Visual Trend Analysis with Digital Libraries

Kawa Nazemi[1], Reimond Retz[1], Dirk Burkhardt[1], Arjan Kuijper[1,2], Jörn Kohlhammer[1,2], and Dieter W. Fellner[1,2]
[1]Fraunhofer IGD
[2]Technische Universität Darmstadt
Fraunhoferstr. 5, D-64283 Darmstadt, Germany
{kawa.nazemi, rretz, dburkhar, arjan.kuijper, joern.kohlhammer, fellner}@igd.fhg.de

ABSTRACT

The early awareness of new technologies and upcoming trends is essential for making strategic decisions in enterprises and research. Trends may signal that technologies or related topics might be of great interest in the future or obsolete for future directions. The identification of such trends premises analytical skills that can be supported through trend mining and visual analytics. Thus the earliest trends or signals commonly appear in science, the investigation of digital libraries in this context is inevitable. However, digital libraries do not provide sufficient information for analyzing trends. It is necessary to integrate data, extract information from the integrated data and provide effective interactive visual analysis tools. We introduce in this paper a model that investigates all stages from data integration to interactive visualization for identifying trends and analyzing the market situation through our visual trend analysis environment. Our approach improves the visual analysis of trends by investigating the entire transformation steps from raw and structured data to visual representations.

CCS Concepts

•**Human-centered computing** → **Visual analytics; Information visualization; Visualization theory, concepts and paradigms;** Visualization toolkits; •**Information systems** → *Data analytics;* Digital libraries and archives;

Keywords

Visual Analytics, Information Visualization, Datamining, Trend Analysis, Information Extraction

1. INTRODUCTION

The early awareness of upcoming trends in technology enables a more goal-directed and efficient way for deciding future strategic directions in enterprises and research. Possible sources for this valuable information are ubiquitously and freely available in the Web, e.g. news services, companies reports, or social media platforms and blog infrastructures. To support users in handling these information sources and to keep track of the newest developments, current information systems make intensively use of information retrieval methods that reduce the amounts of documents according to a given query. The commonly used search mechanisms are primarily focused on providing the users with easy access to information of their interest and deal with the access to information items and resources [2], but neither provide an overview of the content nor enable the exploration of emerging or disappearing trends for inferring future trends. 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. The interactive overview on data, the continuous changes in data, and the ability to explore data and gain insights are sufficiently supported by visual analytics and information visualization approaches, whereas the appliance of such approach in combination with trend analysis are merely propagated. Another important aspect of trend analysis is the pool of data. If a technology or method is discussed in social media or released in company reports, it can be assumed that these technologies already reached their climax and are popular enough for strategic decisions. However, those early trends of technologies are commonly propagated first in research and are included in scientific digital libraries. Therefore, digital libraries are the "real" information pool for early signals and trends. Although, the characteristics of digital libraries and their value for trend analysis is obvious, a real analysis of trends based on digital library content through visual analytics methods could not be found.

We introduce in this paper an approach for integrating, mining, analyzing and visualizing data from digital libraries with the main goal of providing an efficient visual trend analysis solution. First the overall model will be introduced that investigates several transformation stages from unstructured and structured data to interactive visual representation. Thereafter, the process of exploratory visual trend analysis based on different visual layouts will be described. Our contribution is three-fold: (1) a model for gathering trends from heterogeneous digital library data to visual interactive analysis representations, (2) the combination of visual layouts with data mining approaches for analyzing trends, and a model for assisted search that enables to explore an unknown domain more efficiently. 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.

i-KNOW '15, October 21-23, 2015, Graz, Austria

DOI: http://dx.doi.org/10.1145/2809563.2809569

2. RELATED WORK

Our approach consists of three main parts, namely trend extraction from text, visualizing trends, and using digital libraries as initial data corpus for the trend extraction and visualization. We therefore introduce the related work in these three areas.

2.1 Trend Mining

The origin of trend mining from text can be traced back up to the beginning of the first text mining approaches also known as Knowledge Discovery from Text (KDT) [11]. Different approaches arose that aimed at identifying key topics and discovering their relevance over time. Lent et al. [19] defined a trend as a sequence of frequencies of a specific phrase. Feldman et al. introduced *Trend Graphs* [25] 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 given a specific context. 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 [29]. The starting point for the trend discovery is a normalized topic vector that is extracted from the documents of each time interval. Mei and Zhai introduced two methods for discovering evolutionary theme patterns in text [21]. Their methods are based on a set of salient themes that are extracted from temporal sets of documents using a probabilistic mixture model. Based on these themes per time interval they differentiate between *Theme Evolution Graphs* and *Theme Live Cycles*. 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 [28]. Kim et al. presented an approach for discovering technology trends from patent texts [30]. 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. However their approach is primarily applicable in the specific domain of patent texts. The generation of training examples for learning the classifier is not possible for new and unknown scenarios. Goorha and Ungar presented an approach that discovers emerging trends in text collections by identifying significant phrases associated with user defined known entities [14]. The system extracts the known key-phrases and ranks the interest. The results are visualized in a *Scatterplot* that represents the significance of emerged trends over time. *Tiara* is an approach that allows the identification of topic-related trends. It utilizes the Latent Dirichlet Allocation (LDA) for generating topic models [27]. *Tiara* calculates the strength of the topics over time that is visualized in a stacked graph for identifying peaks and slopes of each topic.

2.2 Trend Visualization

Current trend mining methods provide useful indications for discovering emerging trends. Nevertheless, the interpretation and conclusion for serious decision making still requires human decision making and 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. A more sophisticated approach for representing trends is *ThemeRiver* [26]. It represents thematic variations over time in a stacked graph visualization with a temporal horizontal axis. The visual analytics system *ParallelTopics* [10] 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.

Word Clouds provide an overview of documents' content by visualizing terms that summarize the key issues by using the size as the visual variable of a given document collection. The temporal information are not supported and the visualization is limited to a single instant of time. *Parallel Tag Clouds* (PTC) [9] enhance this with multiple word clouds that represent the contents of different facets in the document collection. Temporal facets can be used to identify the difference of certain keywords over time. Another extension of word clouds are *SparkClouds* [18] that includes a sparkline for each term or trend. These sparklines indicate the temporal distribution of each term.

Beside topic-, animation- and word-based trend visualizations there are also other approaches from the area of text visualization that offer potential for designing novel trend visualizations. For instance map-based approaches [12, 24] provide an overview of large document collections by visualizing clusters of documents. For analyzing trends, these approaches can be used to select specific areas of interest for further analysis. Graph-based methods are well suited for visualizing associations that are latent in the underlying text collections and have been successfully used to visualize co-occurrences [15] and to compare semantic profiles [32].

2.3 Digital Library Visualization

The visualization of bibliographic entries is a broad area with many efficient and useful techniques, especially in the field of Digital Libraries. The *BiblioViz* system by Shen et al. [31], the *Citation Map* of *Web of Knowledge* [20] and a methodology based on *Power Graphs* [13] may serve as examples.

BiblioViz does not use a specific existing data collection [31]. It requires detailed information as input, e.g. lists of papers, venues, authors etc. Neither data normalization nor data mining techniques are used and the visualizations are limited to the overview in form of table and networks. Chou and Yang combined in *PaperVis* a modified version of Radial Space Filling and Bullseye View to arrange papers as a node-link graph with the static set of the InfoVis 2004 Contest Dataset [8]. *PaperVis* uses the keywords of the data collection to provide a semantically meaningful hierarchy for facilitating literature exploration. Lee et al. introduced *PaperLens* a visualization for the same contest data [1]. No efforts were made to enrich the data quality and get further insight through data mining approaches. Bergström and Atkinson [3] introduced with *PaperCube* a web-based approach for visualizing the metadata of *CiteSeer*. It does not use any kind of topic or trend extractions or classification methods.

Chen introduced with *CiteSpace* a system to detect and visualize emerging trends and transient patterns with a cluster and a time-zone view [7]. The system is limited by the lack of direct access to the data repository. For each trend analyzing, the users have to perform a new search on the *Web of Science*, download the result set and import this data set into the *CiteSpace* application. van Eck and Waltman intro-

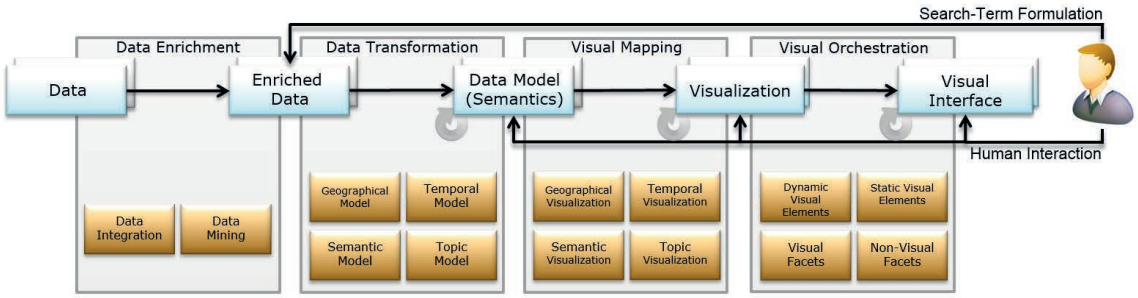

Figure 1: Our Conceptual Model includes the five transformation steps: *Data* is enriched, structured, mapped to *Visualizations*, and orchestrated into a *Visual Interface*.

duced *CitNetExplorer* for analyzing and visualizing citation networks [16]. Similar to *CiteSpace*, the users have to search, download, and import the data for visualization. We introduced with *SemaVis* a system that uses BibTeX entries from the Eurographics Association (EG) Digital Library [17]. No effort is being made to use data mining techniques.

Based on our literature review we can outline that there exist no system or approach that uses the existing information in digital libraries to provide a sufficient solution for mining and visualizing trends. Trend mining approaches use predefined key phrases that require prior knowledge about the domain and are not sufficient for identifying upcoming and unknown trends. Those methods that are able to extract information with probabilistic methods, e.g. LDA [4] are not applied to visualizations in digital libraries. Existing trend visualizations do not make use of digital libraries as data corpus and do not provide comprehensible and interactive views on unknown trends. Digital library visualizations focus commonly on citation networks and correlations of meta data. Even if they propagate to visualize trends, data mining approaches are either not used at all or the process of analyzing is difficult and limited to a downloaded subset of data.

3. VISUAL TRANSFORMATION MODEL

Our model builds upon on our previous work on adaptive visualization [22] that used the reference model of Card et al. [6] as foundation. According to our previous model, we subdivide the transformation process for visual trend analysis in digital libraries into the steps of (1) *Data Enrichment*, (2) *Data Transformation*, (3) *Visual Mappings*, and (4) *Visual Orchestration* as illustrated in Figure 1. *Data Enrichment* gathers additional data from external repositories to enhance the quality of data and uses text analysis techniques to extract valuable information from these data. *Data Transformation* structures the data for a proper visualization. It detects relevance, amount and content of queried data and uses these features to create models revealing certain aspects of the data. *Visual Mappings* transforms the data models to appropriate visualizations. *Visual Orchestration* uses textual and visual information gathered in the previous transformation steps to create static and dynamic elements for human interaction.

3.1 Data Enrichment

We could outline in our literature review that the data quality in digital libraries is very heterogeneous and not always coherent. For a proper analysis of the given data, enhancements of data quality are necessary. To enhance the quality of data, we first use *Data Integration* techniques to gather additional data from Web. Thereafter, *Data Mining* (Text Mining) is used on the data to generate valuable information from the data collection and therefore create enriched data.

The data collection used by our model 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 three million entries. All entries in this data set are without text, e.g. abstracts, which are necessary to enable an analysis of trends by information extraction methods. We therefore balance out the limitation of the basis data collection 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 Web or which online digital library has more information about a certain publication. We integrate data from *Springer*, *IEEE*, and *Computer Org.* The basis data collection contains a link to the publishers' 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. With this first step we enrich the data of DBLP with further meta data and abstracts directly from the publishers (e.g. Springer or IEEE).

During the first step, we gather at least abstracts for a major part of the DBLP entries and 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 [23]. For this purpose, we have integrated in the step of *Data Mining* the Latent Dirichlet Allocation (LDA) [4] 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

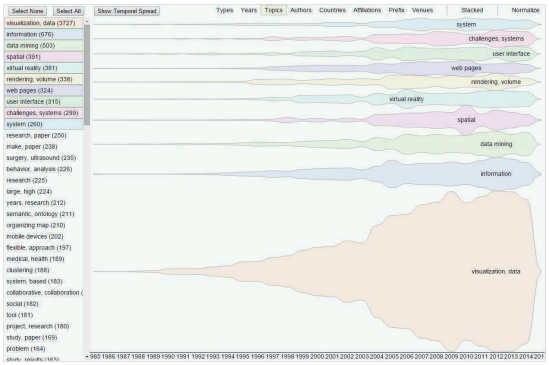

Figure 2: Temporal River comparison of topics in publications for the search term "Information Visualization"

the accuracy of the resulting model. 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*. Together with the *Temporal Model*, emerging trends can be visualized as illustrated in Figure 2.

3.2 Data Transformation

In our first transformation step, we enhanced the information quality and extracted topics from the abstracts of publications. In this second transformation step, the *Enriched Data* are transformed into aspect-oriented *Data Models*. These allows to visualize different aspects of the underlying data to enable a proper analysis of emerging trends. Our assumption for identifying and analyzing trends refer mainly to the 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 evolve, and *which* technologies or topics are relevant for an enterprise?

The aspect-oriented data models focus on the above questions with the particular aspects that are given in the data. Therefore, we generate four data models, *Semantics Model*, *Temporal Model*, *Geographical Model*, and *Topic Model*. The basis for the creation of the models are the *Enriched Data* from the previous step. To create our data models, it is necessary to limit the data to the queried search, as the process of analysis starts always with a formulated search. For this purpose, we generate an index over the entire data. The data models contain entries for the search-terms. The gathered abstracts, as well as titles and authors, are indexed to help answering the analysis questions.

The generation of a semantic model serves as the primary data model for storing all information. It adds structure and semantics to the data for an easier extraction of information. This model is used in the textual list presentation, where all available information about each publication are presented, and the generation of facet information for filtering purposes (see Figure 3). To accomplish this, a Data Table is generated including all publications with their attributes and relations. The information required for facets is also included as part of the results. This includes the *type of publication*, *publication year*, extracted *topics*, *authors*, *countries*, *affiliations*, and *venues*. This primary data model is the foundation for all other data models.

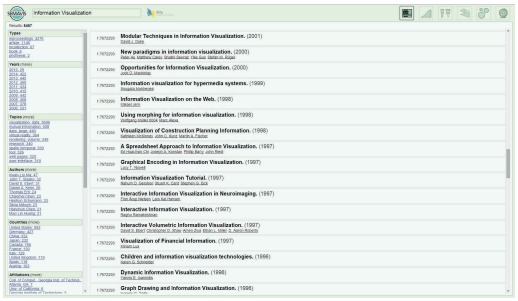

Figure 3: List-based representation of results with facets for refining the results.

The *temporal model* is used by multiple temporal visual layouts. Here, multiple aspects of information in the data collection need to be accessible based on the time property. The temporal analysis is important to analyze the trends and emerging technologies and all related *when* questions. Based on faceted attributes, detailed temporal spreads are part of the temporal model for all attributes of each facet. The complexity of the geographical model is lower and refers to the *where* questions. The geographical layout only needs quantity information for each country. We extract for this purpose the authors' affiliation in relation to the country or use the meta tag "country". The topic model is built by integrated machine learning algorithm for text mining. Additionally, the topic model contains top 20 most used phrases as N-Grams with their usage probability. The inclusion of most used phrases can help the user to reformulate the search query and find additional information.

3.3 Visual Mapping for Analysis

In the first steps of our model, we enhanced the data quality by gathering additional information. The enhanced information were analyzed and topics were generated. These topics and different aspects of data were modeled into data models that enable a more efficient visual analysis. The transformation process of *Visual Mappings* aims at creating visual layouts by using appropriate positioning and layout algorithms. According to the introduced questions and related data models, we identify *Semantic Visual Layouts*, *Temporal Visual Layouts*, *Geographical Visual Layout*, and *Topic Visual Layout* to support the process of visual trend analysis.

Our approach starts with a bottom-up search term formulation [22], whereas the results are presented first as temporal overview to illustrate the temporal spread of the searched key-term and indicate upcoming trends (see Figure 4). It consists of a line chart showing the amount of publications for each year. This aids in understanding the evolution of the popularity of the search-term. It can also help in making predictions about rise or or digression of topics and technologies.

Although the users are able to get the searched results as data entities with our list-based view, the main intention is to enable first an overview of the entire searched results and enable to view the overall temporal spread of the searched term. Further, the various visual layouts enable the refinement, reduction, and inspection of the result set for analyzing the relevance of certain terms, technologies, or topics. These visual layouts offer great insights into the entire re-

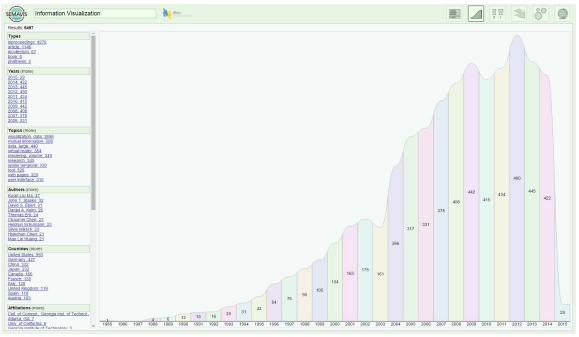

Figure 4: Temporal overview of the search "Information Visualization" as starting visual layout to indicate the overall trend spread.

sult set and help to understand the area of interest (AOI) or knowledge domain. The acquired knowledge enables reformulating or refining the search query. In our model, we provide multiple layouts that enlighten several specific aspects of the result set. Temporal visual layouts (see Figures 2 and 4) help to understand the chronological properties of technologies and approaches ranging from purely quantity information to semantic properties or extracted topics of publications over time. Our *temporal stacked chart* consists of two configuration areas and a view area for the visual layout (see Figure 5). 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 left) allows to select the number of visualized entities. It lists all available items for the chosen facet type.

Figure 5 shows another *temporal visual layout*, *Temporal Stacked Comparison*. Displaying the temporal information for multiple facet items at the same time empowers the user to analyze the result set in more detail by comparing relevant faceted information.

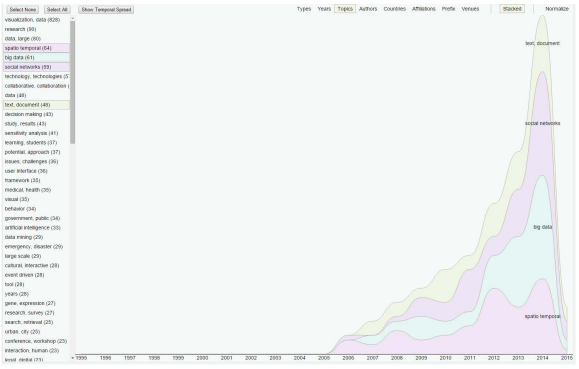

Figure 5: Temporal stacked view on the search results for the term "visual analytics"

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 (see Figure 5). 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 2 illustrates this approach. 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.

For analyzing the trends it is important to gather the knowledge which of the underlying topics, technologies etc. emerged during the time and which one lost relevance. To enable a fast and comprehensible analysis view on this issue, we further integrated a *temporal ranking* (see Figure 6). 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 the 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 (e.g. topics, authors or venues) with high influences in each year.

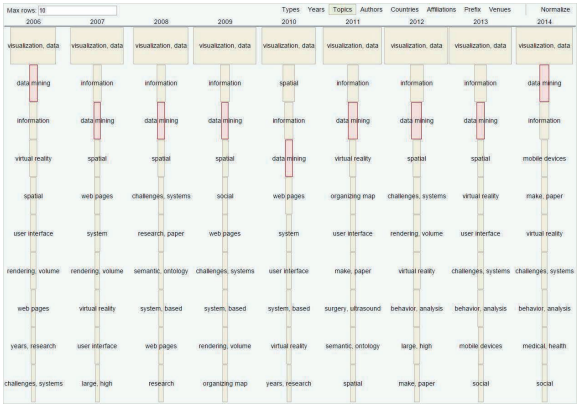

Figure 6: Temporal ranking layout of the search term "information visualization". The selected facet type is *Topics* and the selected topic is "data mining".

The *Geographical Visual Layout* provides topological information about the origin of the published articles. It consists of the world map flattened with the Mercator projection [5], showing only the outlines of countries (see Figure 7:a). The color saturation of country areas acts as an information vessel for the view. This is used to display the amount of publications written by authors who reside within each country. We use a linear saturation scale, ranging from low saturation for countries with few publications to a high saturation for countries with a lot of publications. Using user interaction techniques, it is possible to display the name and amount of publication of individual countries. Quantitative geographical information about residents of authors in general can help in understanding which topic is of great interest in which country.

The *Topic Visual Layout* (Figure 7:b) offers the foundational information of the topic model. The Latent Dirichlet Allocation [4] creates a statistical topic model, where each topic is characterized by a set of word probabilities. In turn, each document is modeled as a set of topic probabilities. For trend analysis the information about the topics assigned to publications are of interest, thus they might illustrate which topic is emerging in the area of visual analytics or in context of intelligent environments.

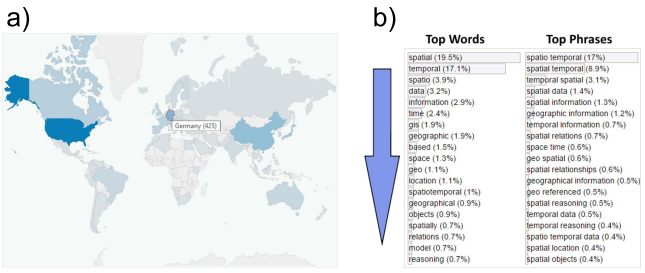

Figure 7: a: *geographical visual layout* encodes the amount of publications per authors' country through saturation and b: topic visualization giving insights into the topic model.

The *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 illustrates the semantic relations based on the facet type author. Thereby the spiral arrangement is displayed with lines. 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.

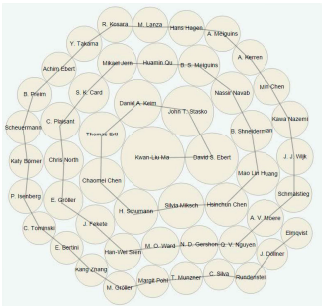

Figure 8: *Semantic Visual Layout* with the spiral arrangement.

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. Further, users are able to get an insight about correlations within the semantic relations through mouse-over (see Figure 9). This relational visualization can be enhanced with further interactive views.

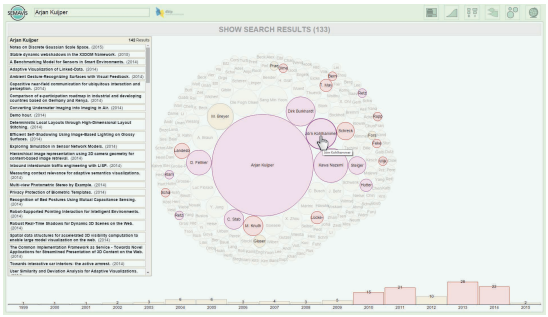

Figure 9: *Semantic Visual Layout* with enhanced relation views through different colors.

The introduced visual layouts make use of visual variables to enhance cognitive perception of illustrated information. The usage of visual variables in meaningful ways improves the speed in which analysis tasks can be accomplished. We already illustrated that size, color and arrangement of visual entities represent different meanings, e.g. different colors in the *Semantic Visual Layout* for different kind of relations and size or color for quantity measure. This leads to a faster identification and consequently retrieval of important information.

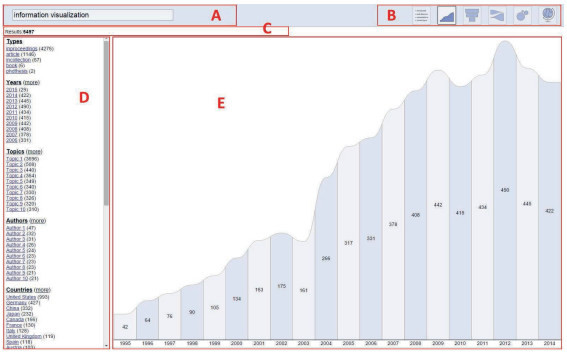

Figure 10: The general visual interface with the different regions.

3.4 Visual Orchestration

The last step, visual orchestration, is responsible for the placement of different visual layouts on screen and the changes of layouts to support the process of trend analysis. Figure 10 illustrates the general design of the visual interface with the following areas: *Region A* consists of the input field for the search query. *Region B* contains a visual configuration switch, which determines the visual layout to be shown in *Region E*, the main content area in the center. *Region C* informs the user about the total amount of hits found for the specified search query. *Region D* lists all facets in textual form.

The region *D* in Figure 10 lists the extracted facets from the entire search result in a textual form. It is divided into groups of listings for each facet type. Each group shows the name of the facet type as a header for easier navigation, followed by a list of facet items. These items are typically

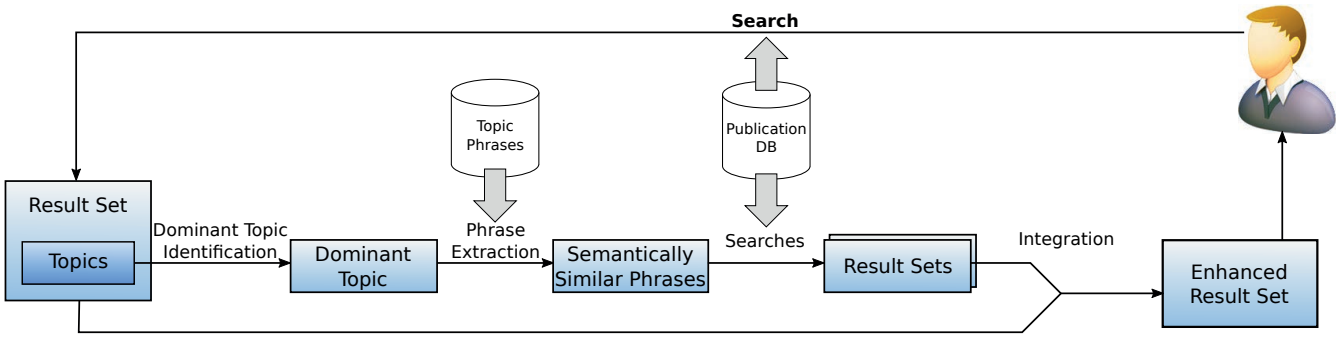

Figure 11: Assisted search Approach: Semantically similar phrases are extracted based on the most dominant topic for enhancing the search.

ordered by amount of publications. Except in the facet type *Year*, where a chronological order makes more sense. Initially, the amount of facet items in the listings is limited to a predefined amount. If the amount shown is not enough, the user is always able to expand the list to include all facet items with a quick interaction of clicking on the title header. This truncation in combination with the ordering by amount gives the user a quick overview over quantity information in regard to each individual facet type category. Facet items of each facet type reveal information about specific aspects of the entire result set and consequently the search term knowledge domain. ***Types***: What kind of publication type is dominant within the search result? ***Years***: This category offers a great insight into the temporal history of the search query domain. ***Topics*** Information about prominent topics from publications of the entire search result set helps immensely when trying to get a quick content overview. ***Authors***: Getting a list of the prominent authors within the search result offers a great way to find people with vast knowledge in the search term domain. ***Countries***: Geographical information about the residence of the authors and the associated countries. ***Affiliations***: Information about prominent institutes, which invest resources into the research of specific topics is offered in this category. ***Venues***: The listing of the most prominent venues (journals and conferences) can also act like the topic facet type.

The displayed facet information in textual form does not only offer insight into the data collection limited by the search terms and consequently is not only a vessel for comprehending the search result. During the analysis process, the user examines the result set using both the textual form of the facet listing and the presented visual layouts. It would be a nuisance for facet filtering, if the user needed to find facet items in the textual view after identifying them during the exploration in one of the visual layouts. So in order to enable fluid and intuitive interaction with the model, the visual layouts also offer a possibility to apply facet filtering, like it is already possible in the textual facet view. Each visualization and the interaction with it leads to reducing the result set.

4. ASSISTED SEARCH

The analysis of trends requires deep knowledge in a particular domain. In case of visual trend analysis a way of extending the result set is required to provide even to non-expert a way of gathering the technological trends in a domain and enable the foresight of technologies. We therefore extended the search functionalities of our approach beside traditional linguistic methods, e.g. usage of thesaurus for British and American English or retrieving plural forms of search terms, with a new topic-based approach. Our model incorporates the information within the *Topic Model* to enhance search-terms from the query. The *Topic Model* provides us with N-Grams (with $n \in N > 1$), which are the most used phrases within each topic. These phrases often represent different ways to articulate the idea of the topic and consequently can be used as a key phrase to represent the topic. We use data from the *Topic Model* and choose the top five most used phrases as additional search-terms to extend the result set. Figure 11 illustrates our assisted search approach. Based on the initial users' search the most dominant topic in the result set is identified. In the next step, semantically similar phrases are extracted based on the identified dominant topic. Additional searches are performed by the system using the semantically similar phrases as search-terms.

For example the user tries to find information about "intelligent environment". Our model recognizes the topic of the provided search term within our *Topic Model* by choosing the most assigned topic from the publications within the initial result set and enhances the search with the terms "smart home", "ubiquitous computing", "pervasive computing", and "Internet of Things". This leads to fewer iterations in search, more results in analysis and enables the process of learning. Additionally, the content of the topic model is available for inspection.

5. CONCLUSIONS

We introduced in this paper an approach for integrating, mining, analyzing and visualizing data from digital libraries with the main goal of providing an efficient visual trend analysis solution. Our contribution in this paper was three-fold: (1) a model for gathering trends from heterogeneous digital library data to visual interactive analysis representations, (2) the combination of visual layouts with data mining approaches for analyzing trends, and (3) an assisted search approach by using data mining methods. For the fist contribution, we introduced a model that enables the visual analysis

of structured and unstructured data through five transformation steps. Each step was introduced. For our second contribution, we introduced a number of visual layouts that make use of probabilistic information extraction methods and enables a sufficient visualization of the gathered information. These extracted information are the baseline for our third contribution that enables a search by making use of the extracted topics as complimentary search terms for enhancing the result set with semantically similar terms. Our approach is implemented with ground truth data (see video of system: (http://www.media.semavis.net/dblp)).

6. ACKNOWLEDGMENTS

This work was partially funded by the European Commission under the grant agreement no. 287119 of the 7th Framework Program. This work is part of the SemaVis technology, developed by Fraunhofer IGD (http://www.semavis.net).

7. REFERENCES

[1] B. Lee et al. Understanding research trends in conferences using paperLens. In *Extended Abstracts Proceedings of CHI 2005*, 2005.

[2] R. Baeza-Yates and B. Ribeiro-Neto. *Modern Information Retrieval*. Addison-Wesley Publishing Company, 2nd edition, 2010.

[3] P. Bergström and D. C. Atkinson. Augmenting the exploration of digital libraries with web-based visualizations. In *IEEE ICDIM*, 2009.

[4] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3, 2003.

[5] R. Buchholz and W. Krücken. *Die Mercator-Projektion: zu Ehren von Gerhard Mercator (1512 - 1594)*. Becker, 1994.

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

[7] C. Chen. CiteSpace II: Detecting and Visualizing Emerging Trends and Transient Patterns in Scientific Literature. *Journal of the American Society for Information Science and Technology*, 2006.

[8] J. K. Chou and C. K. Yang. PaperVis: Literature Review Made Easy. In *Proceedings of the 13th Eurographics / IEEE - VGTC Conference on Visualization*, 2011.

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

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

[11] 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.

[12] B. Fortuna, M. Grobelnik, and D. Mladenic. Visualization of text document corpus. *Informatica*, pages 497–502, 2005.

[13] G. Tsatsaronis et al. How to become a group leader? or modeling author types based on graph mining. In *Proceedings of TPDL*, 2011.

[14] S. Goorha and L. Ungar. Discovery of significant emerging trends. In *Proceedings of the 16th ACM SIGKDD*, 2010.

[15] M. Grobelnik and D. Mladenic. Visualization of news articles. In *SIKDD 2004 at multiconference IS 2004*, 2004.

[16] N. Jan van Eck and L. Waltman. CitNetExplorer: A new software tool for analyzing and visualizing citation networks. *Journal of Informetrics*, 8:802–823, 2014.

[17] K. Nazemi et al. Adaptive Semantic Visualization for Bibliographic Entries. In *Advances in Visual Computing*. Springer, 2013.

[18] B. Lee, N. H. Riche, A. K. Karlson, and S. Carpendale. Sparkclouds: Visualizing trends in tag clouds. *IEEE TVCG*, 16, 2010.

[19] B. Lent, R. Agrawal, and R. Srikant. Discovering trends in text databases. In *Proceedings of KDD Š97*, 1997.

[20] T. Matthews. *Citation Map Visualizing Citation Data in the Web of Science*. Thomson Reuters, 2010.

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

[22] K. Nazemi. *Adaptive Semantics Visualization*. PhD thesis, TU Darmstadt, Eurographics Association, 2014.

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

[24] F. Paulovich and R. Minghim. Text map explorer: a tool to create and explore document maps. In *IV 2006*, 2006.

[25] R. Feldman et al. Trend graphs: Visualizing the evolution of concept relationships in large document collections. In *Princ. of Data Mining and Knowl. Disc.*, LNCS. Springer, 1998.

[26] S. Havre et al. Themeriver: Visualizing thematic changes in large document collections. *IEEE TVCG*, 8(1):9–20, 2002.

[27] S. Liu et al. Tiara: Interactive, topic-based visual text summarization and analysis. *ACM Trans. Intell. Syst. Technol.*, 3(2):25:1–25:28, Feb. 2012.

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

[29] M. M. y Gomez, A. Gelbukh, and A. Lopez-Lopez. Mining the news: Trends, associations, and deviations. *COMPUTACION Y SISTEMAS*, 5(1):14–24, 2001.

[30] Y. Kim et al. Automatic discovery of technology trends from patent text. In S. Y. Shin and S. Ossowski, editors, *Proceedings of the 2009 ACM Symposium on Applied Computing (SAC)*. ACM, 2009.

[31] Z. Shen et al. BiblioViz: A System for Visualizing Bibliography Information. In *Proceedings of the 2006 Asia-Pacific Symposium on Information Visualisation - Volume 60*, 2006.

[32] C.-N. Ziegler and M. Skubacz. Towards automated reputation and brand monitoring on the web. In *Mining for Strategic Competitive Intelligence*. Springer, 2012.