An Interactive System for Visual Analytics of Dynamic Topic Models

Nikou Günemann · Michael Derntl · Ralf Klamma · Matthias Jarke

Received: xxx / Accepted: xxx

Abstract The vast amount and rapid growth of data on the web and in document repositories make knowledge extraction and trend analysis a challenging task. A well-proven approach for the unsupervised analysis of large text corpora is dynamic topic modeling. While there is a solid body of research on fundamentals and applications of this technique, visual-interactive analysis systems for allowing end-users to perform analysis tasks using topic models are still rare. In this paper we present D-VITA, an interactive text analysis system that exploits dynamic topic modeling to detect the latent topic structure and dynamics in a collection of documents. D-VITA supports end-users in understanding and exploiting the topic modeling results by providing interactive visualizations of the topic evolution in document collections and by browsing documents based on keyword search and similarity of their topic distributions. The system was evaluated by a scientific community that used D-VITA for trend analysis in their data sources. The results indicate high usability of D-VITA and its usefulness for productive analysis tasks.

1 Introduction

Monitoring and analyzing the large, constantly growing volumes of data being offered through various channels nowadays is beyond human capability. Many data mining algorithms have been proposed for dealing with this problem; a recently popular approach for automatically distilling the themes in the content of unstructured document collections is topic modeling, which uses statistical methods to classify the documents on the basis of the topics exposed in the text [2]. Many document collections—for instance scientific journals or social web data—are highly dynamic, being permanently extended with new documents and content updates. These dynamically changing document collections call for dynamic topic modeling approaches, since the topics exposed in the document collections evolve over time [3].

In today’s era of fast-paced developments in industry and science, gaining an overview of large and dynamic collections of text documents is becoming increasingly important for establishing and maintaining a competitive advantage. In this context a line of research has focused on the development of algorithms for automatic categorization of documents based on their latent, inherent topics (e.g., [4,3,21]). However, the generated results—long lists of numbers and statistics—are difficult to interpret by a “normal” human end-user, who is not familiar with the details of the topic modeling technique, but who has a strong stake in making sense of the data. Therefore, finding methods to provide intuitive and interactive visualizations of the results of topic modeling methods is of fundamental interest nowadays. Such visualizations should enable users to explore the temporal evolution of topics, which plays a central role for dynamic data collections, e.g., for trend prediction or weak signal analysis [19].

Based on this motivation we present in this paper an interactive text analysis tool for dynamic document collections called D-VITA (Dynamic Visual Interactive Text Analysis). D-VITA supports the user, for example, in identifying recent trends in a documents collection and in visually interacting with the results in a web-based application. The pilot application case was
motivated by the TEL-Map support action\(^1\), which was funded by European Commission to provide impact analyses and organize roadmapping activities in the area of Technology Enhanced Learning (TEL). There are many technology providers and technology adopters in Europe that have a stake in TEL research, and these actors are interested in staying up to date with current topics in their community. The TEL-Map Mediabase [9, 8] provided the raw data sets for this pilot case. These data sets were crawled and indexed from several TEL related web sources—including R&D project descriptions, research paper bibliographies, and blogs. D-VITA was deployed for TEL-Map stakeholders and evaluated for its usability and usefulness.

The rest of this paper is structured as follows: In Section 2, we present relevant related work on topic modeling. We then introduce in Section 3 the system architecture of D-VITA and demonstrate different visualization concepts used in the system. In Section 4 we evaluate D-VITA and discuss the results. We focus on two evaluation perspectives, namely an evaluation of the runtime performance as well as end-user studies. We conclude in Section 5 with a summary and an outlook on future work.

## 2 Related Work

### 2.1 Topic Modeling

Topic modeling is the unsupervised discovery of thematic information hidden in a set of documents. Intuitively, a topic provides a compact description of the content of documents belonging to this topic. Technically, topics are often modeled as distributions of words over pre-defined number of topics [2]. The basic idea is that documents discussing the same topic also primarily use the same words. Thus, these words can be used to describe the common topic and they give the topic its meaning. To generate effective summarizations of the topic models, we have researched for different state-of-the-art summarization techniques for document databases.

The most widely used and implemented method to determine topics is Latent Dirichlet Allocation (LDA) [4]. LDA models the process of generating the words of each document. Let \( K \) be a specified number of topics and \( V \) the total number of words. For each document the categorical distribution over topics is generated by the Dirichlet distribution \( \text{Dir}_K(\alpha) \) controlled by \( \alpha \), a positive \( K \)-vector. Furthermore, LDA uses a Dirichlet distribution controlled by \( \eta \) to generate the actual topics (distributions over words). We let \( \text{Dir}_V(\eta) \) denote a \( V \)-dimensional Dirichlet with scalar parameter \( \eta \). The overall generative process of LDA is given as follows:

1. For each topic \( k \),
   (a) Draw a distribution over words \( \beta_k \sim \text{Dir}_V(\eta) \).
2. For each document \( d \),
   (a) Draw a vector of topic proportions \( \theta_d \sim \text{Dir}_K(\alpha) \).
   (b) For each word,
      (i) Draw a topic assignment \( Z_{d,n} \sim \text{Mult}(\theta_d) \), \( Z_{d,n} \in 1, \ldots, K \).
      (ii) Draw a word \( W_{d,n} \sim \text{Mult}(\beta_{Z_{d,n}}) \), \( W_{d,n} \in 1, \ldots, V \).

This process is schematically shown in Figure 1. Each circular node in the figure corresponds to a random variable and edges denote dependencies between these variables.

### 2.2 Dynamic Topic Modeling

While static topic models consider a particular snapshot of a document collection without modeling the change over time, dynamic topic models consider the temporal dynamic by using a time-stamped collection of documents. Several dynamic topic models have been proposed in the literature (e.g. [18, 3, 21, 22, 12, 16, 1, 14, 13]). We use one of the earliest approaches [3] to describe the general idea behind these models. In this model the temporal dimension is addressed by two steps: First, the documents are partitioned into so called “epochs” with each epoch containing all documents of a certain time slices, e.g., of a specific year or a specific point in time. In a second step, the dependency between epochs is modeled. The generative process of this model


\[\begin{align*}
\alpha & \quad \Theta, \\
\beta_k & \quad \text{Dir}_V(\eta) \\
\theta_d & \quad \text{Dir}_K(\alpha) \\
Z_{d,n} & \quad \text{Mult}(\theta_d) \\
W_{d,n} & \quad \text{Mult}(\beta_{Z_{d,n}})
\end{align*}\]

**Fig. 1** The graphical model of LDA
is depicted in Figure 2, which contains four instances of the static LDA illustration (cf. Figure 1), each of them representing one epoch.

The general idea of the dynamic LDA model is that the content of each topic, described by the categorical distribution $\beta_k$ over words, changes smoothly over time. That is, words important for a topic at a specific point in time are very likely important also in the next time step. In Figure 2 this dependency is shown by the horizontal arrows pointing from left to right. Formally, we introduce for each topic $k$ and epoch $t$ a distribution $\beta_k^{(t)}$ and we model the smooth evolution between $\beta_k^{(t+1)}$ and $\beta_k^{(t)}$. We refer to [3] for more technical details about the actual evolution. Similarly, we model the evolution of the topic proportions by introducing $\alpha^{(t+1)}$ and $\alpha^{(t)}$.

3 The D-VITA System

In this section we present the D-VITA system which exploits the ideas of dynamic LDA and offers a web-based user interface to visually interact with the results of dynamic LDA on arbitrary data sources. We divided the system architecture in three layers: data layer, application layer, and presentation layer. Figure 3 shows an overview of the system architecture and Figure 4 shows a screenshot of the D-VITA main window.

3.1 Data Layer

Besides different kinds of raw data (e.g., blogs, research papers, or web content) the data layer stores the results generated by the application layer. This includes the source document collection in a preprocessed format for dynamic LDA as well as data on the detected topics. All relevant information is stored in a relational database.

3.2 Application Layer

As illustrated in Figure 3, we divide the application layer into two main components for data processing: an online and an offline component. The offline component performs tasks that are only executed infrequently and potentially require lengthy processing. In the offline component we integrate the data preprocessing, the execution of topic modeling algorithms, and the computation of the documents’ similarity. Data preprocessing is required since the architecture is intended to be independent of particular data schemas in the raw source data. In the TEL-Map Mediabase, for instance, blog posts in the blog database are used as input documents for the topic modeling algorithms. The preprocessing step also extracts the set of distinct words from the raw data, it applies stop-word list, and performs stemming to reduce inflected or derived words to their stem.

In contrast to the offline component, the tasks of the online component, are executed frequently in response to user actions in the D-VITA graphical user interface and should instantly retrieve and display results in the browser. The online component contains several tools that support the user in interacting with the topic modeling results, which relate to three core concepts: documents, topics and words. The topic modeling results are visualized by the presentation layer which is described in the following section.

Technical Realization. The application layer is implemented based on Apache Tomcat\(^2\). Apache Tomcat is
an open source software implementation of the Java Servlet and JavaServer Pages technologies. It supports the implementation and execution of Java programs (servlets) in a web server environment. Each servlet provides publicly accessible methods which can be invoked by the clients. The client-server communication is automatically handled using protocols like HTTP.

3.3 Presentation Layer

The key idea in representing dynamic LDA results in D-VITA was to exploit the paradigm of Visual Analytics [15]. Visual analytics aim at integrating the human capabilities into the data analysis process by using visual representations and interaction techniques. Thus, the user is involved in the analysis process and data interpretation and reasoning are supported by visualizing the important aspects of the data. Various methods of visual analytics for the analysis of document content have been proposed in previous work, e.g., for topic-based navigation in Wikipedia [5] and in TIARA [17]. Existing systems, however, often fail to support important use cases like handling of dynamic data or detecting and highlighting emerging topics in the data. Furthermore, none of the existing tools decouples the topic modeling process from the data sources in a way that allows “plugging in” arbitrary databases, which is achieved in D-VITA by mapping the source data schema to a simple, well-defined target schema that can be used by the offline component for data preprocessing.

Visual Analysis of Dynamic Topic Models. In D-VITA, topic analysis can be performed in different ways. One possibility is to present topics in conjunction with the words assigned to the corresponding topic. Based on the LDA procedure, each topic $k \in 1, \ldots, K$ can be described by its most frequently occurring words. For topic labeling, we use the four most important words in the word distribution of a topic. This is a common approach in existing topic modeling systems. Formally, the four most important words $\text{words}_k \subseteq \{1, \ldots, V\}$ for topic $k$ can be computed based on the word distributions $\beta_k^{(t)}$:

$$\text{words}_k := 4 \arg \max_{w \in \{1, \ldots, V\}} \left\{ \sum_{t=1}^{T} \beta_k^{(t)}[w] \right\},$$

where the function $4 \arg \max$ denotes the generalization of the arg max function to return the four elements with the highest values. An example of a list of topics and their four-word labels identified in a data source is depicted in the left-hand pane of Figure 4.

While labeling topics based on frequent words serves for defining topics, it does not provide any information whether a topic is relevant for the document collection. That is, while some topics might be discussed by many documents, other topics might be less important. Since each document contributes to each topic at a different degree, the relevance of topic $k$ at time $t$ can formally be defined as

$$\text{relevance}(k, t) := \frac{\sum_{d \in DB_t} \theta_d[k]}{|DB_t|},$$

where $DB_t$ is the set of documents belonging to time $t$ and $\theta_d$ is the topic distribution for document $d$. Following this formalism a relevant topic is a topic which is important for many documents at the current time interval. That is, many documents are well described by the considered topic. Consequently, the higher the relevance, the more important is a topic. As an example, consider a time interval containing 100 documents. Assuming 25 of these documents are completely described by topic 1, the relevance of topic 1 at this time point would be 0.25. If 10 additional documents belong to this topic with 90% (i.e., a small fraction of each document refers to other topics) the relevance of the topic would be 0.1 + 0.09 = 0.39.

The above definition of relevance follows the idea presented in [5], with the difference that we normalize the relevance value using the term $1/|DB_t|$. In [5] only a static document collection is considered, i.e. only a single database $DB_1$ is given. In our model, however, we have to consider multiple databases $DB_t$ of potentially different sizes. Thus, to ensure a fair comparison of relevance values among different time intervals, we normalize by $1/|DB_t|$. This ensures for each time $t$ the property $\sum_{k \in K} \text{relevance}(k, t) = 1$, i.e. the relevance summed over all topics at a specific time is normalized to 1.

Each topic’s relevance might greatly vary over time. This kind of topic evolution is an important characteristic for the visual presentation of the topic evolution. D-VITA allows the selection of multiple topics for visualizing the topic evolution using a themeriver graph [11]. This graph is illustrated in Figure 4 in the upper part. Each current in the themeriver presents the dynamic development of a selected topic in a particular time interval. The wider the current, the higher the relevance of a topic at that time, i.e. the more documents “contribute” to the topic.

To facilitate this visual analysis of the topic evolution, D-VITA provides topic ranking functionality as a parameter to analyze the relevance of topics in comparison to each other. In the following, we briefly describe these topic ranking functions.
Fig. 4 D-VITA’s visual summary of topics. In the visualization, the themeriver [11] (top-right-pane) represents the selected topics and their evolution over different time slices. By selecting a topic in the themeriver, a list of documents exhibiting the selected topic is shown in the bottom-right pane. By selecting a document from this list, e.g., the conference paper “Unsupervised Discovery of Student Learning Tactics” as in the figure, the distribution of topics in the document is visualized as a pie chart. The document’s content can be inspected by clicking the button “content”. Additionally, a user can retrieve for each document a list of similar documents based on shared topics.

Topic Ranking. The first two rankings we propose are based on the average relevance and variance in relevance of the topics, respectively. In the case of average variance, we compute the arithmetic mean of a topic’s relevance values over all time intervals. This corresponds to

\[ \mu_k := \frac{1}{T} \sum_{t=1}^{T} \text{relevance}(k, t). \]

Given the values \( \mu_k \) for each topic, we can easily sort the topics based on this property allowing to capture the most relevant or least relevant topics in terms of average absolute relevance.

Another metric is to rank the topics based on the variance in their relevance, i.e.

\[ \text{Var}_k = \frac{1}{T} \sum_{t=1}^{T} (\text{relevance}(k, t) - \mu_k)^2. \]

The greater the variance, the higher the relative deviations in a topic’s relevance over time. This can be used to spot volatile or stable topics, respectively, depending on the sort order.

Focusing on the temporal aspect of the topics (e.g., for trend analysis), an interesting metric is to determine topics which have a high increase in relevance in their evolution. We consider this aspect by estimating the sum of the absolute differences between the topics’ relevance at several time points. Let \( \Delta_t = \text{relevance}(k, t+1) - \text{relevance}(k, t) \) we define

\[ \text{rising}_k := \sum_{t=1}^{T-1} \begin{cases} \Delta_t & \text{if } \Delta_t > 0 \\ 0 & \text{else} \end{cases} \]

The differences refer to the ascent of included documents in a topic at time \( t+1 \) compared to the included documents of same topic but at the previous time point. The greater the sum of the ascent, the more important the topic will be in this regard.

Accordingly, it might also be of interest to know which topics are of decreasing relevance in a dataset. To realize this, we reformulate the above equation to only consider time intervals with a negative ascent:

\[ \text{falling}_k := \sum_{t=1}^{T-1} \begin{cases} |\Delta_t| & \text{if } \Delta_t < 0 \\ 0 & \text{else} \end{cases} \]

It is worth mentioning that the above two equations only accumulate values \( \Delta_t \) that are either positive or negative, respectively. They do not consider the overall sum \( \sum_{t=1}^{T-1} \Delta_t \). This sum would simply resolve to \( \sum_{t=1}^{T-1} \Delta_t = \text{relevance}(T) - \text{relevance}(1) \), i.e. only the difference between the first and last time intervals is considered. The remaining time intervals would be neglected. The equation \( \sum_{t=1}^{T-1} |\Delta_t| \), in contrast, is already well captured by the definition of \( \text{Var}_k \). Considering these definitions, it also becomes apparent that
the topic with the highest value of \textit{rising}_k is not necessarily the topic with the lowest value of \textit{falling}_k. Thus ranking the topics according these measures provides two complementary views on the data.

To consider both rising and falling trends in relevance of a topic over time we use the principle of decay functions [6, 7]. The idea is that late time points (containing the most recent documents) are often more interesting than early time points (containing old documents). Thus, when considering the differences \( \Delta_t \) it is suitable to give a higher importance to the values obtained for late time points. Thus, instead of considering the unweighted and less informative sum \( \sum_{t=1}^{T-1} \Delta_t \), we use

\[
\text{rising Decay}_k := \sum_{t=1}^{T-1} e^{-\lambda(T-t)} \cdot \Delta_t.
\]

Here, the value at time \( t \) is weighted by the exponential decay function \( e^{-\lambda(T-t)} \). The smaller \( t \), the smaller the weight. In contrast to the value \textit{rising}_k that ranks the topics according to their increase in relevance independent of the times when the increase appeared, this measure prefers topics which recently showed an increase in their relevance. The value of \( \lambda \) controls how strong the preference for recent time points is. By choosing a very large \( \lambda \) only the most recent point in time is considered. For D-VITA we choose a default value of \( \lambda = 0.1 \) to ensure a reasonable focus on more recent documents without neglecting the topic evolution history in the dataset.

Overall, performing a ranking based on the above metrics allows to automatically spot the most important topics according to different characteristics.

\textbf{Presentation of Document Analysis.} In D-VITA document analysis can be used to analyze the topic-document correlations. By selecting a topic in the themeriver corresponding to a certain point in time a list of documents is displayed in the “Related Documents” pane at the bottom of the D-VITA window. These documents are the most representative documents for the selected topic at the selected point in time. Note that the representative documents for the selected topic may change when considering different points in time. Formally, given a selected topic \( k \) at time \( t \), the set of documents \( \text{Repr}_{k,t} \subseteq DB_t \) is determined based on the following equation:

\[
\text{Repr}_{k,t} := r \arg \max_{d \in DB_t} \{ \theta_d[k] \}
\]

Again, \( r \arg \max \) denotes the function returning the \( r \) elements with the highest value and \( DB_t \) is the set of all documents belonging to time \( t \). Although these documents are the most representative ones for the selected topic, the documents will typically expose several additional topics. Thus, for each document we represent the share of topics by colored slices in a pie chart as shown in Figure 4. Additionally, the original content of each document can be viewed by clicking on the respective button in the document table.

Topics distributions also allow to determine the similarity of documents. To achieve this we compare the documents’ topic distributions \( \theta_{d_1} \) and \( \theta_{d_2} \). If the distributions are very similar, both documents belong to the same topics with a similar degree. Since \( \theta_{d_1} \) and \( \theta_{d_2} \) are distributions over the same domain, it is useful to measure their similarity based on the Jensen-Shannon Divergence [10]. Thus, formally, the (dis-)similarity between two documents \( d_1 \) and \( d_2 \) is defined as

\[
\text{JSD} \prime (d_1, d_2) := \text{JSD}(\theta_{d_1}, \theta_{d_2}) = \frac{1}{2} D(\theta_{d_1} \parallel M) + \frac{1}{2} D(\theta_{d_2} \parallel M),
\]

where \( D \) is the Kullback-Leibler-Divergence and

\[
M_i := \frac{1}{2} (\theta_{d_i} + \theta_{d_i}).
\]

We are free to compare documents belonging only to the same time interval or to different intervals. Note that the smaller the Jensen-Shannon Divergence value, the higher the similarity. Based on these similarity values for pairs of documents, our system allows to determine the most similar documents w.r.t. a specific query document (cf. button for “Similar Docs” in Figure 4). More precisely, given a document \( d \) it is interesting to find the most similar documents located in a specific time interval \( t \) by computing

\[
\text{simDocs}(d, t) := r \ arg \min_{d' \in DB_t} \{ \text{JSD} \prime (d, d') \}.
\]

It is worth mentioning that the document \( d \) itself does not necessarily need to belong to time interval \( t \).

\textbf{Presentation of Word Analysis.} Finally, D-VITA allows to perform analyses of the word distributions in the dynamic topic models. The word distribution of a topic may change over time. This word evolution can be illustrated in the lower part of our system as an alternative to the related documents browser (cf. Figure 5). Based on the dynamic nature of the data, also the above mentioned keyword search is time dependent, i.e. the lists of topics and documents containing the keyword may change at different times. Additionally, the topic list can be filtered using a keyword in the search field. The filtered list will only contain topics whose word list contains the keyword. The word comparison is performed using word stems.
4 Case Study and Evaluation

4.1 Datasets

The D-VITA system was deployed using different production databases maintained in the TEL-Map project [9]. These databases index blogs, papers and research projects on Technology Enhanced Learning (TEL). The TEL blogs dataset contains 11218 blog posts from the years 2011 and 2012. The TEL papers dataset contains paper abstracts from the International Conference on Advanced Learning Technologies from 2003-2012 (ICALT; 2227 papers), from the European Conference on Technology Enhanced Learning from 2007-2012 (ECTEL; 282 papers), and from the International Conference on Web-Based-Learning 2003-2012 (ICWL; 444 papers). The TEL projects dataset contains project descriptions of EU projects funded under Technology Enhanced Learning related objectives. This dataset covers the description of 26 projects funded under the EU’s Seventh Framework Programme between 2008 and 2011. In addition, D-VITA was used to analyze the LAK dataset [20], which includes the full text of all papers published in the conference proceedings of the Educational Data Mining (EDM) conferences 2008-2012, the Learning Analytics and Knowledge (LAK) conferences 2011-2012, and the papers of the 2012 Special Issue on Learning Analytics in the Educational Technology & Society (ETS) journal. Overall, the LAK dataset comprises 280 papers.

4.2 Runtime Analysis

To evaluate the runtime of D-VITA’s offline component, each of the above datasets was processed on the basis of the methods presented in Section 3.2: preprocessing, dynamic topic modeling, computation of similar documents, and topic ranking. The datasets were analyzed using epochs of one year, except the blog dataset, for which we used monthly epochs since the overall time span is only two years. We used 15 topics for the blog and LAK datasets and 10 topics for the other datasets. To ensure comparability, all experiments were conducted on a machine with Intel Core2 Quad 2.83 GHz CPU with Java7 64bit and 8GB RAM.

The results of the runtime analysis are presented in Figure 6. On the horizontal axis the name of the datasets are shown while the vertical axis highlights to the runtime of the different offline processing steps for each dataset. It is apparent that the LDA component is the most time consuming part of the offline component. Its runtime is orders of magnitude longer than the runtime of the other components. The preprocessing, the computation of similar documents, and topic ranking are very efficient to compute with runtimes in most cases lower than one minute. Since topic modeling is the dominating part for the efficiency of our method, we elaborate on this part in more detail.

Besides the general complexity of the statistical approach used in the dynamic topic modeling method, a further analysis showed which factors lead to the longer runtime of this part. Interestingly, the overall number of documents is not necessarily the most important factor; though, the length of each individual document affects...
the runtime. In Figure 6 we see that the runtime of LDA in the case of the LAK dataset is longer than, e.g., for the ECTEL or ICWL dataset. Although the number of documents in ICWL is higher than the number of documents analyzed for the LAK data, the documents in ICWL are shorter and thus contain fewer words: ICWL contains abstracts, while LAK contains full papers.

The relevant factor for the responsiveness of the online component is the latency of the internet connection, since D-VITA will fetch data for user interactions by executing fast queries on the preprocessed data to update the visualization.

4.3 User Studies

This section reports on a study involving end-user evaluations of D-VITA using the datasets described in the previous section. The objective of the study was to determine the usability and usefulness of D-VITA for productive use. The users for the study were recruited from the TEL-Map project.

Each user had to perform two tasks and complete a short questionnaire after each task containing Likert-scale questions about usefulness and usability, test questions to see whether the participant successfully completed the task, and open ended questions for general feedback. The first task used the ICALT dataset with the aim to find the “most representative” ICALT paper. The second task used the TEL blogs dataset with the objective of inspecting the relevance of “analytics” in the blogosphere.

38 participants evaluated D-VITA and completed the questionnaire. Their personal background and their expertise in TEL is shown in Figure 7. The majority of the participants stated to have a high expertise with TEL. In the following the three tasks with the corresponding results are presented.

We will visualize the results via box plots. The items used a Likert scale with values ranging from 1 (strongly agree) to 5 (strongly disagree). The description how to read these box plots is shown in Figure 8.

Task 1: Representative ICALT Paper The objective of this task was to identify the “most representative” ICALT paper. After opening the dataset the task description provided to the participants was given as follows:

- Use the sort buttons to sort the topic list by highest average relevance overall for ICALT.
- Select the topmost topic and identify the point in time with where this topic peaks in relevance.
- Identify the paper abstract (“document”) which is most closely related to this topic at its peak.
- View the abstract (“content”) of this paper and follow the link to the paper on the publisher website.

The functionality of D-VITA and its usability for this task were evaluated based on the users’ agreement with the following statements:

S1.1) Searching papers based on topics is generally relevant to my work in TEL
S1.2) Selecting the data set to work with was easy
S1.3) Sorting the topic list was easy
S1.4) Identifying the topic’s peak was easy
S1.5) Finding the most relevant document was easy
S1.6) Getting to view the paper abstract was easy
S1.7) D-VITA’s three-pane window layout is easy to understand

The results are shown in Figure 9. While the majority of participants were able to use the offered functionalities for identifying the correct results and to analyze the topics and documents, only a small number of participants had difficulties in the usage of D-VITA.

The subtask S1.5, referring to the simplicity of finding the most relevant documents, showed an increase in the number of participant who experienced difficulties. But even in this case, the mean percentile of users evaluated this subtask with “agree”.

Task 2: Blog Posts on “analytics”. In this task, the objective was to inspect the relevance of “analytics” in the TEL blogs dataset. After giving participants an instruction to change the dataset, to filter the topic list using the keyword “analytics” and to finally select all resulting topics, the following subtasks were asked to be performed by participants:

- Inspect the Word Evolution for this topic at the beginning of the topic evolution timeline (very first
marker, early 2011). At which position (if at all) does “analytics” show up in the top-10 word list in the Word Evolution?

- Inspect the Word Evolution for this topic at the last marker of the timeline (end 2012). At which position does “analytics” show up in the top-10 word list?
- Visualize the relative word relevance of “analytics” for the topic. Does the relevance of analytics rise or fall in the end?

After these subtasks, the participants were asked to rate D-VITA based on the following statements:

S2.1) Being aware of relevance of topics/words over time is relevant to my work in TEL
S2.2) Changing the data set to work with was easy
S2.3) Filtering the topic list using a keyword was easy
S2.4) Looking for a particular word in the topic evolution visualization was easy
S2.5) Visualizing the evolution of word relevance for a topic was easy
S2.6) The visualizations help in identifying rising and falling tendency of topics/words

The aim of the tasks refers to the usability of the word evolution chart as well as the topic filtering in D-VITA. While the topic filtering and its functionality received favorable ratings, inspecting the word evolution in topics and looking for a particular word in a topic evolution was a challenge for some participants (cf. Figure 10). The overall result of the evaluation indicates that the majority of participants show a positive satisfaction with the topic analysis functionality of D-VITA.

Overall Assessment. At last the participants were asked for evaluating D-VITA referring to its usage in the participants research in general and the degree of its usefulness regarding to topic and document exploration. The statements to be assessed were:

S3.1) I can imagine productive uses of D-VITA for researchers and/or practitioners
S3.2) D-VITA offers intuitive ways of visual interaction with the data
S3.3) The Topic Explorer was useful

Figure 11 shows the results of the evaluation in regard to the overall assessment of D-VITA. Considering the statements S3.1 and S3.3, it is noticeable that the majority of participants were satisfied with the performance of D-VITA in regard to the topic explorer functionality and its usage in a productive environment. As illustrated by the box plot, even the upper quartiles for these statements show ratings better than 2, i.e. at least “agreement”. Statement S3.2 which refers to the intuitive way of visual interaction, is also evaluated favorably. Most of the participants gave a rating between strongly agree and agree. The 97.5th percentile shows that the users consistently gave positive ratings better than the neutral value of 3.

By asking the participants to state one use case of D-VITA which they can imagine in their work, several points were addressed. One of the most notable points was to search for relevant key ideas by inspect the overall trends in increasing or decreasing research topics, e.g., by analyzing different conference papers. Also searching for documents and their interaction with
different topics were of importance for practical applications for some of participants.

5 Conclusion

In this paper we presented D-VITA, a web-based system for text analysis based on dynamic topic modeling algorithms. The system was designed to support users in visually exploring and interacting with topics and documents contained in large document collections. D-VITA presents an overview of the topics hidden in the documents, highlights the evolution of selected topics, and also presents the evolution of words that establish a particular topic.

To increase the flexibility of the system, D-VITA allows the processing of arbitrary data sources by transformation to a simple target schema. We have deployed and evaluated the D-VITA prototype in the TEL-Map project, which maintains several databases related to the research area of Technology Enhanced Learning (TEL), including TEL related papers, blogs and projects. The runtime evaluation showed that there is space for improving the performance of the offline components, which will be tackled in future work. The usability and usefulness of D-VITA were evaluated in end-user studies involving researchers and stakeholders from the TEL-Map project. The results of this evaluation showed that the participants rated both usability and usefulness of D-VITA highly positively.

Current work on extending D-VITA proceeds along several threads. On the one hand, the system is currently extended community-oriented features to offer user interfaces for managing the data sources, topic models, and access to analysis results based on different user roles in a user community. On the other hand, we are planning to equip the system with a monitoring mechanism to allow for automatic updating of topic models based on updated data sources. Another interesting direction for future research is to apply methods for topic modeling in streams to provide realtime visualizations of topic evolution in web data streams.

References