of proximity.

**Displaying the affiliations of topics and their discriminative relations**

Next, we have to show which topic(s) discriminate which class(es) from which other class(es). This requires to encode (a) *the affiliation of topics to classes or sets of classes* and (b) their *discriminative relations* (i.e. which class(es) a topic discriminates from the rest). The first requirement (a) relates to finding a sufficient representation of set relations, which is a challenging task for which multiple solutions have been proposed in the past. What makes our situation special is that we have to deal with complex objects (the word clouds) and their spatial extension. In the following we will briefly review the different solutions for displaying overlapping sets and assess them in terms of their suitability for our scenario. Finding an effective mapping of the affiliations and discriminative relations is especially important for tasks T2 and T3 as well as the overview required by task T6.

Euler diagrams are a common technique to represent sets and their relationships by depicting each set as a closed region in the plane, which is similar to what we want to convey. An intersection between sets is displayed as an overlap between their respective regions. Thanks to their long tradition and intuitiveness, Euler diagrams are wide spread and well known. However, with larger numbers of sets the representation soon gets very complex (cf. [RD10]) which is why it is mostly used with few sets only (typically 3). Furthermore, in our scenario the number of elements in each set can differ significantly which results in a loss of the regular structure that would be an ideal aid for understanding the discriminative relations.

Alsallakh et al. [AAMH13] suggest Radial Sets, a system that combines multiple approaches for displaying overlapping sets. The main visualization is the Radial Sets view in which the different classes are each assigned a position on a ring similar to [Mis06]. Overlaps of degree = 2 are visualized as arcs between two classes. Overlaps of degree $\geq 3$ are represented with a circle of proportional size that is linked to the different classes. Adapting the technique for our scenario would mean to exchange the circles inside the ring with word clouds. This would further aggravate the problem of clutter caused by overlapping lines and objects in this technique.

Displaying set memberships only on demand is not an option if we want to support tasks T3 and T6.

Other approaches assume that there are predefined positions for the elements (such as the location on a map, a graph structure or a tabular representation) and that the set relations have to be overlaid. [CPC09] does so with enclosing contours (bubbles) whereas [ARRC11] uses lines to connect the elements. However, both approaches do no show discriminative relations and are therefore not able to fulfill the requirements of T2, T3 and T6 well. An even more specialized technique is presented in [XDC*13] that deals with encoding set relations in a graph.

The DiTop-View suggested in this paper follows the idea of the Euler diagram in the sense that we assign each class (or set) a closed region in the plane. However, in our case those regions do not overlap but have common borders. Topics that are shared by several classes are overlaid at the border of the two sets (see Fig. 1). We tested design alternatives for emphasizing this containment in two sets. Enclosing all shared topics of one group with an outline or coloring their background with a fourth color broke the visual clarity and over-emphasized these groups. We also thought about rendering these topic coins differently by using blended colors or a zebra pattern with the colors of the involved classes. However, human perception of blended colors is very limited and zebra patterns seemed too distracting. We finally decided to use position (and whitespace) homogeneously for encoding all group containment. Our design ensures a coherent structure, i.e., the class(es) a topic is discriminative for are always opposite of the class(es) that they are discriminated from. A limitation of our design is that at most three classes can be compared at a time.

Additionally, the class membership is encoded in the *topic coin* - a glyph that is used to show the discrimination properties (details below). This allows us to go also beyond designs that use the position in the 2D plane to encode set membership. In an alternative view we arrange the topic coins line by line (see Fig. 2). This representation comes with the advantage that it is more space-efficient and thus more scalable in terms of the number of topics that can be displayed. It is especially beneficial when the focus of the analysis is on the discrimination properties which can be used as sorting criteria in this view. However, it is more tedious in this view to see the discriminative relations and to compare different classes. Other arrangements, e.g., using projection techniques like MDS to position the coins could also be thought of as they would permit to take the topic similarities into account. Note that in the free arrangement or in an MDS projections, the visualization can easily be extended to work with more than three classes (as long as the number of classes still allows to assign each class enough space on the circle to distinguish the lengths of different bars).

**Visualizing discrimination properties**

We use a special glyph representation that we call *topic coin*

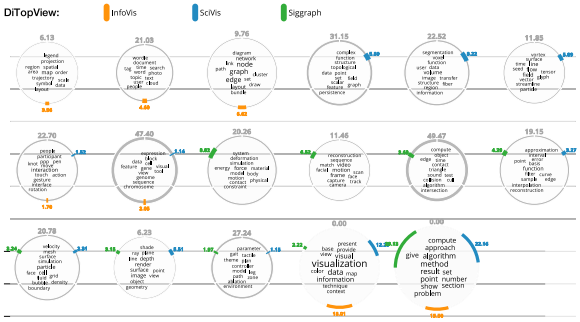


Figure 2: Topic coins arranged line-by-line and sorted according to their class membership.

that encodes the class membership but also the discrimination strength and the degree of characteristicness of a topic as required by tasks T4/T5 (see inlay of Fig. 1). Each class is assigned a section of the circular borderline of the topic coin and a color. If a topic is assigned to a class, a colored bar is shown in this area. The longer this bar is the more characteristic is the topic for the specific class. Additionally, the transparency shows the average characteristicness of the topic for the classes it was assigned to. The degree of distinctiveness of a topic is mapped to the thickness of the borderline of the coin (metaphor of "protecting walls"). On demand the encoded values can additionally be displayed as text. In the middle of each topic coin a word cloud with the most descriptive terms of the topic is shown as detailed above.

### 4.2. Technical implementation

The algorithm performs the following steps to achieve the layout of our visualization:

1. **Create topic word clouds** – We use the RWordle [SSS*12] algorithm to generate compact collections of terms. RWordle representations come with the advantage that they are space efficient and form outer shapes of aspect ratios close to one. The occurrence probability of each term is mapped to font size.
2. **Place topic word clouds at template positions** – Since we have a fixed number of possible positions for the classes, we use a template for their initial positions. The template contains seven positions namely three for the sets themselves and the four intersection areas. The positions are scaled from the centre with distance $s = \sqrt{n} \cdot k + l$ with $n$ being number of topics and $k, l$ being two scaling constants.
3. **Remove overlap and recenter** – After positioning, topics might overlap. This overlap is removed by applying RWordle on the convex hull of topic word clouds. The resulting groups of clouds are then re-centered around the template positions. A possible extension of this step is to additionally let the characteristicness of topic coins influence their position within the group of shared topics.

The layout is computed server-side in a Java Servlet and is sent to the requesting web client. In the web browser a D3.js javascript is decorating each topic with its respective Topic Coin (as shown in Figure 1). The operations of sorting and further interaction are handled by the javascript as well. If the layout is not sufficient, the user can re-arrange (drag) the coins manually.

## 5. Expert Case Study

We evaluated the system with two political science researchers working on text-based analysis of political discourse in their research. For the study a pre-version of the final system was used that is however in the most important aspects consistent with the current version (differences: arbitrarily shaped boundaries instead of circles, only avg. characteristicness shown, i.e. no colored bars indicating the characteristicness for each class and no redundant encoding with numbers). The data, proceeding, and lessons-learned will be described in the following:

**Data** The domain experts were given two datasets relating to political negotiations or mediations for exploration. The first dataset consists of the US Presidential and Vice Presidential TV Debates from 2012. Each turn within one of the debates is considered as a separate document. The turns of Obama and Biden form one class, the turns of Romney and Ryan another class, and the turns of the moderators a third class (see Fig. 3). The second dataset was from a controversial public mediation process in Germany, the Stuttgart 21 mediation (Stgt21), which was about the construction of a new underground station in Stuttgart and was broadcasted in TV over several days. Here, the turns of all project supporters are one class, the turns of all opponents another class, and the turns of the mediator and neutral experts a third class.

**Proceeding** The domain experts were invited separately and were first carefully explained the system, the rationale behind it, the design decisions, and the visual mappings. They were then asked, whether they thought they had understood the explanation and whether they found the system components intuitive or not. Next they were provided with the visualization of the presidential debates dataset (afterwards the Stgt21 dataset) and could interact with four versions (two versions for Stgt21) with different parameterizations. They were asked to investigate the data and formulate their thoughts and findings, while being observed by a visual analytics researcher who also wrote down a think-aloud protocol and interviewed them afterwards. For both experts the whole procedure took almost one hour. DiTop-Views for all datasets and parameterizations used in the expert study, additional application examples and a video showing the system in use can be found in the supplementary material.

**Findings & Lessons learned** The main outcomes of the expert study, i.e. the observations, the think-aloud protocol, and the structured interview, are summarized according to different criteria.

*Understandability & Intuitiveness*: Whereas one of the experts was familiar with topic modeling and felt he was clear about the system, the other one had doubts about the functionality of the topic modeling. He mentioned that he would have to gain a better understanding about what topic modeling algorithms exactly do, before productively working with such a system in his research. In principle, both experts found the tool intuitive. Yet, one had to ask again what the thickness of the borders around topics meant. The other one was not sure whether he had missed in the explanation what the semantics of the exact location of a topic within the given area was. He also was not sure whether the distances between different topics would carry any meaning.

*Usefulness & Scope*: Both experts expected that such a tool would be useful within their domain. One expert mentioned that he liked the quick overview on the topics and believed it to be a good starting point for qualitative researchers. In any case, analysts should be enabled to drill down to the underlying text sources in order to gain a better understanding why certain topics pop up.

*Hypotheses & Interpretations*: When analyzing the given datasets both experts found that the topics of the moderator and the topics shared among all classes actually did not really carry content, but rather joined words used to structure the discourse and relate to other stakeholders. For the overlaps between the moderators with one of the parties, both experts supposed that those were topics where the moderators explicitly asked only one of the parties. In the case of the presidential debates the experts were able to identify several topics that would discriminate one party from the other reflecting their different ideological viewpoints. Both experts found also those topics interesting that were quite similar in content, but assigned to the different parties. They assume that the word usage again reflects differences in ideology, but the basic topic is discussed by both parties. In the case of the Stuttgart 21 mediations both experts independently mentioned the same result as strikingly interesting to them. All topics assigned to the opponents of the project referred to issues of the financing and costs of the project. Yet, some of the well-known topics of the opponents as for example issues about the groundwater and a beetle species under protection, appeared on the side of the supporters of the project. The experts hypothesized that this reflects the different negotiation strategies of the counterparts. The supporters tried to weaken the known arguments of the opponents, such as environmental issues. On the other hand, the opponents focused on the main public criticism to the project, the unclear costs and doubtful profitability.
One of the experts specifically noted that he found the thickness of the topic borders a good and intuitive visual sign for the discriminativity. He reported that many of the topics most interesting to him had thick borders.

*Parameterization*: Both experts preferred configurations that resulted in a moderate number of topics. Yet, it was in-

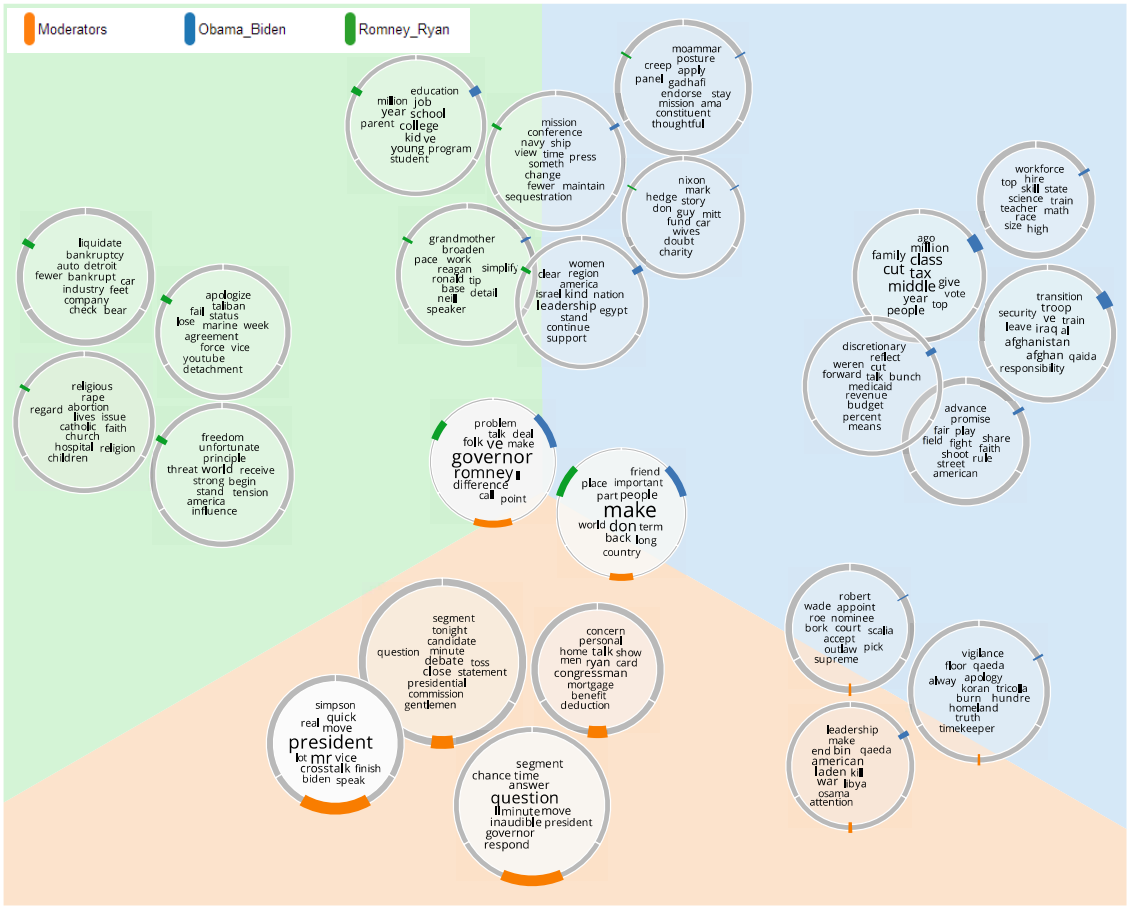


Figure 3: Presidential Debates of 2012 (discrimination threshold = 2, number of topics = 80)

teresting to observe that both experts pursued the same strategy for the Stgt21 dataset in that they used the version with more topics to investigate in more detail on the hypotheses formed with the version containing fewer topics.

## 6. Conclusions

In this paper we present a visual analytics approach that helps to detect and explore discriminative and common topics when comparing several classes of documents. We suggest an automatic method for extracting discriminative and common topics as well as a visual representation called DiTop-View that enables analysts to explore the results in an intuitive way. Our approach complements the previous line of research that aims to provide insight into single document collections. In contrast, we focus on a clear-cut task when dealing with different classes of documents and aiming at a comparison of differences and commonalities in content. The presented technique is widely applicable and can be used in scenarios like comparing publications of different conferences, books/papers written by one author, speeches held by politicians, open access course material, etc.

## Acknowledgment

This work was partially funded by the German Federal Ministry of Education and Research (BMBF) under grant 01461246 “VisArgue” and by the Volkswagen Foundation as part of the Lichtenberg-Professorship Program under grant No. I/82806.

## References

[AAMH13] ALSALLAKH B., AIGNER W., MIKSCH S., HAUSER H.: Radial Sets: Interactive Visual Analysis of Large Overlapping Sets. *IEEE Trans. on Visualization and Computer Graphics 19* (Dec. 2013), 2496–2505. 6

[AdOP12] ALENCAR A. B., DE OLIVEIRA M. C. F., PAULOVICH F. V.: Seeing beyond reading: a survey on visual text analytics. *Wiley Int. Rev. Data Min. and Knowl. Disc. 2*, 6 (Nov. 2012), 476–492. 3

[ARRC11] ALPER B., RICHE N., RAMOS G., CZERWINSKI M.:

Design Study of LineSets, a Novel Set Visualization Technique. *IEEE Trans. on Visualization and Computer Graphics 17*, 12 (2011), 2259–2267. 7

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

[CB12] CHANEY A. J.-B., BLEI D. M.: Visualizing Topic Models. In *Proc. of the 6th Intern. Conf. on Weblogs and Social Media* (2012). 3

[CGSQ11] CAO N., GOTZ D., SUN J., QU H.: DICON: Interactive Visual Analysis of Multidimensional Clusters. *IEEE Trans. on Visualization and Computer Graphics 17*, 12 (Dec 2011), 2581–2590. 3

[CLT*11] CUI W., LIU S., TAN L., SHI C., SONG Y., GAO Z. J., TONG X., QU H.: TextFlow: towards better understanding of evolving topics in text. *IEEE Trans. on Visualization and Computer Graphics 17*, 12 (2011), 2412–21. 3

[CPC09] COLLINS C., PENN G., CARPENDALE S.: Bubble Sets: Revealing Set Relations with Isocontours over Existing Visualizations. *IEEE Trans. on Visualization and Computer Graphics 15*, 6 (Nov. 2009), 1009–1016. 7

[CS13] CASTELLA Q., SUTTON C. A.: Word Storms: Multiples of Word Clouds for Visual Comparison of Documents. *CoRR abs/1301.0503* (2013). 3

[CSBT09] CHEN Y.-X., SANTAMARÍA R., BUTZ A., THERÓN R.: TagClusters: Semantic Aggregation of Collaborative Tags beyond TagClouds. In *Proc. of the 10th Intern. Symp. on Smart Graphics* (2009), SG '09, Springer-Verlag, pp. 56–67. 3

[CVW09] COLLINS C., VIÉGAS F. B., WATTENBERG M.: Parallel Tag Clouds to explore and analyze faceted text corpora. In *IEEE Symp. on Visual Analytics Science and Technology* (2009), VAST, pp. 91–98. 3

[CWL*10] CUI W., WU Y., LIU S., WEI F., ZHOU M. X., QU H.: Context-Preserving, Dynamic Word Cloud Visualization. *IEEE Comput. Graph. Appl. 30*, 6 (Nov. 2010), 42–53. 3

[DDL*90] DEERWESTER S. C., DUMAIS S. T., LANDAUER T. K., FURNAS G. W., HARSHMAN R. A.: Indexing by Latent Semantic Analysis. *JASIS 41*, 6 (1990), 391–407. 3

[DWCR11] DOU W., WANG X., CHANG R., RIBARSKY W.: ParallelTopics: A probabilistic approach to exploring document collections. In *IEEE Conf. on Visual Analytics Science and Technology* (2011), VAST, pp. 231–240. 3

[DYW*13] DOU W., YU L., WANG X., MA Z., RIBARSKY W.: HierarchicalTopics: Visually Exploring Large Text Collections Using Topic Hierarchies. *IEEE Trans. on Visualization and Computer Graphics 19*, 12 (2013), 2002–2011. 3

[GTZ*12] GAO H., TANG S., ZHANG Y., JIANG D., WU F., ZHUANG Y.: Supervised Cross-collection Topic Modeling. In *Proc. of the 20th ACM Intern. Conf. on Multimedia* (2012), MM '12, ACM, pp. 957–960. 3

[Ins] Pacific northwest national laboratory, http://in-spire.pnnl.gov/. 2

[KKEE11] KIM K., KO S., ELMQVIST N., EBERT D. S.: WordBridge: Using Composite Tag Clouds in Node-Link Diagrams for Visualizing Content and Relations in Text Corpora. In *Proc. of the 44th Hawaii Intern. Conf. on System Sciences* (2011), HICSS '11, pp. 1–8. 3

[KLKS10] KOH K., LEE B., KIM B., SEO J.: ManiWordle: Providing Flexible Control over Wordle. *IEEE Trans. on Visualization and Computer Graphics 16*, 6 (Nov. 2010), 1190–1197. 3

[KOR10] KEIM D. A., OELKE D., ROHRDANTZ C.: Analyzing Document Collections via Context-Aware Term Extraction. In *Proc. of Natural Language Processing and Information Systems*, vol. 5723 of *Lecture Notes in Computer Science*. Springer Berlin / Heidelberg, 2010, pp. 154–168. 4

[LKC*12] LEE H., KIHM J., CHOO J., STASKO J., PARK H.: iVisClustering: An Interactive Visual Document Clustering via Topic Modeling. *Comp. Graph. Forum 31*, 3 (2012), 1155–1164. 3

[LKK04] LAGUS K., KASKI S., KOHONEN T.: Mining massive document collections by the WEBSOM method. *Inf. Sci. 163*, 1-3 (June 2004), 135–156. 2

[LZP*12] LIU S., ZHOU M. X., PAN S., SONG Y., QIAN W., CAI W., LIAN X.: TIARA: Interactive, Topic-Based Visual Text Summarization and Analysis. *ACM Trans. Intell. Syst. Technol. 3*, 2 (2012), 25:1–25:28. 3

[Mal] Mallet. http://mallet.cs.umass.edu/topics.php. 4

[Mis06] MISUE K.: Drawing bipartite graphs as anchored maps. In *Proc. of Asia-Pacific Symp. on Information Visualisation* (2006), pp. 169–177. 6

[New] Newsmap, http://newsmap.jp. 3

[OST*10] OESTERLING P., SCHEUERMANN G., TERESNIAK S., HEYER G., KOCH S., ERTL T., WEBER G. H.: Two-stage framework for a topology-based projection and visualization of classified document collections. In *IEEE Symp. on Visual Analytics Science and Technology* (2010), pp. 91–98. 2

[PM08] PAULOVICH F. V., MINGHIM R.: HiPP: A Novel Hierarchical Point Placement Strategy and Its Application to the Exploration of Document Collections. *IEEE Trans. on Visualization and Computer Graphics 14*, 6 (Nov. 2008), 1229–1236. 2

[PTT*12] PAULOVICH F. V., TOLEDO F. M. B., TELLES G. P., MINGHIM R., NONATO L. G.: Semantic Wordification of Document Collections. *Comp. Graph. Forum 31*, 3 (2012), 1145–1153. 3

[RD10] RICHE N. H., DWYER T.: Untangling Euler Diagrams. *IEEE Trans. on Visualization and Computer Graphics 16*, 6 (2010), 1090–1099. 6

[RHNM09] RAMAGE D., HALL D., NALLAPATI R., MANNING C. D.: Labeled LDA: a supervised topic model for credit attribution in multi-labeled corpora. In *Proc. of the Conf. on Empirical Methods in Natural Language Processing: Volume 1* (2009), EMNLP'09, pp. 248–256. 3

[RMD11] RAMAGE D., MANNING C. D., DUMAIS S.: Partially labeled topic models for interpretable text mining. In *Proc. of the 17th ACM SIGKDD intern. conf. on Knowledge discovery and data mining* (2011), KDD'11, pp. 457–465. 3

[SSKK10] STOFFEL A., SPRETKE D., KINNEMANN H., KEIM D. A.: Enhancing Document Structure Analysis using Visual Analytics. In *Proc. of the ACM Symp. on Applied Computing* (2010), SAC, ACM, pp. 8–12. 4

[SSS*12] STROBELT H., SPICKER M., STOFFEL A., KEIM D., DEUSSEN O.: Rolled-out Wordles: A Heuristic Method for Overlap Removal of 2D Data Representatives. *Comp. Graph. Forum 31*, 3 (2012), 1135–1144. 7

[XDC*13] XU P., DU F., CAO N., SHI C., ZHOU H., QU H.: Visual Analysis of Set Relations in a Graph. *Comp. Graph. Forum 32*, 3pt1 (2013), 61–70. 7

[ZVY04] ZHAI C., VELIVELLI A., YU B.: A Cross-collection Mixture Model for Comparative Text Mining. In *Proc. of the 10th ACM SIGKDD Intern. Conf. on Knowledge Discovery and Data Mining* (2004), KDD '04, ACM, pp. 743–748. 3