D-VITA: A Visual Interactive Text Analysis System Using Dynamic Topic Mining

Nikou Günnemann
nikou.gholizadeh@rwth-aachen.de

Computer Science 5 - Information Systems & Databases
Prof. Dr. M. Jarke
RWTH Aachen University, Germany

Abstract: Recent developments in web technologies like Web 2.0 have led to the generation of massive amounts of data. The rapid growth of data makes knowledge extraction and trend prediction a challenging task. A recent approach for the unsupervised analysis of text corpora is dynamic topic mining. While there is a growing interest in using this technique, interactive analysis systems for dynamic topic mining are still in an early stage.

In this paper we present D-VITA, an interactive text analysis system that exploits dynamic topic mining to detect the latent topic structure and topic dynamics in a collection of documents. D-VITA supports end-users in understanding and exploiting the topic mining results, in visualizing the topic dynamics within document collections, and in browsing of documents based on shared topics. We present an application case for a scientific community that uses an instance of D-VITA for trend analysis in their data sources.

1 Introduction

There are many scientific, industrial or even historical issues which are researched and published online. Managing these large amounts of data is beyond human capability. It is almost impossible to gain an overview by hand. A possible solution to create an overview of large collections of documents is to use topic mining. Topic mining classifies the documents in topics on the basis of their content. The already existing documents can be substantially enhanced by publishing new documents. These dynamically changing document collections even complicate the topic mining since the aspects discussed in the stored documents might evolve.

A concrete example is given by the TEL-MAP research project [DK12, DCP+11], associated to the FP7 Cooperation Program of the European Commission, which tackles the challenge of analyzing the landscape of the Technology Enhanced Learning (TEL) research field. There are many technology providers and technology adapters in Europe that have a stake in the TEL research area. These actors are interested in finding topics which have recently attracted attention in the TEL community and are worth to be financially support in the future. Obviously, the recent trends in the TEL area change over time. By analyzing the TEL-MAP Mediabase, which stores information about TEL-related projects,

papers, and blogs [DK12, DCP+11], such trends might be automatically detected and can support the stakeholders decisions.

Overall, gaining an overview over large and dynamic collections of electronic documents is an increasing need in the area of industry and science. On this ground, a line of works has been focused on the development of algorithms for categorizing documents based on their inherent topics [BNJ03, BL06, WBH08].

Grouping of documents in topics by these mining methods is a prerequisite for gaining an overview over documents and finding hidden thematic structures. However, the generated results - huge lists of mere numbers - are still difficult to interpret by the user. Thus, visualizing the result of a topic mining method in a easily comprehensible way is also of fundamental importance. Via visualizing the result we finally enable the analysis of large document collections to interpret the right context of topics in different documents. Such a visualization should also be able to depict the temporal evolution of topics which plays a central role for dynamic data collections.

Based on this motivation we present in this paper an interactive text analysis tool for dynamic document collections. It supports the user, for example, in finding recent trends in a set of documents or visualizing the corresponding results. The remainder of this paper is structured as follows: In Section 2, we discuss relevant related work on topic mining. We then introduce in Section 3 the system architecture of D-VITA and we demonstrate different visualization concepts used in our system. In Section 4 we present results of a preliminary evaluation of D-VITA. We conclude in Section 5 with a summary and an outlook on further work.

2 Related Work

2.1 Topic Mining

Topic mining 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 over words. The basic idea is that documents discussing the same topic also primarily use the same words to a certain extent. Thus, these words can be used to describe the common topic and they give the topic its meaning. To generate an effective topic mining summarization, we have researched for different state-of-the-art summarization techniques for document databases. A prominent method to determine topics is Latent Dirichlet Allocation (LDA) [BNJ03]. Therefore, we employ LDA (and its extension for dynamic data; cf. next section) in our system to summarize topics for the whole database collection. In the LDA model, each document is allowed to belong to multiple topics. More specifically, each document is associated with a distribution over topics, i.e. documents belong to topics with different weights.

Since our system exploits the ideas of LDA, we first review the underlying statical assumption of the topic model described by [BNJ03]. 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 $Dir_K(\vec{\alpha})$ controlled by $\vec{\alpha}$, a positive K-vector. Furthermore, LDA uses a Dirichlet distribution controlled by η to generate the actual topics (distributions over words). We let $Dir_V(\eta)$ denote a V-dimensional Dirichlet with scalar parameter η. The overall generative process of LDA is given as follows:

1. For each topic k,

(a) Draw a distribution over words $\vec{\beta}_k \sim Dir_V(\eta)$.

2. For each document d,

(a) Draw a vector of topic proportions $\vec{\theta}_d \sim Dir_K(\vec{\alpha})$.

(b) For each word,

(i) Draw a topic assignment $Z_{d,n} \sim Mult(\vec{\theta_d}), Z_{d,n} \in 1, ..., K$.

(ii) Draw a word $W_{d,n} \sim Mult(\beta_{\vec{Z_{d,n}}}), W_{d,n} \in 1, ..., V$.

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.

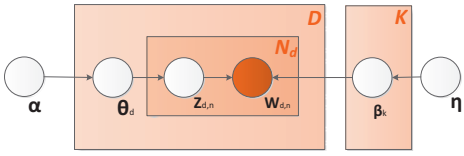

Figure 1: The graphical model of LDA

2.2 Dynamic Topic Mining

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 [MZ05, BL06, WBH08, WM06, HCP+09, LBK09, AX10, JHL11, HYGD11]. We use one of the earliest approaches [BL06], to describe the general idea behind these models.

Considering the dynamic LDA in [BL06], 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 is depicted in Figure 2. Figure 2 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 β_k over words, changes smoothly over time. That is, words important for a topic at a specific 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

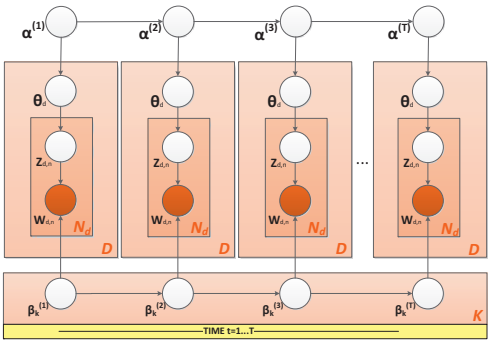

Figure 2: Graphical model of the dynamic LDA approach

smooth evolution between $\beta_k^{(t+1)}$ and $\beta_k^{(t)}$. We refer to [BL06] 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 results of dynamic LDA on different 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 a screenshot of the current version of D-VITA.

3.1 Data Layer

Besides the different kinds of raw data (e.g. blogs, posts, research papers, web content), the data layer stores the results generated by the application layer. This includes the documents in a form preprocessed for dynamic LDA as well as the detected topics. All relevant information is stored in a relational database management system.

3.2 Application Layer

As illustrated in Figure 3, we divide the application layer into two main components for data processing: online and 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,

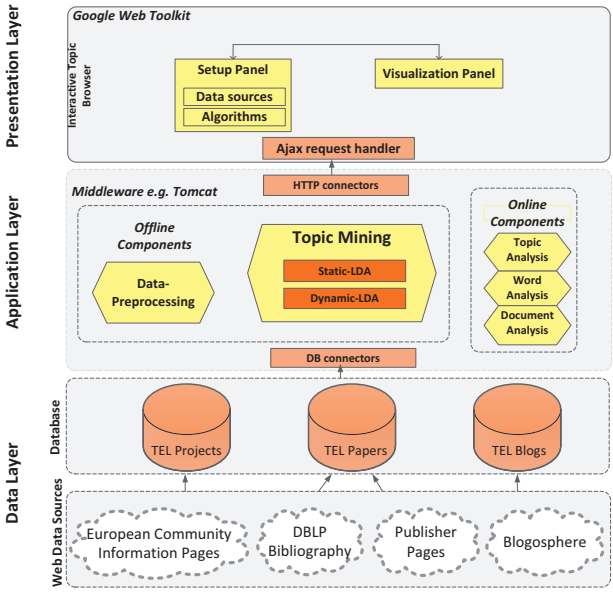

Figure 3: System Architecture of D-VITA

the execution of topic mining 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 TEL Mediabase, for instance, blog posts in the blog database are used as input documents for the topic mining 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 mining results. The user interacts with three general concept (i.e. documents, topics and words). Subsequently, the results are visualized by the presentation layer which is described in Section 3.3.

Technical Realization The application layer is implemented based on Apache Tomcat[1]. Apache Tomcat is an open source software implementation of the Java Servlet and JavaServer Pages technology. It supports the implementation and execution of Java programs (servlets) in a webserver environment. Each servlet provides publicly accessibly methods which can be invoked by the clients. The client-server communication is automatically handled using protocols like HTTP.

[1] http://tomcat.apache.org/

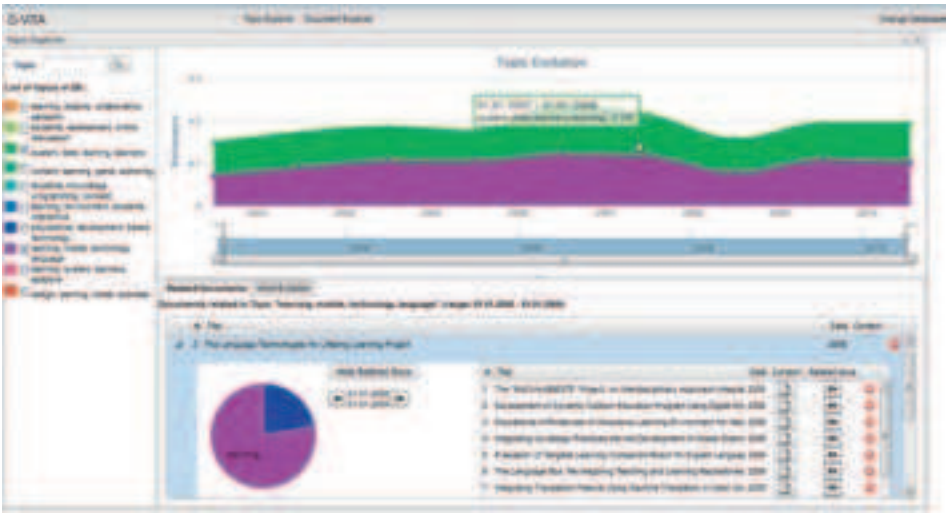

Figure 4: *D-VITA's visual summary of topics. In the visualization, the themeriver (top-right-pane) represents the selected topics and their evolution over different time slices. By selecting (i.e. clicking) a topic in the themeriver, a list of documents exhibiting the selected topic is shown in the bottom-right pane. By selecting a document from this list, e.g. the conference paper "The Language Technologies for Lifelong Learning Project" as in the figure, the distribution of topics in the document is visualized as a pie chart. In the example, the selected document is associated with 77.4 percent to the topic "learning, mobile, technology, language" and 22 percent with the topic "educational, development, based, technology". 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.*

3.3 Presentation Layer

The key idea in representing LDA's results and the correlation between the results is to exploit the paradigm of *Visual Analytics* [KMS+08]. 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 its interpretation is supported by visualizing important information. 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 [CB12] and in TIARA [LZP+12]. Existing systems, however, often fail to support important use cases like handling of dynamic data or the detection and highlighting of emerging topics in the data. Furthermore, none of the existing tools decouples the topic mining 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 schema that can be used by the offline component for data preprocessing.

Presentation of Topic Analysis 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 $k \in 1, ..., K$ can be described by the most occurring words in it. In D-VITA, we consider the most four occurring words to name a topic. Formally, the four most important words $words_k \subseteq \{1, \ldots, V\}$ for topic k can be computed based on its word distributions $\beta_k^{(t)}$:

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

where the function $4 \arg\max$ denotes the generalization of the $\arg\max$ function to return the *four* elements with the highest values. In doing so, D-VITA generates at each start a list of selectable topics from the database. On this way, each user has an overview of topics in the database, as depicted in Figure 4 (left part).

By selecting several topics, the user can analyze the topics evolution, e.g. whether they are emerging or declining. For this purpose D-VITA visualizes the topics by a themeriver graph [HHN00]. 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 time interval. The wider the current, the more relevant is a topic, i.e. the more documents contribute to this topic. Since each document contributes to each topic with a different degree, the relevance of topic k at time t is formally defined as

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

where DB_t is the set of documents belonging to time t and θ_d is the topic distribution for document d. A user can also interact with the content of a certain topic in a dynamic topic view. Each current in the themeriver is selectable to allow an in-depth analysis. This is an another possibility to present topics in conjunction with documents that is explained in the following.

Presentation of Document Analysis In D-VITA document analysis can be used to analyze the topic-document correlations. By selecting a topic in the theme river corresponding to a certain point in time a list of documents is displayed in the "Document Browser" panel (at the bottom). These documents correspond to 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 $Repr_{k,t} \subseteq DB_t$ is determined based on the following equation:

$$Repr_{k,t} := r \arg\max_{d \in DB_t} \{\theta_d[k]\}$$

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

Figure 5: Word evolution pane of D-VITA

represent the share of topics by colored slices in a pie chart as shown in Figure 4. Additionally, the content of each document can be inspected by clicking on the button content from the table.

D-VITA also offers a way to analyze the document-document correlation. To achieve this, we first measure the similarity between the selected document d_1 and the whole documents $d_i \in DB_t$ in the respective time slice by using Jensen Shannon Divergence which is formally defined as:

$$JSD(d_1 \parallel d_i) := \frac{1}{2}D(d_1 \parallel M_i) + \frac{1}{2}D(d_i \parallel M_i)$$

where D is Kullback-Leibler-Divergence and

$$M_i := \frac{1}{2}(d_1 + d_i) \quad , \quad d_i \in DB_t$$

As explained earlier, the computation of similarity scores for pairs of documents is executed in the offline component, while searching for similar documents to the selected document and listing the result is executed by the online component. Note that the related documents might belong to the same time step as the input document as well as to different time steps. Our GUI enables the user to browse between the documents in different time steps (cf. buttons below "Hide Related Documents" in Figure 4). In doing so, a user navigates through future or previous time slices different than the current time slice.

Presentation of Word Analysis Finally, the D-VITA system allows to perform a word analysis. Giving a certain keyword in a search field, D-VITA returns the results in two aspects: as a list of documents and as a list of topics with the keyword in their context. Additionally, based on the properties of dynamic LDA, the word distribution of a topic may change over time. The word distribution depends on the number of produced articles and papers that contain the word. For this reason, each topic evolves in its relevant words. This word evolution can be illustrated in the lower part of our system as an alternative to the document 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.

4 Runtime Metrics

The D-VITA system was deployed using different production databases from the TEL-Map project [DK12]: a DB2 database that stores blogs crawled from the web and an Oracle database that stores data based on DBLP and CiteSeerX. Additionally, D-VITA was tested on an extract of the 20 Newsgroups dataset[2]. The table below shows the runtime of the offline component for each data set.

Database	Size	Time steps	# Topics	Offline Components		
				LDA	Similar docs	Preprocessing
ICALT Papers	2,227	4 (Year)	15	4.1 h	1.9 min	18 sec
20Newsgroup	10,757	9 (Year)	6	16.4 h	9.4 min	2.5 min
Blogs	11,221	23 (Month)	15	6.6 days	10.3 min	9.7 min

Considering the elapsed time for the offline components, our results show that the run time increases depending on different factors. Besides the size of data, the number of time steps on which LDA runs, increases the runtime. LDA requires 6.6 days for the blogs data, which covers a period of 23 time steps, while it needs 4.1 hours for the ICALT papers, covering a period of 4 time steps only. Additionally, a high number of selected topics, influences the run of the LDA algorithm to fit the topics as precisely as possible. This, also increases the runtime as shown in the case of 20Newsgroup and blogs. Note that these experiments were conducted using the single-threaded code provided by the authors [BL06]. By using parallelized implementations of LDA, a higher efficiency can be achieved.

In contrast to the offline components, the online component needs only seconds of time to present the results on demand. The relevant factor for the run time of the online component is the speed of the internet connection since for each query D-VITA executes a connection to the database.

5 Conclusion

In this paper, we present D-VITA, a novel interactive visual text analysis system based on dynamic topic modeling that is designed to support users exploring and interacting with numbers of documents. It presents an overview of the topics hidden in the documents, highlights the evolution of selected topics, and also displays the evolution of words that establish a particular topic. To increase flexibility of D-VITA, we allow to process arbitary data sources. We have deployed a prototype of D-VITA in the TEL-Map project which maintains in its "Mediabase" several databases related to the research area of technology enhanced learning (TEL), including TEL-papers, blogs and projects. The runtime analysis showed that there is room for improving the performance of the offline components, which will be tackled in future work. Currently we are focusing on improving and empirically evaluating the usability of the web application for both data providers and end-users in an empirical study with TEL-Map partners.

[2] http://kdd.ics.uci.edu/databases/20newsgroups/20newsgroups.html

References

[AX10] Amr Ahmed and Eric P. Xing. Timeline: A Dynamic Hierarchical Dirichlet Process Model for Recovering Birth/Death and Evolution of Topics in Text Stream. In *UAI*, pages 20–29, 2010.

[BL06] David M. Blei and John D. Lafferty. Dynamic topic models. In *ICML*, pages 113–120, 2006.

[BNJ03] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.

[CB12] Allison June-Barlow Chaney and David M. Blei. Visualizing Topic Models. In *ICWSM*, 2012.

[DCP+11] Michael Derntl, Adam Cooper, Manh Cuong Pham, Ralf Klamma, and Dominik Renzel. Mediabase Ready and First Analysis Report, TEL-Map Deliverable D4.3. 2011.

[DK12] Michael Derntl and Ralf Klamma. A Mediabase for Technology Enhanced Learning in Europe. *IEEE Learning Technologies Newsletter*, 14(3):2–5., 2012.

[HCP+09] Qi He, Bi Chen, Jian Pei, Baojun Qiu, Prasenjit Mitra, and C. Lee Giles. Detecting topic evolution in scientific literature: how can citations help? In *CIKM*, pages 957–966, 2009.

[HHN00] Susan Havre, Elizabeth G. Hetzler, and Lucy T. Nowell. ThemeRiver: Visualizing Theme Changes over Time. In *INFOVIS*, pages 115–123, 2000.

[HYGD11] Liangjie Hong, Dawei Yin, Jian Guo, and Brian D. Davison. Tracking trends: incorporating term volume into temporal topic models. In *KDD*, pages 484–492, 2011.

[JHL11] Yookyung Jo, John E. Hopcroft, and Carl Lagoze. The web of topics: discovering the topology of topic evolution in a corpus. In *WWW*, pages 257–266, 2011.

[KMS+08] Daniel Keim, Florian Mansmann, Jrn Schneidewind, Jim Thomas, and Hartmut Ziegler. Visual analytics: Scope and challenges. *Visual Data Mining*, pages 76–90, 2008.

[LBK09] Jure Leskovec, Lars Backstrom, and Jon M. Kleinberg. Meme-tracking and the dynamics of the news cycle. In *KDD*, pages 497–506, 2009.

[LZP+12] Shixia Liu, Michelle X. Zhou, Shimei Pan, Yangqiu Song, Weihong Qian, Weijia Cai, and Xiaoxiao Lian. TIARA: Interactive, Topic-Based Visual Text Summarization and Analysis. *ACM TIST*, 3(2):25, 2012.

[MZ05] Qiaozhu Mei and ChengXiang Zhai. Discovering evolutionary theme patterns from text: an exploration of temporal text mining. In *KDD*, pages 198–207, 2005.

[WBH08] Chong Wang, David M. Blei, and David Heckerman. Continuous Time Dynamic Topic Models. In *UAI*, pages 579–586, 2008.

[WM06] Xuerui Wang and Andrew McCallum. Topics over time: a non-Markov continuous-time model of topical trends. In *KDD*, pages 424–433, 2006.