2019 23rd International Conference Information Visualisation (IV)

Visual Analytics for Analyzing Technological Trends from Text

1st Kawa Nazemi
Human-Computer Interaction & Visual Analytics
Darmstadt University of Applied Sciences
Darmstadt, Germany
kawa.nazemi@h-da.de

2nd Dirk Burkhardt
Human-Computer Interaction & Visual Analytics
Darmstadt University of Applied Sciences
Darmstadt, Germany
dirk.burkhardt@h-da.de

***Abstract*—The awareness of emerging technologies is essential for strategic decision making in enterprises. Emerging and decreasing technological trends could lead to strengthening the competitiveness and market positioning. The exploration, detection and identification of such trends can be essentially supported through information visualization, trend mining and in particular through the combination of those. Commonly, trends appear first in science and scientific documents. However, those documents do not provide sufficient information for analyzing and identifying emerging trends. It is necessary to enrich data, extract information from the integrated data, measure the gradient of trends over time and provide effective interactive visualizations. We introduce in this paper an approach for integrating, enriching, mining, analyzing, identifying and visualizing emerging trends from scientific documents. Our approach enhances the state of the art in visual trend analytics by investigating the entire analysis process and providing an approach for enabling human to explore undetected potentially emerging trends.**

***Index Terms*—Visual Analytics, information visualization, trend analytics, emerging trend identification, visual business analytics**

I. Introduction

Technological developments have an essential impact on strategic decision making. The early awareness of possible upcoming or emerging technological trends could lead to strengthening the competitiveness and market positioning of enterprises. However, if innovation-driven companies would ignore emerging technological developments, they may not tap the full potentials of their own products or technologies. Larger enterprises already consider this essential aspect and create so called ”Innovation Centers” that try to foresight future developments and upcoming innovations. Commonly either fully automated predictions are performed or manually created ”technology-roadmaps”. A combination of the humans' intelligence with machine learning methods are rarely implemented in this process, although Visual Analytics provides exactly this combination [1]. Currently existing information systems make use of different information retrieval methods in combination with different algorithms to extract trends from text. However, the related visual representations do not really enable the exploration, identification and detection of emerging trends for inferring future technological developments. The interactive overview on data, the continuous changes in data, and the ability to explore data and gain insights are essential to identify upcoming trends. Another important aspect for identifying emerging technological trends are the data. Social media, news, company reports and blogs refer commonly to those technologies that already reached their climax or are already available at the market. Early technological trends are often propagated first in research and scientific publications. Therefore, these data are the ”real” information pool for early signals and trends [2]. Although, scientific publications and their value for identifying early trends is obvious, a real analysis and in particular identification of emerging trends out of textual scientific publications are rarely proposed. The gathering and analysis of this continuously increasing knowledge pool is a very tedious and time consuming task and borders on the limits of manual feasibility.

We introduce in this paper an approach for integrating, enriching, mining, analyzing, identifying and visualizing emerging trends from scientific documents. We will first introduce the state of the art in trend mining, and trend and text visualization. Thereafter, we introduce our general approach that investigates the several steps and give an overview of the architecture. Based on this model, we will introduce each step of our model beginning with indexing, data enrichment, data mining over trend identification and data modelling to interactive visualizations. Our contribution is three-fold: (1) a model for gathering trends from text to visual interactive analysis representations, (2) the identification of upcoming or emerging trends based on text, and (3) an approach for visual interaction through different data models and related interactive visual representations to explore the potentials of technologies and detect new insights. With our contributions the process of trend analysis is supported in a more efficient and effective way and enables finding undetected patterns in data.

II. Related Work

The approach presented in this paper uses trend mining approaches from text, and trend and text visualizations. We therefore introduce the related work in trend mining and trend and text visualization. These two categories are strongly related to each other and can not be distinguished strictly. So each section may include approaches from both areas. Whereas

2375-0138/19/$31.00 ©2019 IEEE
DOI 10.1109/IV.2019.00041

we carefully tried to focus on the main contributions and assign the works based on the proposed scientific contributions.

A. Trend Mining from Text

Discovering trends from text was assigned as one of the most important tasks in the early works of text mining also known as Knowledge Discovery from Text (KDT) [3]. In these early states, the problem of identifying trends was described as the task of discovering deviations from expected models that are constructed from past data. Building on this advance towards trend mining, different approaches arose that aimed at identifying key topics and discovering their relevance over time. One of the first approaches for discovering trends in text was proposed by Lent et al. [4]. They defined a trend as a sequence of frequencies of a specific phrase. For identifying these kinds of trends, the document corpus was divided in several temporal sets. After this segmentation the key phrases were extracted with Sequential Pattern Mining (GSP-Algorithm) [5] and a history of each phrase was generated as a sequence of their occurrences per time interval. With this representation, trends could be identified by means of shape queries [6]. Users were able to formulate specific shapes of a trend and receive the corresponding phrases that comply with these shapes. *Trend Graphs* [7] was introduced by Feldman et al. [7] that provides an overall picture of all major trends and focuses on concept relations and their evolution and define a trend as a change of the relations between the terms in the corpus in a specific context, rather than sequences over time. They define a trend as a change of the relations between the terms in the corpus and the given specific context, for which they introduce the notion of a *Context Graph*. The vertices of a Context Graph correspond to terms found in the documents that are connected with an edge if both terms co-sufficiently occur in the given "context". Based on this notion a *Trend Graph* is generated from documents in a certain time interval and the trends are indicated by means of different edge representations according to the predecessor graph and the type of the identified trend. Their approach incorporates first techniques of information visualizations for representing trends in a temporal manner. Montes-Gómez et al. distinguished between *change* and *stability* trends and introduced analysis methods for identifying the key topics that contribute to a trend between two time intervals in a document collection [8]. The starting point for the trend discovery was a normalized topic vector that was extracted from the documents of each time interval. Based on this representation and a symmetric similarity measure for comparing the topic distributions, they defined different metrics for discovering key *Change* and *Stability* Factors and introduced the concept of *Topic Deviations* that allows the identification of anomalous instances that do not fit in the standard case. *BlogPulse* was introduced as a toolkit for analyzing online collections of time-stamped documents in order to automatically detect trends [9]. The system harvested daily blog articles and transferred the text documents in sets of tokens that include different types of annotations (e.g. part-of-speech-tags, sentence boundaries, etc.) that facilitate further processing steps. Beside the extraction of key phrases and key persons per day the system extracted key paragraphs that include the majority of phrases of identified key topics for indicating the current trend in the blogging community. It also included a trend search tool that visually showed the hits of a search query over time.

Mei and Zhai introduced two methods for discovering evolutionary theme patterns in text [10]. Their methods are based on a set of salient themes that are extracted from temporal sets of documents using a probabilistic mixture model [11]. Based on these themes per time interval they present two different methods for identifying latent trends. The first method called *Theme Evolution Graphs* represents the change of themes over time in a graph constructed with the themes as vertices and the relations determined using the Kullback-Leibler divergence. The evaluation on two different data sets shows that this method is suitable for discovering how themes in one time period influence other themes in later periods and to provide a temporal overview. The second method called *Theme Live Cycles* is based on Hidden Markov Models (HMM) and aims at discovering globally interesting themes. The HMM is used as a generative model for determining the strength of trans-collection themes over time. The evaluation reveals that *Theme Live Cycles* allows users to not only see the trends of strength variation but also provides a method for comparing the relative strengths of different themes over time. In contrast to the approach from Mei and Zhai, Viermetz et al. utilized temporal granularity and a density-based clustering algorithm for extracting short and long term topics as keyword vectors [12]. Each topic is described as a keyword vector that is transformed to a representative keyword vector by contrasting the foreground model corpus to a background model. By linking short term to long term topics the method allows to discover topic trends as the evolution of long term topic clusters, expressed by the emergence and disappearance of short term topics. Kim et al. presented an approach for discovering technology trends from patent texts [13]. They defined a technology trend as several salient technologies sharing the same problem or solution. Their approach is divided into the two steps of (1) semantic key-phrase extraction and (2) technological trend discovery. In the first step key-phrases in each patent are identified that are classified either a solution or a problem using a learned *Support Vector Machine* (SVM). In the second step time spans and their most salient trends are identified for discovering technology trends. However their approach is primarily applicable in the specific domain of patent texts. Goorha and Ungar presented an approach that discovers emerging trends in text collections by identifying significant phrases associated with user defined known products, companies or people of interest [14]. The system extracts key-phrases found in the context of user specified keywords and ranks the interestingness of these key-phrases with an empirically verified formula. The results are visualized in a scatterplot that represents the significance of emerged trends over a specific time period. *Tiara* is an approach that allows the identification of topic-related trends.

It utilizes the *Latent Dirichlet Allocation* (LDA) for generating a topic model [15]. *Tiara* calculates the strength of the topics over time that is visualized in a stacked graph for identifying peaks and slopes of each topic. Nguyen et al. proposes a two-step approach for detecting hot topic and technology trend tracking in patent data [16]. In a first step the terms are extracted through TF*PDF [17] and in a second step the variation of the extracted terms are measured over time. This second step is based on the works of Chen et al. [18] who proposed an aging theory with the four cycles of birth, growth, decay, and death. Nguyen et al. adapt the proposed calculations to get the Energy of a topic, which is defined by the frequency of a term in a specific time slot and indicates if a term is hot followed by the energy function that converts a term energy value into a life support value. Hurtado et al. proposed an approach for topic discovery and future trend forecasting in text documents using sentence level pattern mining [19]. For topics discovery they introduce a fine granular process for converting sentences into a transaction format and association rule mining to discover frequent patterns in the document. They use association rules for correlating topics to each other and Pearson's correlation to correlate the topics with the temporal dimension. Their forecasting is based on linear regression with the assumption that all topics correlates to each other. The results are visualized as nodes and edges, whereas shaded regions denote to communities with strongly correlated topics. The introduced visual representation does not really allow to gather the trend of the extracted topics, thus the temporal dimension is not visualized.

B. Trend and Text Visualization

Current trend mining methods provide useful indications for discovering trends. Nevertheless, the interpretation and conclusion for serious decision making still requires the human and his knowledge acquisition abilities. Therefore, the representation of trends is one of the most important aspects for analyzing trends. Common approaches often include basic visualization techniques. Depending on the concrete results, line graphs, bar charts, word clouds, frequency tables, sparklines or histograms are utilized to impart different aspects of trends. *ThemeRiver* represents thematic variations over time in a stacked graph visualization with a temporal horizontal axis [20]. The variation of the stream width indicates the strength of a specific topic over time. A similar approach for visualizing trends is utilized in *Tiara* with the difference that it includes additional features like magic lenses and an integrated graph visualization [15]. *ParallelTopics* [21] includes a stacked graph for visualizing topic distribution over time. Although the system was not designed for discovering trends but rather for analyzing large text corpora, it allows users to interactively inspect topics and their strengths over time and thus allows the exploration of important trend indicators in the underlying text collection. *Parallel Tag Clouds* (PTC) is based on multiple word clouds that represent the contents of different facets in the document collection [22]. Temporal facets can be used to identify the difference of certain keywords over time and to infer the dynamics of themes in a text collection. Another extension of word clouds is *SparkClouds* that includes a sparkline for each word [23]. These sparklines indicate the temporal distribution of each term and allow conclusions about the topic trends. A user study reveals that participants are more effective with SparkClouds compared to three other visualization techniques in tasks related with trend discovery [23]. A similar approach [24] also includes co-occurrence highlighting. In contrast to SparkClouds, this technique includes a histogram for representing the temporal relevance of each tag. Additional overlays in the histograms show the co-occurrences over time for a selected word to enable a more comprehensive analysis of trend indicators. Han et al. introduce with *PatStream* a visual trend analysis system for technology management [25]. Their system measures similarity between pairs of patents using the cosine metrics and extents the work of Heimerl et al [26] in particular in regards of visualization. The evolution and structure of topics that indicates the trends is visualized through a *Streamgraph*, which was already proposed in the previous works of Heimerl et al. [26]. In contrast to this previous work, *Patstream* breaks down the streams into vertical time slices, which represents periods of years. These time slices are based on their introduced concept that uses the term score, the ratio between the radiative frequency of a term in the given patent collection and its relative frequency in a general discourse reference corpus [25]. Although, their concept makes use of term frequencies, *title score* and *claims score* [25], the most useful approach seems to be the term score, thus it relies on a relative score and investigates the entire document or patent corpus. The topic stream visualization is similar to a stacked-graph with included term (topics) in the area-based visual representation. As they hierarchically cluster patents according to their textual similarities, users are able to zoom-in into a cluster through a level-slider. Beside the main visual representation, the stream visualization, they provide four further visual representation, such as a scatterplot with brushing and linking [25].

The most advanced interactive visual representation is *PatStream*. It provides more than one view, makes use of relative scores and co-occurrences and visualize the temporal spread of the topics with the related categories. However, the approach does not really visualize emerging trends, overview of upcoming trends in a certain field and does not support an in-depth analysis through users' interaction.

As the literature review in both fields revealed, there are already a number of algorithms for gathering trends from text and different approaches to visualize the extracted terms and trends. We could outline that the existing approaches are commonly using different algorithms and approaches to define a trend as hot or emerging, whereas commonly the frequency of terms indicate this information. Thereby it is not really considered that scientific publications, patents and user generated content are increasing rapidly. This correlation between the frequency and changing amount of documents is not really investigated in the calculations. From the visualization point of view the systems are commonly designed to illustrate either

the trend or frequency of terms (even without the temporal dimension). An analytical visualization system that enables the human to analyze through different data models and visual structures could not be found.

III. General Approach

Our goal is to enable an exploratory approach to identify trends and their probable future potentials through users' interaction with different visual structures that enable viewing the data from different perspectives. Thus, we focus our attention primarily upon enabling users to interactively gather an overall topic trend evolution and different perspectives (e.g. geographical or semantic) on data to inspect and analyze potential technological trends with regards to the following questions: (1) *when* have technologies or topics emerged and when established? (2) *where* are the key-players and key-locations, (3) *who* are the key-players, (4) *what* are the core-topics (4) *how* will the technologies or topics probably evolve, and *which* technologies or topics are relevant for a certain enterprise or application area? Addressing these question requires the ability to get an overview of core topics that are relevant at the moment, navigate through the different perspectives, analyze the results, and reason about probable evolving of trends. Based on these requirements, we developed our approach with the following steps (see Figure 1).

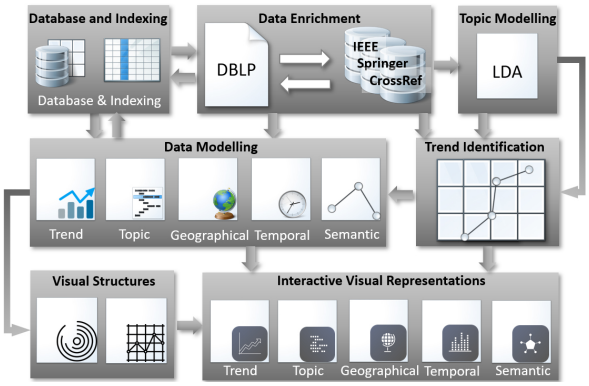

Fig. 1. Our approach with the main steps of *Indexing*, *Data Enrichment*, *Topic Modelling*, *Trend Identification*, *Data Modelling*, and *Interactive Visual Representation*.

IV. Data Indexing

On the server-side there is a database with indexing functionality that stores all relevant information. These information are gathered through the initial data of DBLP that provides some rudimentary metadata in the area of computer science and related fields. Through the *Document Object Identifier* (DOI) the data are enriched with data from several additional databases. The enriched data are used to extract and model topics that enables identifying trends and in particular emerging trends. Based on the stored data, the continuously enriched results and the identified trends data models are generated and stored in the database. The data models are the foundations for choosing visual structures according to Card et al. [27] and enable in the last step to either choose appropriate interactive visualizations or a juxtaposed visual dashboard for the above mentioned tasks and questions.

V. Data Enrichment

For a proper analysis of the given data, enhancements of data quality are necessary. To enhance the quality of data, we first use *Data Enrichment* techniques to gather additional data from Web. The data collection used as basis is a combination of multiple different data sets. The individual data sets offer data of varying quality and content in terms of available meta information. We use in our approach the DBLP data set as initial data pool with about 4.5 million entries. All entries in this data set are without text, e.g. abstracts or full texts, which are necessary to enable an analysis of trends with information extraction methods. We therefore balance out the limitation of the orginial data basis of DBLP by augmenting the available data with additional information for each publication. For this purpose, the system has to figure out, where data resources are located on the Web or which online digital library has more information about a certain publication. We integrate data from *Springer*, *IEEE*, *IEEE Computer Society*, and *CrossRef*. The basic data collection contains a link to the publisher's resource and is used to identify the digital library and location of additional information. These information can be gathered either through a web-service or crawling techniques. The resulting response of web-services is well structured and commonly contains all required information, while crawling techniques require a confirmation of robot policies and the results have to be normalized. But the data may contain duplicates, missing or faulty data. Therefore, common data cleansing techniques are applied. With this step we enrich the data of DBLP with further metadata including abstracts and text directly from the publishers and include some citation information through *CrossRef* that should enable to identify the most relevant papers in a field with regards to citation count.

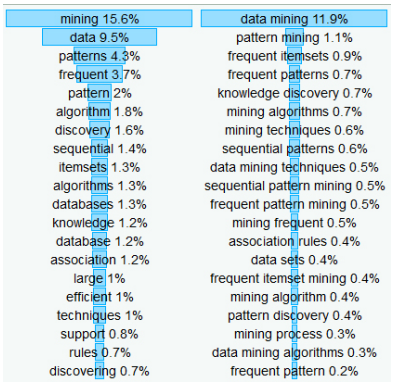

Fig. 2. Example of an automatically generated topic "data mining" with the related words and phrases (N-Grams).

VI. Topic Modelling

In the previous step, we gathered at least abstracts for a major part of the DBLP entries and some open access full text for some entries from general public source like CEUR-WS or the *Springer* database. Based on these enriched data we are able to perform information extraction from text to generate topics. For topic generation, we apply learned probabilistic topic models as topic classification. Studies have shown that this approach is a viable alternative and can even outperform subject heading systems when evaluating similarities between documents clustered by both systems [28]. For this purpose, we have integrated in the step of *Topic Modelling* the Latent Dirichlet Allocation (LDA) [29] with the major advantage of a full automatic topic classification and assignment. Since one classifier is in control of assigning all the topics, the classification is done consistently across all publications. There is no need for any kind of normalization process for the generated topics. The amount of documents have a significant impact on the accuracy of the resulting model. In this step, each document gets assigned to one or multiple topics, which are typically represented by the top 20 used words within the topic. Additionally, we also generate most used phrases for each topic in form of N-Grams. In summary 500 topics are generated with 20 word and 20 phrases through 4000 iterations of the LDA-algorithm. As facets, they offer a great way to filter search results based on the extracted topics. Further, they are used to construct the *Topic Model*. Figure 2 illustrates an example for the topic generation. Thereby the words and phrases for the topic "data mining" is visualized.

VII. Trend Identification

In the second step, we extracted topics with the LDA-algorithm across all publications in the database. If we would try to identify trends based on the frequency of the topics over the years, we would not get any appropriate trends. Nearly the number of all topics will increase through the years. This is just because the number of publications increased in the last years dramatically. Table I illustrates the real number of publications at the time of writing this paper in the DBLP database for every four years.

TABLE I
Number of publications in DBLP every four years

	Documents in DBLP					
Years	1998	2002	2006	2010	2014	2018
Documents	58875	89547	165596	218991	275790	308448

To get the real trends over time, the first step is the normalization of the topic frequencies. We therefore calculate for each year the normalized number of documents containing a topic. Let d_y be the total number of documents in a year y, and t_y is the number of documents in year y that contain a certain topic t. Then, $\tilde{t}_y$ is the normalized topic frequency in the given year y, and is computed as

$$\tilde{t}_y = \frac{t_y}{d_y} \tag{1}$$

After having the normalized frequency of documents containing the topic, the entire years with documents with a certain $\tilde{t}$ are split into periods of a fixed length $x > 1$, limiting the length of the period to the time of the first occurrence of the topic, if necessary. So at the current year y_c, each period p_k covers the previous years $[y_c - x \cdot (k+1), y_c - x \cdot k]$. For example, in the year 2019, for $x = 5$, we have the periods $p_0 = [2015, \ldots, 2019]$, $p_1 = [2010, \ldots, 2014]$, $p_2 = [2005, \ldots, 2009]$, up to the period where the topic appeared for the first time.

For each period, we calculate the regression of the normalized topic frequencies, and take the gradient (slope) as indicator for the trend. The following equation (2) calculates the slope for a topic t in a period p_k, based on the normalized topic frequencies $\tilde{t}_y$, where $\bar{t}$ is the mean of the normalized topic frequencies and $\bar{y}$ is the mean of years in the time period.

$$b_{\tilde{t},k} = \frac{\sum_{y \in p_k} (y - \bar{y}) \cdot (\tilde{t}_y - \bar{t})}{\sum_{y \in p_k} (y - \bar{y})^2} \tag{2}$$

Each calculated slope $b_{\tilde{t},k}$ is weighted through two parameters. The first parameter is the coefficient of determination R_k^2 of the regression. The second parameter is a weight ω_k that is determined with a function that decreases for earlier periods.

For example, the weight ω_k that is used for one period can be defined using a linearly decreasing function:

$$\omega_k = max(0, 1 - \frac{k}{4}) \tag{3}$$

This means that the weight is 1 for the most recent period p_0, decreases linearly about 0.25 for each earlier period, and becomes 0 for period p_4 and beyond.

Alternatively, the weight can decrease exponentially:

$$\omega_k = \frac{1}{2^k} \tag{4}$$

In this case, the weight is 1 for the most recent period p_0, then 0.5 for period p_1, and 0.25 for period p_2.

The final weighting for a topic t is then computed from the slopes $b_{\tilde{t},k}$, the coefficients of determination R_k^2, and the weights ω_k of each of the K periods as follows:

$$\omega = \frac{1}{K} \cdot \sum_{i=1}^{K} b_{\tilde{t},k} w_k R_k^2 \tag{5}$$

We have integrated both, the linear and the exponential measurements for the weight ω_k and are evaluating those through two different systems. The first results shows that the linear calculation seems to lead to more appropriate results, due to the fixed time periods of overall 20 years.

The weighting of the trends, the slopes in different time periods and the regression allow us to identify trends with better results compared to trend identification methods illustrated in the literature review, although the method is quite simple.

VIII. Data Modelling

The *Data Modelling* step of our approach aims at creating data models according to Card et al. [27] for different aspects of the data that are relevant in the analysis process. We identified in Section III questions that should be answered through the interaction with our system. The aspect-oriented data models focus on those questions with the particular aspects that are given in the data. Therefore, we generate five data models, *Semantics Model*, *Temporal Model*, *Geographical Model*, *Topic Model* and *Trend Model*. The basis for the creation of the models are the *Enriched Data* and the *Trend Identification*. The generation of a semantic data model [30] serves as the primary data model for holding all information. It adds structure and semantics to the data for easier extraction of needed information in order to generate the visual representations. This model is mostly used in the textual list presentation, where all available information about each publication needs to be presented, and the generation of facet information for filtering purposes. To accomplish this, a data table is generated including all publications with their attributes and relations.

The temporal data model is used by multiple temporal visualizations. Here, multiple aspects of the information in the data collection need to be accessible based on the time property. For the overview of the whole result set in a temporal spread, the temporal model needs to map publication years to the amount of publications in the certain year. This temporal analysis is not only necessary for the entirety of the available result set, it is also important to analyze specialized parts of faceted aspects. Based on these faceted attributes, detailed temporal spreads need to be part of the temporal model for all attributes of each facet type. The temporal spread analysis needs to be available for each facet in the underlying data. With these information the temporal visualizations can be built more easily. These are then able to show a ranking over time, or demonstrate comparisons of popularity over time. The geographical data model holds the geographical aspect of the available data. The complexity of this model is lower than that of the temporal model, as the geographical visualization only needs quantity information for each country. The data in this model provides the information about the origin country of the authors of publications. Although, the data are enriched with information from different databases as described, there are a lot of data entities without the country information. To face this problem, we introduced two approaches, if no country information could be gathered: (1) we take the affiliation of the authors for gathering the country and (2) we take publications from the same author from the same discipline based on the extracted topics and the same year plus and minus one to assume the country. The year of publication is important, thus many researchers change the affiliation and with the affiliation the country. The topic model contains detailed information about the generated probabilistic topic model. The semantic model contains publications with all assigned properties and relations, which includes topics, and the topic model supplements this data by offering insights into the assigned topics. Like already mentioned in Section VI, the information about each topic contains the top 20 most used words and phrases with the assigned probability of usage for each word (see Figure 2).

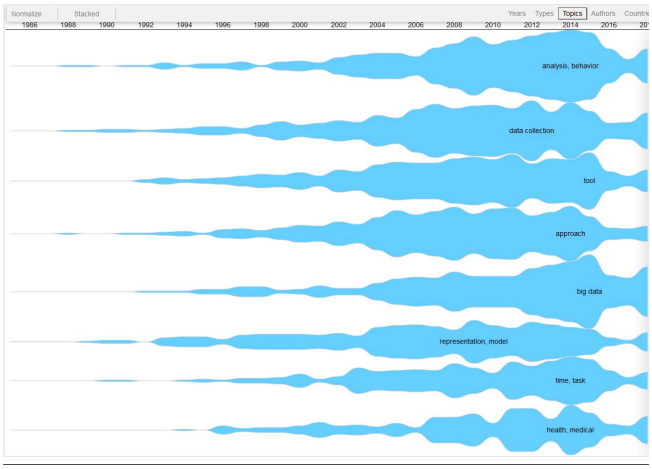

Fig. 3. Example for temporal spread of topics for the search term ”Information Visualization”.

The inclusion of most used phrases can help the user immensely in the reformulation of the search query to find additional information on topics of interest. But the main purpose of the topic model is to gather relevant information about technological developments and the used approaches within a development. The topic model is commonly correlated to the temporal model and provide also the temporal spread of topics. Figure 3 illustrates the temporal spread of topics related to the search term ”Information Visualization”. The trend model is generated through the trend identification process described in Section VII in combination with the temporal model. It illustrates the main trends either as an overview of ”top trends” identified through the described weight calculation or after a performed query. In the second case the same procedure is applied with the difference that the document corpus is not the entire database but only the results referring to the queried term. Figure 4 illustrates our overview visualization of trends for the entire document corpus of the DBLP database. The visualization method is based on the introduced method of *SparkClouds* [23] with nine temporal clouds and a list of all identified trends (right) ranked through the measured weight.

IX. Interactive Visualizations

A. Visual Structure

The visual structure enables a full automatic selection of visual representations based on the underlying data model. We applied the procedure of visual adaptation according to Nazemi [31, p. 256] with the three steps of *semantics*, *visual layout* and *visual variable*. As proposed in [31], we start the visual transformation for generating a visual structure with the semantics layer. Thus our system is not yet adaptive, we

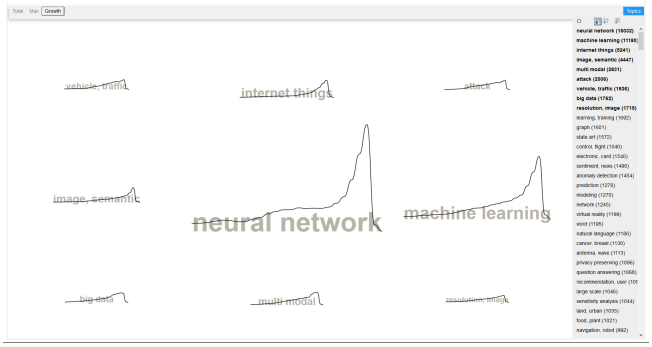

Fig. 4. Example for most emerging trends in the entire document corpus based on the *SparkClouds* approach.

investigate the data characteristics for choosing an appropriate visual layout. Based on the chosen visual layout, we defined a number of visual variables according to Bertin [32] that are applied to a certain visual layout. This procedure allows us to enhance the system with an adaptive behavior and reduces the complexity of integrating new visualizations.

B. Visual Interaction

Interaction with the system should enable users to gather the required information regardless of their domain knowledge or knowledge about the system. We have therefore applied two complimentary approaches that support users' information acquisition and analysis process. First, the *information seeking mantra* by Shneiderman [33] is applied to provide an overview followed by zoom and filter and then details on demand. This procedure allows to see either emerging trends as illustrated in Figure 4 or the most recent search terms typed by other users as a wordcloud to gather first information of the underlying data and interact until they reached their intended goals.

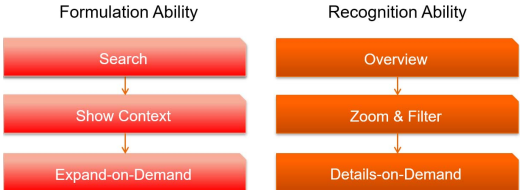

Fig. 5. The complementary interaction approaches applied in our system.

Second, the approach proposed by van Ham and Perer [34] was applied that enables searching the database, getting the context and get more detailed information. Although, the second approach was designed for graph exploration, we think that the entire process of interacting with visualizations can profit by this approach, due to its complimentary interacting process compared to the overview-first approach. Figure 5 illustrates both approaches in an abstract way.

With this procedure users are able to start their analytical tasks and get either after querying a search term or clicking on a trend or a search phrase, first a temporal overview of the entire documents in the database. Figure 6 illustrates one initial

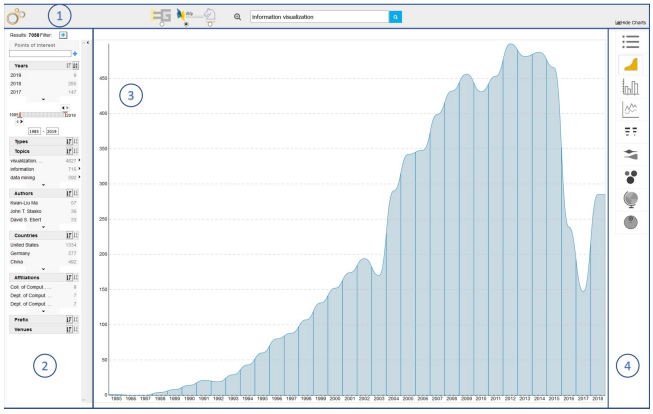

Fig. 6. The UI of our system with its four areas.

screen with the entire user interface and the temporal overview of the underlying documents after a query for "Information Visualization!".

The user interface is built with four main areas. At the top (1), users are able to search (including advanced search formulation capabilities), activate assisted search, select a database or hide the visualization selection area. The assisted search is implemented according to our previous work [2] that enhances users' performed query based on the resulted top five phrases of the top ranked topic [2]. At the left (2) the facets of the underlying data are generated and visualized dynamically. This area also includes number of results, that is automatically adapted to the selected facets, a logical facet selection, and the search-in-search functionality (see Figure 9). The logical facet selection allows users to reduce the amount of the results to get the most appropriate documents for a certain task. In Figure 7 the user has selected after the query for "Information Visualization" the facets *data mining*, *social networks* and *government, public* and combined them with a logical OR-operator with other facets. The amount of the results are reduced from 7058 to 8. The list-view shows the final result set of documents.

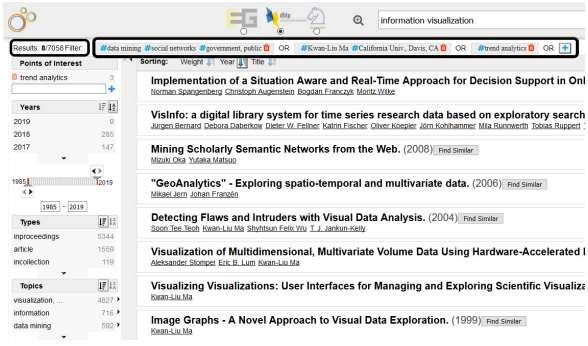

Fig. 7. Facets with the list-view. With highlighted facets chosen by the user and the number of results.

In the center area (3) the main visualization(s) are placed

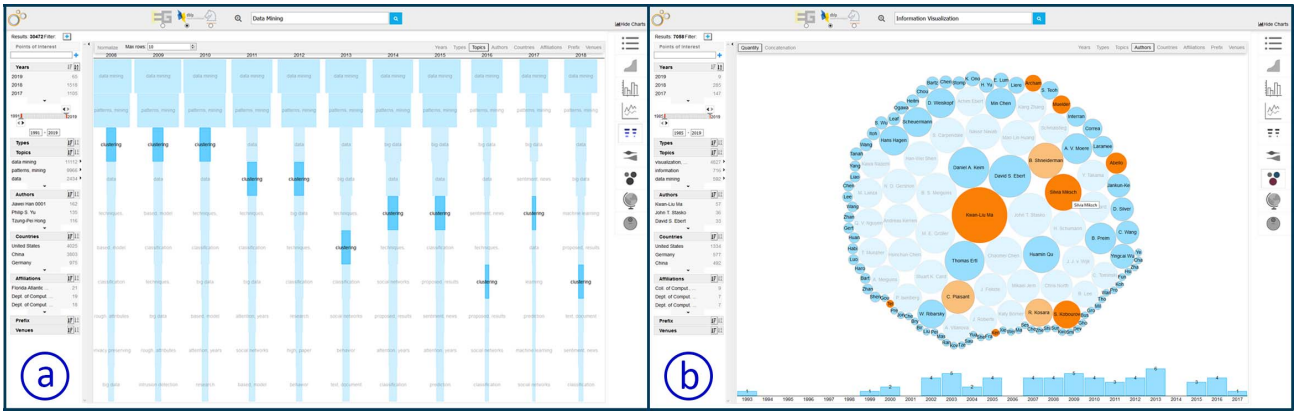

Fig. 8. The visual POIs for enabling a search within the result set.

that are either automatically selected by the type of data, the search query (see Section VIII) or by the user himself in the right area (4), where a dynamic set of visualization are available based on the data and theirs structure. In Figure 6 the temporal overview of the entire data is visualized while the Figure 7 illustrates a list-view with a small number of documents that are refined through the facets.

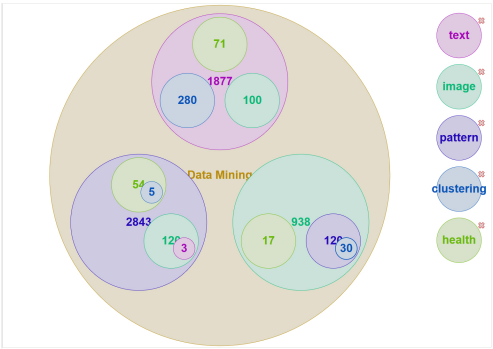

Fig. 9. The visual POIs for enabling a search within the result set.

The search-in-search functionality allows users to formulate terms that are relevant for him, create so called visual *points-of-interests* (POI) and see at a glance the number of documents with that contain the created points of interest. In Figure 9 the user searched for ”data mining” and created a number of visual POIs. The defined POIs are visualized at the right side and can be included into the main search term ”data mining” per drag and drop. The color is the indicator for a certain POI and allows users to see how many publications are in the data base with the created POIs. The number represents the search result quantity within the search result set, so the user is able to define and redefine such POIs for his purposes. In Figure 9 the main search term is ”data mining”. The user is able to see at a glance the results for data mining documents containing the text, image, pattern, clustering or health. With a nested method they are able to see that there are 938 publications containing the phrase ”Image” withing the result corpus ”Data Mining” (circle at right-bottom) with 17 publications about data mining of images for the domain health, 120 results containing pattern and within this search 30 documents that are using clustering methods.

C. Visual Representations

As described in Section VIII, we integrated several data models that enables users to interact with different aspects of the underlying data. The integrated data models allow us to provide several interactive visual layouts that enables information gathering from different perspectives. A simple temporal visual layout for overview purposes that visualizes the amount of documents over the years of the entire search results is illustrated in Figure 6. Another temporal visual layout is the stacked chart consisting of two configuration areas and a view area for the visual layout (see Figure 10). The first configuration area (on top) allows choosing facet type, e.g. topics, authors, countries, affiliations etc. for visualization. After the selection of the facet type, the second configuration (on right) allows to select the number of visualized entities. It lists all available items for the chosen facet type (in that example topics related to the search query). Although the stacked chart is a well established visual layout for temporal data, the perception quality might get difficult, if more information entities are illustrated. The differences between multiple data sets or even changes within the same data-set over time might get difficult to identify. We therefore integrated beside the stacked layout, the temporal river layout that separates all the topics and trends for a more comprehensible view. Instead of layering (stacking) the items on top of each other with no space between them, we represent each facet item with a ”river”. Figure 5 illustrates this. Each river has a center line and a uniform expansion to each side based on frequency distribution over time. Additionally, placing multiple rivers next to each other makes spotting differences in temporal data sets straightforward. Tasks like comparing the impact of various authors, topics, or trends on a search-term become easier.

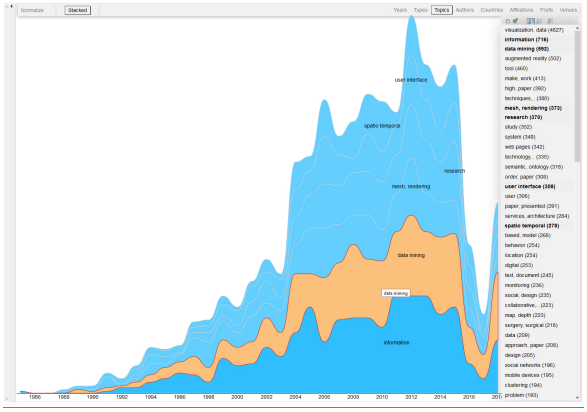

Fig. 10. Temporal visual layout: stacked-chart.

For analyzing trends, it is important to gather the knowledge of the underlying topics, technologies etc. emerged during the time or lost its relevance. Figure 8 illustrates two visual layouts that make use of the different data models. To enable a fast and comprehensible analysis, we integrated a temporal ranking (see Figure 8-a). This visual layout offers beside the introduced configuration areas the ability to specify the amount of rows to be visualized. The visual layout is divided horizontally into columns for each year of the analyzed time span. The arrangement is based on the amount of publications having a facet item as a property of the selected facet type, sorted in descending order from top to bottom. The order only represents the ranking, additional more concrete information about the relative amount is represented by the width of each rectangle. With these position and form indicators, the user can quickly determine facet items with high in influences in each year. In Figure 8-a the temporal data model, the topic model and the semantic model are merged to get the visualized information. Thereby, top related topics of documents from the result set of the query ”data mining” are visualized and by selecting one topic, the temporal ranking for each year is highlighted. In Figure 8-b a semantic visual layout offers the ability to understand relations between authors (co-author information) and between topics (which topic relations) and the semantic correlation between the information entries. Commonly semantic relations are visualized with node-link graphs, which may lead to complex visualizations and reduce the analysis capability. We integrated beside such node-link visualizations a circle-layout that arranges the entities as a spiral starting from the center of the screen. Figure 8-b illustrates the semantic relations based on the facet type author.

The size of each element indicates the amount of publication per author, whereas the degree option indicates the amount of distinct relation targets within the facet type.The Semantic Visual Layout is used to provide detailed relational information about individual facet items, which can be accessed through user interaction. After selecting a circle all relational information within the same facet type are highlighted. This leads to a real-time loading of all co-authors in the result set.

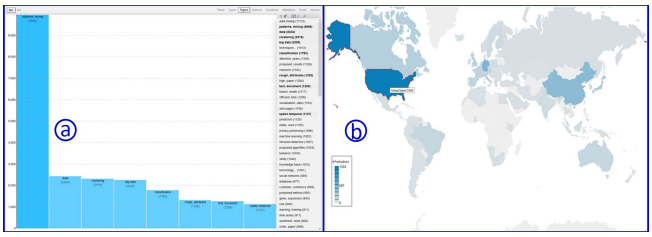

Fig. 11. (a): topic visual layout and (b): geographical visual layout.

Further, users are able to get an insight about correlations within the semantic relations through mouse-over that refines the in the figured cased the co-authorship of certain authors through the color. Figure 11-a illustrates a topic visual layout that visualizes topics related to search term based on the frequency of their appearance, and Figure 11-b illustrates a geographical visual layout that encodes the amount of publications per authors' country through saturation.

D. Visual Orchestration

The described visual layout are just a set of possible interactive visualization that should enable solving analytical tasks with regards to technological trends and related questions as described in Section III. With our proposed approach varieties of visual interfaces can be integrated for various tasks. Figure 12 illustrates some more examples designed for different tasks.

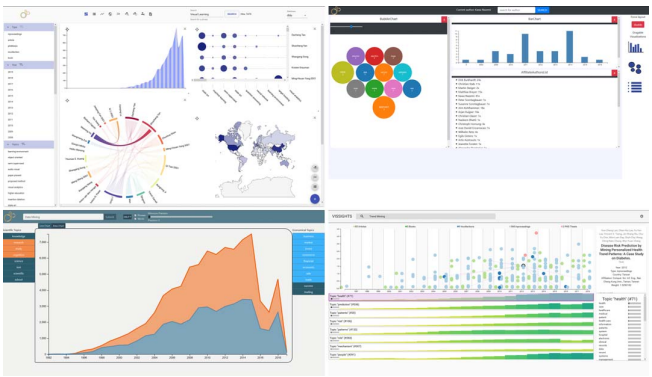

Fig. 12. Example of an alternative visual interface.

X. Conclusions

We introduced in this paper an approach for integrating, enriching, mining, analyzing, identifying and visualizing emerging trends from scientific text collections. We first introduced the state of the art in trend mining and trend and text visualization. Our contribution was three-fold: (1) a model for gathering trends from text to visual interactive analysis representations, (2) the identification of upcoming or emerging trends based on text, and (3) an approach for visual interaction through different data models and related interactive visual representations to explore the potentials of technologies and detect new insights. With our contributions

the process of trend analysis is supported in a more efficient and effective way and leads to finding undetected patterns in data. For the fist contribution, we introduced a model that enables the visual analysis of text documents through different transformation steps. Each step was introduced in detail. For our second contribution, we introduced a novel method to measure emerging and decreasing trends. The calculation was described in a replicable way. For our third contribution, we introduced a number of visual interaction techniques that allows a comprehensive exploration and detection even of undetected trends in text collections. Overall, an approach was provided that enables the main goal of detecting emerging or decreasing trends through humans' interaction with visual interactive systems.

ACKNOWLEDGMENT

This work was partially funded by the Hessen State Ministry for Higher Education, Research and the Arts within the program "Forschung für die Praxis" and was conducted within the research group on Human-Computer Interaction and Visual Analytics (https://vis.h-da.de). The presentation of this work was supported by the Research Center for Digital Communication & Media Innovation of the Darmstadt University of Applied Sciences.

REFERENCES

[1] D. Keim and F. Kohlhammer, Jörn; Ellis Geoffrey; Mansmann, Eds., *Mastering the information age : solving problems with visual analytics*. Goslar : Eurographics Association, 2010.

[2] K. Nazemi, R. Retz, D. Burkhardt, A. Kuijper, J. Kohlhammer, and D. W. Fellner, "Visual trend analysis with digital libraries," in *Proceedings of the 15th International Conference on Knowledge Technologies and Data-driven Business - i-KNOW '15*. ACM Press, 2015.

[3] R. Feldman and I. Dagan, "Knowledge Discovery in Textual Databases (KDT)," in *Proceedings of the First International Conference on Knowledge Discovery and Data Mining*, 1995.

[4] B. Lent, R. Agrawal, and R. Srikant, "Discovering trends in text databases," in *Proceedings of KDD '97*, 1997.

[5] R. Agrawal and R. Srikant, "Mining sequential patterns," in *Proceedings of the Eleventh International Conference on Data Engineering*, 1997.

[6] R. Agrawal, G. Psaila, E. Wimmers, and M. Zait, "Querying shapes of histories," in *Proceedings of the 21st International Conference on Very Large Databases*, 1995.

[7] R. Feldman, Y. Aumann, A. Zilberstein, and Y. Ben-Yehuda, "Trend graphs: Visualizing the evolution of concept relationships in large document collections," in *Principles of Data Mining and Knowledge Discovery*, J. M. Żytkow and M. Quafafou, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 1998, pp. 38–46.

[8] M. M. y Gomez, A. Gelbukh, and A. Lopez-Lopez, "Mining the news: Trends, associations, and deviations," *COMPUTACION Y SISTEMAS*, vol. 5, no. 1, pp. 14–24, 2001.

[9] N. S. Glance, M. Hurst, and T. Tomokiyo, "Blogpulse: Automated trend discovery for weblogs," in *In WWW 2004 WS on Weblogging*. ACM, 2004.

[10] Q. Mei and C. Zhai, "Discovering evolutionary theme patterns from text: An exploration of temporal text mining," in *Proceedings of ACM SIGKDD*, 2005.

[11] C. Zhai, A. Velivelli, and B. Yu, "A cross-collection mixture model for comparative text mining," in *Proceedings of the Tenth ACM SIGKDD*, ser. KDD '04. NY, NY, USA: ACM, 2004, pp. 743–748. [Online]. Available: http://doi.acm.org/10.1145/1014052.1014150

[12] M. Viermetz, M. Skubacz, C.-N. Ziegler, and D. Seipel, "Tracking topic evolution in news environments," in *10th IEEE Conference on E-Commerce Technology*, 2008, pp. 215–220.

[13] Y. Kim et al., "Automatic discovery of technology trends from patent text," in *Proceedings of the 2009 ACM Symposium on Applied Computing (SAC).*, S. Y. Shin and S. Ossowski, Eds. ACM, 2009. [Online]. Available: http://doi.acm.org/10.1145/1529282.1529611

[14] S. Goorha and L. Ungar, "Discovery of significant emerging trends," in *Proceedings of the 16th ACM SIGKDD*, 2010. [Online]. Available: http://doi.acm.org/10.1145/1835804.1835815

[15] S. Liu et al., "Tiara: Interactive, topic-based visual text summarization and analysis," *ACM Trans. Intell. Syst. Technol.*, vol. 3, no. 2, pp. 25:1–25:28, Feb. 2012. [Online]. Available: http://doi.acm.org/10.1145/2089094.2089101

[16] K. Nguyen and and, "Hot topic detection and technology trend tracking for patents utilizing term frequency and proportional document frequency and semantic information," in *2016 International Conference on Big Data and Smart Computing (BigComp)*, Jan 2016, pp. 223–230.

[17] K. K. Bun and M. Ishizuka, "Topic extraction from news archive using tf*pdf algorithm," in *Proceedings of the Third International Conference on Web Information Systems Engineering, 2002. WISE 2002.*, Dec 2002, pp. 73–82.

[18] K. Chen, L. Luesukprasert, and S. T. Chou, "Hot topic extraction based on timeline analysis and multidimensional sentence modeling," *IEEE Transactions on Knowledge and Data Engineering*, vol. 19, no. 8, pp. 1016–1025, Aug 2007.

[19] J. L. Hurtado, A. Agarwal, and X. Zhu, "Topic discovery and future trend forecasting for texts," *Journal of Big Data*, vol. 3, no. 1, p. 7, Apr 2016. [Online]. Available: https://doi.org/10.1186/s40537-016-0039-2

[20] S. Havre et al., "Themeriver: Visualizing thematic changes in large document collections," *IEEE TVCG*, vol. 8, no. 1, pp. 9–20, 2002. [Online]. Available: http://dx.doi.org/10.1109/2945.981848

[21] W. Dou, X. Wang, R. Chang, and W. Ribarsky, "Paralleltopics: A probabilistic approach to exploring document collections," in *VAST 2011*, 2011.

[22] C. Collins, F. Viegas, and M. Wattenberg, "Parallel tag clouds to explore and analyze faceted text corpora," in *VAST 2009*, 2009.

[23] B. Lee, N. H. Riche, A. K. Karlson, and S. Carpendale, "Sparkclouds: Visualizing trends in tag clouds," *IEEE TVCG*, vol. 16, 2010. [Online]. Available: http://dx.doi.org/10.1109/TVCG.2010.194

[24] S. Lohmann, M. Burch, H. Schmauder, and D. Weiskopf, "Visual analysis of microblog content using time-varying co-occurrence highlighting in tag clouds," in *Proceedings of the International Working Conference on Advanced Visual Interfaces*, ser. AVI '12. New York, NY, USA: ACM, 2012, pp. 753–756. [Online]. Available: http://doi.acm.org/10.1145/2254556.2254701

[25] Q. Han, F. Heimerl, J. Codina-Filba, S. Lohmann, L. Wanner, and T. Ertl, "Visual patent trend analysis for informed decision making in technology management," *World Patent Information*, vol. 49, pp. 34 – 42, 2017. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0172219017300455

[26] F. Heimerl, Q. Han, S. Koch, and T. Ertl, "Citerivers: Visual analytics of citation patterns," *IEEE Transactions on Visualization and Computer Graphics*, vol. 22, no. 1, pp. 190–199, Jan 2016.

[27] S. K. Card, J. D. Mackinlay, and B. Shneiderman, *Readings in Information Visualization: Using Vision to Think.*, 1st ed. Morgan Kaufmann, 1999.

[28] Y. Noh, K. Hagedorn, and D. Newman, "Are Learned Topics More Useful Than Subject Headings," in *Proceedings of the 11th ACM/IEEE JCDL*, 2011.

[29] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *Journal of Machine Learning Research*, vol. 3, 2003. [Online]. Available: http://www.jmlr.org/papers/v3/blei03a.html

[30] K. Nazemi, D. Burkhardt, R. Retz, A. Kuijper, and J. Kohlhammer, "Adaptive Visualization of Linked-Data," in *Advances in Visual Computing*. Springer, 2014, pp. 872–883.

[31] K. Nazemi, *Adaptive Semantics Visualization*, ser. Studies in Computational Intelligence 646. Springer International Publishing, Studies in Computational Intelligence 646, 2016. [Online]. Available: http://www.springer.com/de/book/9783319308159

[32] J. Bertin, *Semiology of graphics*. University of Wisconsin Press, 1983.

[33] B. Shneiderman, "The eyes have it: A task by data type taxonomy for information visualizations," in *VL*, 1996, pp. 336–343.

[34] F. van Ham and A. Perer, "Search, show context, expand on demand: Supporting large graph exploration with degree-of-interest," *IEEE Trans. Vis. Comput. Graph.*, vol. 15, pp. 953–690, 2009.