CiteRivers: Visual Analytics of Citation Patterns

Category: Research

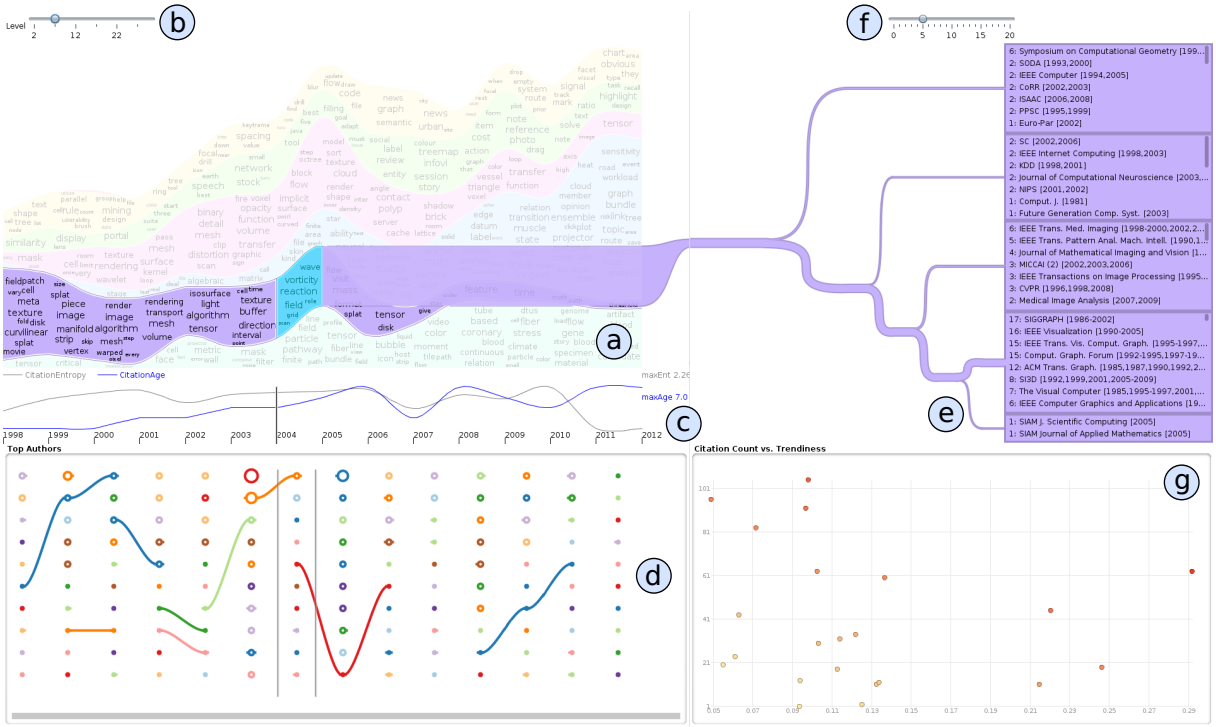

Fig. 1: CiteRivers consists of (a) the stream panel, (b) the level slider, (c) the citation aggregation panel, (d) the author panel, (e) the citation flow panel, (f) the categories slider, (g) the document trend plot.

Abstract—The exploration and analysis of scientic literature collections is an important task for effective knowledge management. Past interest in such document sets has spurred the development of numerous visualization approaches for their interactive analysis. They either focus on the textual content of publications, or on document metadata including authors and citations. Previously presented approaches for citation analysis aim primarily at the visualization of the structure of citation networks and their exploration. We extend the state-of-the-art by presenting an approach for the interactive visual analysis of the contents of scientic documents, and combine it with a new and flexible technique to analyze their citations. This technique facilitates user-steered aggregation of citations which are linked to the content of the citing publications using a highly interactive visualization approach. Through enriching the approach with additional interactive views of other important aspects of the data, we support the exploration of the dataset over time and enable users to analyze citation patterns, spot trends, and track long-term developments. We demonstrate the strengths of our approach through a use case and discuss it based on expert user feedback.

Index Terms—scientific literature, visual document analysis, visual citation analysis, streamgraph, clustering

1 Introduction

An awareness of thematic and structural changes and developments of a scientific community is germane to effective knowledge management. Understanding those dynamics is important for people new to a research field, just as it is for long term members of a community. The former can identify key topics, authors, and publications to gain inspiration for their own work, and learn how to position their publications. The latter are interested in the influence and importance of past developments on current trends to better reflect on and gauge the future evolution of the field. In order to gain new insight into a scientific community, analysts have to base their inquiry on its past and current scientific output in the form of publications. For a comprehensive picture of a field, users require flexible access to various aspects of these publications and interactive visual support to detect patterns. In particular, knowledge about the thematic dynamics of a field and their interactions with other scientific disciplines are crucial factors for the prediction of its future. Previous approaches for interactive visual analysis of scientific literature have ignored this link and consequently provided an incomplete picture of a discipline. With the approach presented in this work, we fill this analysis gap.

Visual approaches for knowledge management have primarily focused on visual analysis of citation networks, their development, topic or thematic evolution, and the creation of science maps for overview. Our approach particularly considers outreach of research to other communities relative to the topic dynamics of a field. We achieve this by taking into account sets of venues that are co-cited from within an area of research. Sets of venues have the advantage of being directly interpretable and they can be derived from an available set of publications, without requiring to get access to and deal with additional, large document repositories. Great value is added if the mentioned outreach can be tracked back to themes and topics developed in a scientific field over time. Previous interactive analysis techniques proposed for scientific documents have concentrated on one of two aspects: document contents and topic structure, sometimes paired with metadata analy-

sis, and citation networks, typically depicted as node-link diagrams. While these approaches are effective for analyzing different facets of the data, they do not support correlating content of publications with their citations. They thus miss the important aspect of integrating the scientific environment of publications and its dynamics.

To address these shortcomings, we link both – topics and venue citation – and support the analysis of this specific combination with adjustable automatic methods for extracting thematic content and citation patterns over time. In addition, we advance the state-of-the-art in scientific literature analysis by presenting a new Visual Analytics approach that enables users to make sense of publication sets by better understanding the dynamics within a scientific field. It is based on a new technique for the visual combination of the contents of publications with their citations and specifically cited venues. We integrate the popular visual streamgraph metaphor to depict thematic developments over time, and augment it with visual links to citations and cited venues. Through interaction with the visual abstractions of the dataset, users can correlate and filter different aspects of the documents. This allows them to develop hypotheses, and test them by iteratively drilling down to different aspects of the dataset.

The approach further comprises automatic methods to join, correlate, and aggregate data from multiple sources. From these sources we extract information about the community structure of neighboring fields, and the popularity of publications. This information is combined with the full text version of the documents containing their authors, abstracts, citations, and their textual contents to give users a full picture of the publications and their scientific context. Users can adapt the granularity of the automatically created abstractions to their analysis goals.

2 Related Work

Organizing, analyzing, and exploring human knowledge has been an active resarch area for a long time. Researchers from many fields have devised methods to categorize and make sense of the massive amounts of available scientific writings. Information science is a field particularly devoted to developing data analysis methods for this goal. Examples of such methods include measures of the prolificacy of authors [27], or the identification of research fronts [22], i.e. current cutting edge research topics. In addition, data mining techniques exist designed for scientific literature analysis, from new search methods [17] to summarization approaches for entire disciplines [39]. Multiple attempts have been made to depict bibliographic data, sometimes coupled with text analysis of various forms. Such techniques often produce visually appealing, static images of multiple disciplines of science called *science landscapes*. A comprehensive overview of them is given in [3]. Techniques to generate science maps have also been explored in the visualization community, e.g., by Fried & Kobourov [21], who use a geographic map metaphor to visualize thematic clusters and combine them with heatmaps to show overlays of subdisciplines on the larger map. This approach is related to ours in that a confined set of publications of a scientific discipline is depicted in relation to a larger, scientific landscape. All of these approaches are either pure data mining methods or offer mostly static visualizations of data aggregation and mining results. Our approach, on the other hand, facilitates the visual interactive analysis of topic dynamics of a field relative to the cited communities from these topics.

Other research into the analysis of scientific literature in the visualization community has been spurred by an InfoVis contest held in 2004 [18]. The contest entries focus on the visualization and analysis of 10 years of InfoVis publications, from its inception in 1994 to 2004. Its dataset contains the full text of all papers including metadata such as authors, titles, keywords, and year of publication. Each paper also includes a full list of references. The entries include systems, such as [28] that focus on the exploration of citation networks and the identification and interrelationships of influential authors. Wong et al. [51] describe the application of their text analytics tool IN-SPIRE to the InfoVis dataset to identify and track research topics and their development over time. Contrary to ours, their approach does not support citation analysis. The PaperLens system [30] also allows to track research topics and their popularity over time and allows the identification of the most often cited authors and papers per year. These three papers are related to our work as they include methods to explore and track topics over time, and identify the most prolific authors from a set of publications. Compared to the three systems, CiteRivers offers a new abstract representation of citations based on community structure that is correlated with the topic dynamics of the analyzed document set. Our approach also features a visualization of the publications per topic for the most prolific authors. Ahmed et al. [1] use a 3D representation of topics over time generated by a clustering technique and depict the citation links between these topics as straight lines. It differs from our approach in that it does not aggregate citations and only handles the ones contained in the dataset at hand.

Other research into the analysis and visualization of scientific literature includes CiteSpace II [7], an approach that is based on citation network analysis. It focuses on the visualization of the interplay between research fronts and their intellectual bases. An intellectual base describes the set of publications that are fundamental to a scientific field. The Eigenfactor project [49] is also based on citation network analysis using spectral methods to identify important journals across disciplines and analyze their intra- and inter-discipline citation links, but it does not take the publication contents into account. Our approach, in contrast, introduces a new dimension by visually linking topic dynamics and the scientific communities that are cited from the publications of a specific topic and facilitating their joint analysis.

The Action Science Explorer [16] supports scientific literature search and the exploration of major topics in a research field. Choo et al. [9] recently introduced UTOPIAN which uses non-negative matrix factorization, a new method for topic extraction, and an improved version of the t-SNE technique to create a 2D mapping of scientific documents. An interactive visualization of paper references organized in a tree structure is presented by Zhang et al. [53]. Stasko et al. [41] show citation links in a matrix-based interactive visualization, and Görg et al. [23] present an approach to correlate documents and other entities based on metadata and content through various visual and interactive means. PivotPaths [14] is an approach that visually integrates citation data with other document metadata. It combines authors, titles, and keywords into an explorable network of information. There are further visual approaches that particularly address scientific document retrieval. Koch et al. [29] support patent retrieval based on content and metadata queries and visual interactive techniques for query widening. Beck at al. [2] devised a web-based approach for browsing and exploring a set of publications. Heimerl et al. [26] present an interactive visual approach for document classifier creation and apply it to scientific literature. These approaches either provide means for interactive analysis of document contents, metadata, or citation networks. Non of them, however, aim to facilitate joint analysis of contents and citations, leaving a crucial analysis gap.

CiteRivers features an extended version of a streamgraph to depict clusters over time. Streamgraphs, or themerivers, are a type of stacked graphs that have been introduced by Havre et al. [25] as a visual metaphor to convey thematic developments in large document collections, and they have quickly proven an effective tool to visualize topic dynamics from a text dataset over time. They have further been applied to various other types data, such as baby name popularity [47], and have even found their way into the mainstream media [5]. Extended versions show thematic changes and interaction between topics in text datasets [13], and tag clouds have been used as topic labels [48]. Dou et al. [15] embed them into a tree structure to allow the exploration of hierarchical topics, and Wu et al. [52] combine them with a visualization of sentiment analysis results encoded by color. We extend streamgraphs with a flowgraph-based exploration technique. Flowgraphs have a long usage history in infographics and are popular as a map overlay to visualize the movement of goods or people. They have recently become popular in combination with new interaction methods for large tabular displays [43]. Phan et al. [36] present a flowgraph creation method that uses hierarchical clustering, similar to ours.

3 Visual Analytics of Topics and Citation Patterns

CiteRivers (Fig. 1) consists of five elements that contain different views of the document set. They are connected by brushing and linking to allow for an easy combination of the different depicted data aspects. The two central views, that implement our approach to link document contents and cited communities, is the *streamgraph panel* (Fig. 1a) and the *citation flow panel* (Fig. 1c). The former is situated in the left upper space of the desktop and contains groups of the publications depicted as streams of varying prominence along a time axis. The latter is situated to its right and shows an aggregation of the citations of the documents in a selected time step of a stream. All views on the left side of the desktop (Fig. 1a/c/d) are aligned to the time axis and contain time-dependent information. The views on the right side (Fig. 1e/g) display additional information for each focused time step of a stream.

3.1 Streamgraph Panel

The *streamgraph panel* (Fig. 1a) visualizes the topic structure of the dataset with the popular streamgraph method [25]. It is an aesthetically pleasing and readily comprehensible visualization scheme that is well established for visually integrating multiple time series, conveying individual values as well as their sums [5]. In addition, the flow metaphor of streamgraphs makes it possible to combine them smoothly with a flowgraph, which we integrated in order to link topics and cited communities, without breaking the metaphor. This results in a comprehensible, organic visualization that users can interact with naturally and smoothly.

Users can explore thematic clusters along the time line. CiteRivers supports two different ways of grouping publications that can be selected according to the user's analysis goals. One clusters documents hierarchically according to content similarity. This technique is further described in Sections 4 and 5. The second method for grouping uses metadata attributes of the documents. It allows users to group documents for example by the conferences they were published at, or the affiliations of their authors. If the user chooses the hierarchical method, the level in the clustering tree can be switched interactively with the *level slider* (Fig. 1b). This allows users to change the number of clusters to a granularity that fits their analysis goals. In case larger time spans are analyzed, the binning of documents can be made coarser than one year to maintain visual scalability.

Analysts can use the mouse to hover and explore the different clusters in the streamgraph. When mousing over a specific stream, it becomes focused and is highlighted by a higher saturation (purple stream in Fig. 1) compared to the other streams. This updates the views on the left side of CiteRivers's desktop to contain the information specific to documents in the highlighted stream. The areas for each of the time slices in a stream are separate visual elements that users can interact with (e.g. the turquoise area in Fig. 1). We call these the blocks of the streamgraph. Each block contains a word cloud that gives an impression of the contents of the publications that it contains. The terms are extracted from the abstracts of the publications, as described in Section 5. We place the terms starting at the center of the block following a spiral path towards its outer bounds [45, cf.], starting with the most frequent term. This results in the most frequent terms being places at the center of the block, and generally in visually appealing word clouds. In case the blocks are so small that only few terms fit into it, users can mouse over a block to highlight it and trigger a tooltip with a larger word cloud in a line by line layout that contains additional terms as depicted in Fig. 5. When a block is highlighted by mousing over it, it is marked with a turquoise background, and extended and attached on its right side to a flow into the *citation flow panel* (Fig. 1e). The left border of the focused block extends to the time line axis below the streamgraph panel, where it marks the corresponding year.

3.2 Citation Flow Panel

The *citation flow panel* (Fig. 1e) is situated to the right of the streamgraph panel. It is visually linked to the streamgraph and shows an aggregation of the citations of the documents in a selected time step of a stream. This is the second component of the central visualization of our approach. It introduces a visual link between the thematic streams and the cited venues, adhering to the stream metaphor. We like to see the citation streams as rivers that flow from the cited conferences or journals, i.e. the leaves of the flowgraph, to the streams of the streamgraph, with small tributaries joining to form a larger river that swells until the root node of the clustering and finally ends within the block of the focused stream. We have decided to use streamgraphs, as they are effective for conveying movement of objects or material from one point to another while keeping clutter at a minimum [36]. To further reduce clutter and increase readability, our layout algorithm introduces only binary splits. The result is easy to interpret and follow, and thus helps users to understand the distribution of citations from a focused block.

When a block is focused, the extension to its right is the end of a flow out of the citation flow panel. This is depicted in Fig. 1 for the highlighted block of the purple flow. The flow depicts knowledge from the cited communities that flows towards the citing documents, with individual streams coalescing on their way to the selected block. In the citation aggregation panel, multiple smaller flows start at clusters of publication venues. These contain similar conferences and journals listed by their names and the years of the citations. The entries in each list are sorted according to citation count, indicated by the number preceding the venue name. Details about each venue can be accessed by double-clicking on a list entry, to open a browser with the DBLP [32] page of the conference or journal. This page offers ample information about the venue. It includes all authors and publications for each installment or issue, and links to the conference or journal web page. Users can also select single venues or whole clusters, showing citation numbers for each block of the streamgraph (cf. Fig. 6).

The split-up of the stream is based on a clustering of all conferences and journals from the DBLP database according to community structure (cf. Section 5). Users can interactively adapt the granularity of the communities according to their analysis goals by setting their number with the *categories slider* (Fig. 1f). For example, on a low granularity level, visualization venues (e.g. IEEE InfoVIS or IEEE VIS) are in the same clusters as venues primarily focused on rendering (e.g. Eurographics or SIGGRAPH), but they get split with increased granularity. Changing granularity results in a smooth relayout of the updated tree, with the clusters remaining in the same order and approximately retaining their positions. The vertical space is distributed to the lists relative to their respective number of elements, with scrollbars for those lists that do not get enough space to show all of their elements at once. In addition, the publications citing the selected venues are highlighted in the document trend plot (Fig. 1g). The community clusters are subsequently connected to the selected block of the streamgraph by a binary cluster tree that represents the flowgraph. Its edges are drawn in a rounded fashion with their thickness scaled according to the number of the cited venues they dominate. The layout algorithm assigns equal horizontal space to both children of each node, starting from the root node. It thus distributes all available space, resulting in a balanced layout of the flowgraph's branches.

3.3 Citation Aggregation Panel

The *citation aggregation panel* (Fig. 1c) is located right below the streamgraph panel. While the citation flow panel visualizes local distribution of citations per block of the streamgraph, users need an aggregated view of the citation behavior and changes therein over time. This panel provides this view by depicting plots of two characteristic values of the citations in each block along a stream highlighted by the user. We have chosen them to help users track the evolution of citation behavior and points them to potentially interesting blocks that can then be analyzed in greater detail.

The blue curve shows average citation age of each of the blocks of a stream. It is the average age of every reference of a publication in a block at the time the document was published. From this curve, users can learn how far back publications in a specific year and stream are sourcing their references. Citation age also indicates how new and trendy a topic of a stream is, as less trendy topics will be based on work that is potentially older, while documents addressing trendy topics will

more likely cite similar recent publications.

The second, gray, curve depicts the citation entropy along the highlighted document stream. It is calculated on the different conferences and journals cited in the focused block according to Equation 1, where V is the set of cited venues, and $\#cites(v)$ denotes the number of citations to venue v in the block. The citation entropy measures the diversity of the citations of the publications along the stream based on cited publication venues.

$$H(V) = -\sum_{v \in V} \frac{\#cites(v)}{\sum_{v' \in V} \#cites(v')} \log \frac{\#cites(v)}{\sum_{v' \in V} \#cites(v')} \quad (1)$$

This helps users assess how widespread documents in a stream cite publications from different academic disciplines and how that particular behavior changes over time. The entropy, for example, rises, if documents in a stream start to cite publications from a new scientific community in addition to their traditionally cited fields. To make orientation easier for the analysts, the citation aggregation panel features a vertical line that marks the year of the currently highlighted block in the streamgraph.

3.4 Author Panel

The *author panel* (Fig. 1d) is situated right below the citation aggregation panel. While we were discussing and designing our approach, it became evident that users who know a community typically connect research topics with authors. Our approach consequently includes a view that lets users refer to the most prolific authors of each block of a stream. We show a selection of the ten most prolific authors of each block n order to give users orientation and support them with understanding and interpreting topic streams. While this number could be increased, we found that ten is a good number in practice that provides ample author references while not overwhelming users with too much choices. We measure author prolificacy by the number of publications authored per year. In the author panel, circles of different sizes are arranged according to a matrix, with each column corresponding to one block. Each column contains up to ten circles that represent authors, ordered according to their rating. The size of the circle is also chosen according to the author's rating.

When a circle is moused over, it displays the name of the corresponding author right above the circle. To color the circles, we relied on ColorBrewer2 [24] for a color mapping of 12 distinct colors. As we do not have a color for each author, we have to use the same color multiple times for different authors. In order to avoid confusion, we link authors that occur in a sequence of adjacent years with a curve of their respective color. If authors do not occur within the matrix in adjacent years, but further away, we add a stub to the left or right of their circle, indicating an outgoing edge to an earlier or later year, respectively. Stubs are connected when the circle is moused over, linking all instances of the corresponding author. Thus, users can easily track a specific author over time. Authors can also be selected by clicking their circle. All blocks in which the selected author has published are then highlighted in the streamgraph panel and show the respective number of publications. This is depicted in Fig. 8. In addition, all publications of the selected author are highlighted in the document trend plot (Fig. 1g).

3.5 Document Trend Plot

The *document trend plot* (Fig. 1g) is situated on the right side of the desktop, below the citation flow panel. When exploring the topic distribution and dynamics of a dataset, users need access to the single documents. This is important to find evidence in favor or against hypotheses, gain new insights by directly referring abstracts or full documents, and find new work that matches a users' research interests. To give users orientation among the document set backing a block of the streamgraph, we show a document scatterplot that, unlike content-based document spatialization techniques [50], layouts the documents according to two dimensions. These give users an idea about the popularity and success of each publications. This is achieved by plotting citation count of each document against a trendiness score. While the first one is a quantification of popularity, the latter quantifies the novelty and success of ideas within a paper, as described in Section 4, and is a local extension to the global topic dynamics in the streamgraph. Seldom cited publications with a low trend value reside in the lower left corner of the plot, while highly cited ones with a high trendiness score populate its upper right corner. This means that documents with less trendy content are situated on the left side of the plot, towards its past, while more trendy ones are closer to their future towards the right edge of the scatterplot.

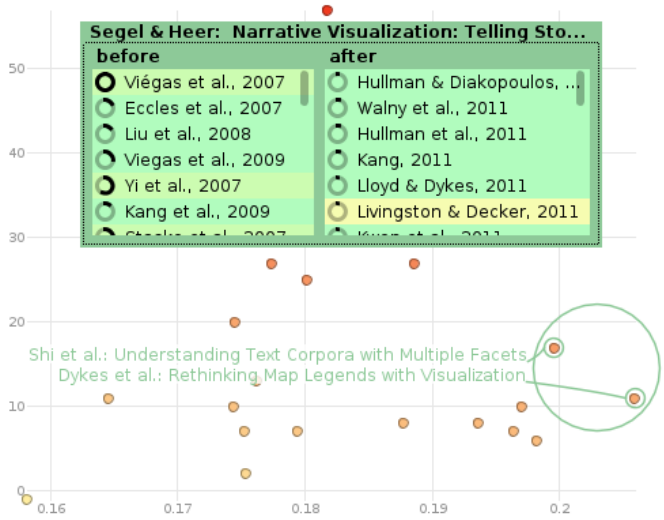

Fig. 2: The document trend plot with the lens showing authors and title for two documents. The detail panel has been activated for the publication right above the panel. It lists its most similar documents.

Users can explore the space by using the radial title lens (cf. Fig. 2), an excentric labeling technique [19], that reduces clutter in the scatterplot by avoiding showing all document titles at once. It can be activated by clicking into the free space of the plot, and shows authors and titles, visually linking them to the respective document. The title lens always has the same color as the currently highlighted stream, and its size can be adapted using the mouse wheel. To learn more about the depicted document set, users can click on the document glyphs to activate a detail panel (cf. Fig. 2). Apart from containing author and title of the document, these panels show two lists of the documents in the dataset that are most similar to the selected one. The left list contains those earlier, while the right one contains those published later. As the trend score for the documents is computed based on these most similar ones (cf. Section 4), the detail panels serve as an explanation for the scores. The listed documents are colored according to the stream they are part of, which gives the user information about the distribution of clusters in this particular set of documents. Relative citation counts for the listed documents are depicted as radial donut charts, normalized to the largest count in the lists of a panel. Absolute counts are available via tooltips for each of the list entries.

4 Data Mining Methods

To help users make sense of the document data, CiteRivers includes data mining and aggregation techniques. We have chosen these methods to facilitate the visual and interactive analysis and interpretation of the data. This section discusses the technical background of the methods and the particular reasons we decided to include them into the approach.

4.1 Spectral Clustering

We use clustering to facilitate the interactive exploration of two different aspects of a set of scientific publications: (1) the thematic structure of the analyzed articles including their dynamics over time, and (2) the conferences and journals cited by a set of articles. By integrating clustering with suitable interactive visualization, we support users

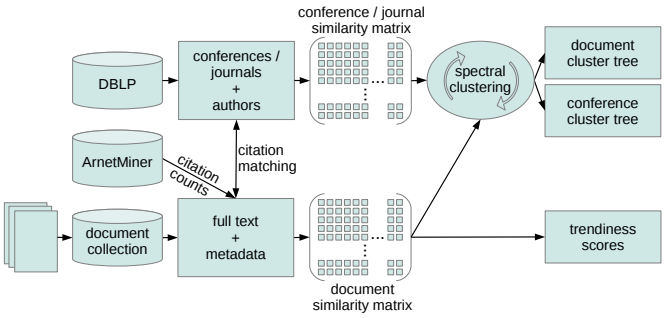

Fig. 3: CiteRivers' data processing pipeline with three data sources: (1) the raw publications for full texts and metadata, (2) the DBLP database for publication venues, and (3) the ArnetMiner database for citation counts. From those, we create similarity matrices of documents and venues to feed them to the spectral clustering algorithm and compute the trendiness scores.

in abstracting from local properties, and exploring higher-level structures of the data. In the following section, we will call all objects to be clustered *instances*, regardless of whether they are documents or conferences and journals. We decided to use spectral clustering [46] for both types of instances for two reasons. Firstly, it is a top-down hierarchical method which generates a cluster tree instead of a flat set of clusters. This lets users interactively adapt the number of clusters by splitting them without modifying their boundaries, which is conducive to creating and keeping a mental map of the dataset. In addition, the cluster tree is binary, i.e. exactly one cluster is split at each level of the tree, allowing for a very fine-grained adjustment of the cluster granularity. Secondly, spectral clustering is not biased with respect to the shape of the clusters that are generated and is thus capable of yielding more natural clusters compared to other methods.

The spectral clustering algorithm recursively searches for the best cuts in a neighborhood graph, i.e. a set of severed edges in a graph that partition the set of nodes. For this, we transform the instances into a k nearest neighbor graph, using $k = 10$ for publications and $k = 14$ for venues. We determined these values using Luxburg's method [46]. One criterion used to identify the best cut is high intra-cluster similarity and low inter-cluster similarity. This is achieved by severing edges whose sum of weights are as low as possible. Using this objective alone would split off many small clusters that have weak links to the remaining instances. To solve this problem, a second objective that balances cluster size is typically included. An objective that incorporates both criteria is the normalized Cheeger cut [6]. We use it for our approach as there are results that suggest that it outperforms other objectives [4]. The Cheeger cut is defined as in Formula 2, where C and $\overline{C}$ represent two candidate partitions of the graph during an iteration of the algorithm.

$$NCC(C,\overline{C}) = \frac{\sum_{i\in C, j\in\overline{C}} w_{ij}}{min(\frac{1}{2}\sum_{i,j\in C} w_{ij}, \frac{1}{2}\sum_{i,j\in\overline{C}} w_{ij})} \quad (2)$$

Finding a solution that minimizes Equation 2 is an NP-hard problem. Luckily, it can be efficiently approximated using spectral methods. Buehler and Hein [4] show that an approximate solution for the optimal Cheeger cut can be found through the second smallest eigenvalue of a Laplacian matrix of the neighborhood graph. We use their implementation of this technique in our prototype.

4.2 Trendiness Score

The trendiness score gives an impression of the freshness of ideas in a publication and their dissemination into later ones. It thus captures one important aspect of publication success that is independent of citation numbers. Our score is based on document similarity and is thus akin to Shaparenko et al.'s lead/lag index [40]. It is, however, more flexible because it considers all available past and future documents, while the lead/lag index relies on a fixed neighborhood of size k in a document's past and future years. Another measure for publication success is presented by Chen et al. [8] which is an adaption of the h-index [27] and is thus entirely based on citations. As the document trend plot (Fig. 1g) already combines raw citation counts with the trendiness score, such a citation-based measure would not introduce a new dimension to the analysis of a publication's impact. Both measures combined give users a deeper insight into the context of a publication than each one individually. An example for this are literature surveys, which tend to attract numerous citations as they summarize the state-of-the-art in a field, but typically do not contain any technical innovations. Literature surveys thus get high citation counts with low trendiness scores. Thus their impact cannot be captured adequately by the individual measures.

To quantify the freshness of ideas of a publication d_i, we take a look at the set of documents published earlier, $D_{before} = \{d_x \mid \text{earlier}(d_x, d_i)\}$. We estimate the influence of earlier documents based on their similarity to d_i and calculate the impact score I_{D_{before},d_i} of D_{before} on d_i according to Formula 3.

$$I_{D,d_i} = \sum_{d_j\in D} e^{-\frac{s(d_i,d_j)^{-2}}{\tau^2}} \quad (3)$$

Depending on whether the documents in D are older or newer than d_i, Formula 3 quantifies the influence of D on d_i, or of d_i on D, respectively. The influence measure is based on the cosine similarity of documents $s(d_i,d_j)$, assuming that influenced publications imitate the influencing ones to a certain extent. Though not capable of capturing all aspects of scientific interaction within a community, this assumption is able to measure a publication's impact without considering citation count. After estimating how much the ideas in d_i become popular in future publications, $D_{after} = \{d_x \mid \text{later}(d_x, d_i)\}$, by computing I_{D_{after},d_i}, the overall trendiness score is determined by weighting the influence on d_i against the influence of d_i on future work (Formula 4).

$$\text{trendiness}(d_i) = \frac{I_{D_{after},d_i}}{I_{D_{before},d_i}} \quad (4)$$

The Gaussian kernel, whose size is defined by τ, determines the similarity level at which we assume that one publication influences the other. We experimentally determined a value for τ by iteratively refining it on our dataset, resulting in $\tau = 10$.

4.3 G2 Keyword Extraction

The stream panel features small word clouds that contain representative terms of the abstracts of the papers in a block. They help users get an impression of the themes that distinguish a block from all others within and outside of the same stream. Multiple schemes exist to select and weigh terms for word clouds from document collections. The most straightforward is term frequency (tf), i.e. raw term counts. Its drawback is that the domain specific vocabulary of the dataset dominates the word clouds, resulting in only minor variations between blocks. For our dataset of visualization papers, the term *visualization* has a high frequency in most blocks, but it is arguably not very informative for users analyzing the clusters. A popular method to downweigh such terms is tf-idf (term frequency / inverse document frequency) [38]. It uses the logarithm of the inverse number of documents a term occurs in to balance frequency with term popularity. While versions of tf-idf have been successfully used to extract distinguishing terms [44], Chuang et al. [10] show that it is significantly outperformed by probabilistic methods in yielding terms expected by humans.

To give users a good impression of a block's idiosyncrasies, we decided to use the G^2 metric [37]. It exhibits high keyword extraction accuracy [10] and has the additional benefit of representing statistical significance values for the different frequencies of a term's occurrence within and outside of the block [11]. The score is calculated by counting the occurrences of each term w in each of the blocks, resulting in a contingency table as shown in Table 1. Based on the frequency of a term in the entire dataset, expected frequency values for each block

	block	¬block	total
freq t	a	b	a+b
freq $\neg t$	c-a	d-b	a+d-a-b
total	c	d	c+d

Table 1: Contingency table for term t, adapted from [37]. The columns *block* and $\neg$*block* denote the current block, and the remaining blocks, respectively.

and the remaining set of documents are computed (Formula 5).

$$E_1 = c \cdot \frac{a+b}{c+d} \quad \text{and} \quad E_2 = d \cdot \frac{a+b}{c+d} \tag{5}$$

$$G^2 = 2a \cdot \log \frac{a}{E_1} + 2b \cdot \log \frac{b}{E_2} \tag{6}$$

The difference between expected and real frequency values for term w is tested for statistical significance under the null hypothesis that differences are of a purely random nature (Formula 6). The G^2 score thereby approximates a χ^2 distribution that makes it possible to derive probability values for the null hypothesis using the standard χ^2 calculators or tables. Higher values of G^2 correspond to lower probabilities for the null hypothesis, allowing us to rank and size terms in the word clouds according to that value.

5 Data Aggregation and Processing

The presented approach includes data from multiple sources that is joined and aggregated. This section presents and discusses the different datasets and the processing and aggregation steps. The chosen steps and techniques are known to work well for extracting and comparing textual content. Fig. 3 depicts the entire process.

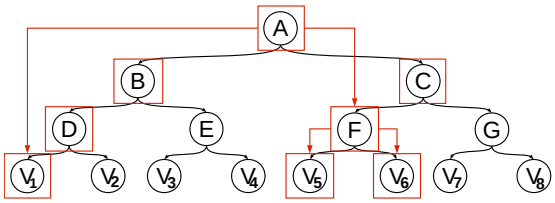

Fig. 4: Example community tree for eight conferences (V_1 to V_8) with an overlay for a block that contains V_1, V_5, and V_6. The marked nodes are dominating a relevant venue. The overlay tree includes all marked nodes that have either two or no marked children.

5.1 Processing of Publications

The central document set of CiteRivers is the collection that the user selects for exploration. It is a collection of scientific publications that are either available in a structured format, or that have to be extracted from pdf files or even scans. Our approach is applicable to all datasets that contain the full texts of the documents, metadata including authors, titles, and abstracts, and complete reference lists. As shown in Fig. 3, each of the documents in this dataset is matched against the ArnetMiner [42] database and its citation counts are extracted from it. The ArnetMiner database comprises all of the entries in DBLP enriched with additional information, such as citation counts.

In the next step, the text content of the documents is transformed into the vector space model [33] which represents documents as vectors of frequency counts for each word. Before we can create these vectors, we linguistically preprocess the texts by applying a tokenizer, a lemmatizer, and a stop word removal scheme. Tokenization separates single tokens (words) from the sequence of characters that represents the text. Then, a lemmatizer transforms these tokens to their lemmas, the base or dictionary form of a word. This step removes e.g. plural forms of nouns, or conjugation of verbs. For both tokenization and lemmatization we use the Stanford CoreNLP package [34]. The next step is a quite aggressive stop word removal method. Instead of traditional stop word removal, which uses a fixed list of words that contain no information in isolation, we remove all terms that occur in more than 60% of the documents. We found that this results in the removal of many frequently used words that introduce noise and results in more compact and informative vectors that help to distinguish better between documents of different topics. Based on the resulting vectors, we compute a matrix of pairwise document similarities using cosine similarity [33]. It defines similarity $s(d_i, d_j)$ of two documents d_i, and d_j as the cosine of the angle between their vectors: $s(d_i, d_j) = (\vec{d_i} \cdot \vec{d_j})/(|\vec{d_i}| \cdot |\vec{d_j}|)$.

The resulting matrix serves as the basis for the document trendiness score, and can be fed into the clustering algorithm to create the thematic clustering of the dataset (cf. Fig. 3). If the user wishes, the clustering algorithm can be disabled, and the documents can alternatively be grouped according to various aspects of the metadata (cf. Section 3).

5.2 Communities and Reference Extraction

In our approach, we aggregate citations by grouping publication venues according to their scientific communities. The references of each of the publications in the dataset are then mapped to the groups that contain the venue it was published at. As mentioned in Section 2, previous works exist that extract scientific communities using either citations, co-authorship, or content, for example [35, 7]. Our approach to detect the influence of different research disciplines based on cited venues and author overlap between venues, is new to the best of our knowledge.

5.2.1 Hierarchical Communities

We extract communities based on the DBLP dataset [32] of computer science publications. It contains roughly 1.6 million entries, and is the largest available dataset of scientific publications. In addition to publications, DBLP has entries for publication venues, linking each of the documents to the conference or journal where it was published. As depicted in Fig. 3, we extract all conferences and journals contained in DBLP. In addition, we collect the author names of each document and the associated venue. We keep each issue of a journal or installment of a conference as separate entities, allowing us to better model the thematic dynamics of venues over time.

In the following step, we create a similarity matrix for the extracted venues based on their author overlap. This is a way of modeling community affiliations of conferences and journals. We assume that the more authors that publish at venue A and at venue B, the more similar both venues are in terms of the scientific communities they are part of. We model this using the Jaccard coefficient as a similarity measure for venues. The Jaccard coefficient is a way of quantifying the overlap between two sets of entities: $jaccard(A,B) = (|A \cap B|)/(|A \cup B|)$. Spectral clustering is applied to the resulting matrix to create a hierarchical community structure.

We chose to use conferences and journals as the base categories for creating a community hierarchy for two reasons. Firstly, we have decided against the obvious alternative of creating and clustering a co-author network from DBLP (such as [35]) because DBLP lists about 1.5 million authors. Partitioning the resulting large co-author graph would take enormous computational resources and thus processing time. In addition, we would only use the first couple of levels of the resulting hierarchy, with the rest being too fine grained and therefore irrelevant. Another disadvantage of co-author networks is that we would have to tackle the hard problem of author name disambiguation [20], as two authors with coincidentally identical names would distort the partition results. Using larger entities, such as publication venues, as a basis is much more robust against this problem, and we found that we can get sufficient results by just treating these ambiguous names as noise. The second reason, we have decided to use conferences and journals as the base categories of our community clustering is that their names are much easier for users to interpret compared to the names of single authors.

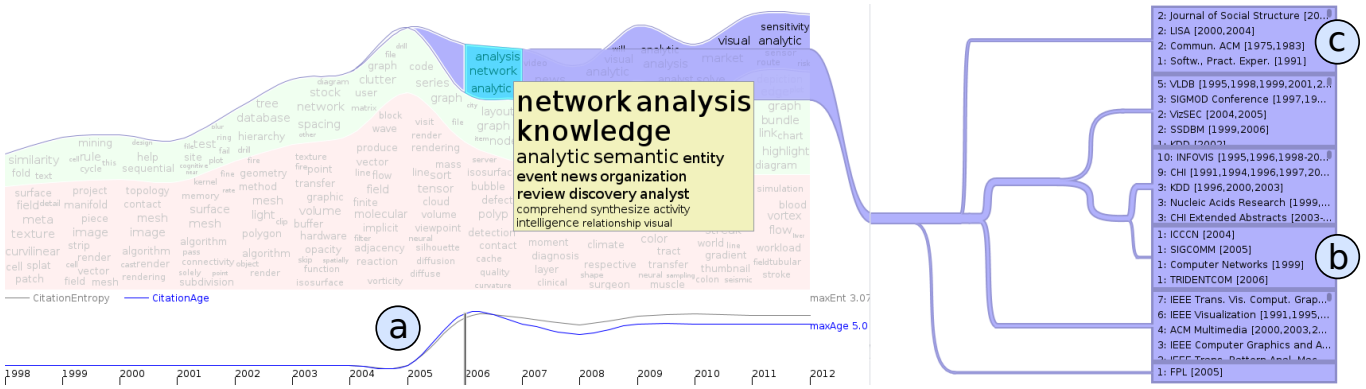

Fig. 5: For the first year of VAST the citation aggregation panel (a) shows a dent for both average citation age and entropy (2007–2008). The tooltip contains a larger word cloud with further terms for the selected block. Citations to computer network venues (b) and the *Journal of Social Sciences* (c) are shown in the citation flow panel.

Community hierarchies are stored, and used to aggregate citations each time a user highlights a block. The result is then visualized in the citation flow panel (Fig. 1e), showing only the venues of references of the highlighted block. The aggregation is created by computing an on-the-fly overlay on the community tree containing only venues relevant for the block. As depicted in Fig. 4 this is done by recursively marking all clusters in the tree that contain at least one of the relevant publication venues. In a second step, a new, temporary tree is created, spanning only those nodes that contain relevant venues and either have no or two children. The partial community hierarchy contained in the new tree is then depicted in the citation flow panel (e) of Fig. 1. As the clustering for each set of references is always based on the same basic tree, contradictory combinations of venues in two different blocks cannot occur. This helps users with their analysis of the data by creating and keeping a consistent mental map of the community structure of publication venues.

5.2.2 Reference Extraction

To be able to map the references of each of the publications in the dataset to their community cluster, we need to extract the conference or journal of each referenced document. Using the venues from the reference strings is not possible, because venues are not referenced in a standard way, and authors use different names and abbreviations for the same venue. To solve this problem, we use the DBLP database again, and find the corresponding entry for each reference. These entries are linked within DBLP to the respective conference or journal where they were published, giving us unambiguous identifiers for each venue.

To find the entry for a reference, we use its title string and compare it to all DBLP entries. For this, we use the Levenshtein distance [31] that measures string distance as the number of character insert, delete, and exchange operations needed to convert one string into the other. A title is matched to its most similar DBLP title above a threshold that we determined iteratively by manually reviewing the results. This fuzzy matching mechanism allows us to handle small differences in the string caused e.g. by orthographic variations. With this technique we were able to match approximately 70% of references to DBLP entries. All unmatched references are currently ignored by our implementation. Once we have mapped a reference to its associated venue in DBLP, we can use this information to create community trees for each block with the method described above.

6 Use Case

This section presents an analysis example that showcases the capabilities of our approach and its implementation. We first describe the preparation of a dataset we created for this use case, and then describe an example analysis session step by step. Although the use case includes all of the features and elements of the approach and our prototype, it focuses on showing the benefits of linking topic dynamics and aggregated citations, as this is the main contribution of this work.

6.1 Dataset

We have prepared a dataset of publications from the IEEE VIS / VisWeek conferences covering the years 1998 to 2011. It includes all full papers for the three main conferences, IEEE SciVis, IEEE InfoVis, and IEEE VAST since 2006. The set of 1336 documents, only available in pdf format, contains 390 InfoVis publications, 797 VIS publications, and 149 VAST publications. All pdfs have been converted to plain text using a commercial OCR solution[1]. We use OCR rather then extracting the text directly from the pdfs because we found that it significantly improves the quality of the resulting documents. Available text extraction tools typically exhibit problems with identifying and extracting whitespace, special characters, and the general structure of documents, especially from a two column paper format. We then used ParsCit [12] to extract the metadata from these documents, including their reference lists. The references have been mapped to their respective DBLP database entries, as described above, to extract corresponding conferences and journals. Other metadata, comprising document titles, abstracts, and author names have been checked manually, and corrected if necessary to eliminate OCR errors and assure high data quality. In addition, citation counts for each of the publication have been extracted from the ArnetMiner database, as described in Section 5.1.

For the analysis of the dataset, we created one metadata grouping and one hierarchical cluster tree for the publications. Both can be displayed in the stream panel (Fig. 1a). The metadata grouping is based on the conference attribute of the documents, grouping them by the conference they were published at, VIS, InfoVis, or VAST, respectively. This grouping gives insight into the development of the three conferences with respect to their topics over time, their interactions in terms of topics and authors, and the dynamics in citation behavior. To cluster publications according to their contents, we used the spectral clustering methods including the pre-processing and term filtering schemes discussed in Section 5.1. The resulting cluster tree contains 50 levels, which we found sufficient for a comprehensive analysis.

6.2 The First Years of VAST

We start our analysis of the VIS dataset by inspecting the streamgraph for the three conferences, as depicted in Fig. 5. It shows a red SciVis stream at the bottom, a green InfoVis stream in the middle, and a purple VAST stream on top. The stream for VAST starts in 2006, its first year. A conspicuous fact shown in the streamgraph is the varying number of papers for the conferences. This is encoded by the height of the respective stream at each timestep. We can see that acceptances

[1] Nuance's Omnipage Professional 18

for SciVis peak in 2004, and then drop again to their past level. InfoVis acceptance numbers, on the other hand, slowly rise over the years. This trend is also reflected by the official acceptance rates for the conference, according to which it has doubled between 1998 and 2010. The acceptance rate of VAST has remained steady during its first five years.

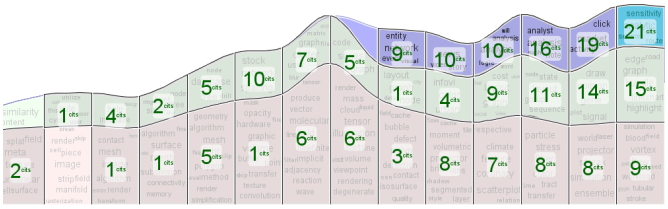

Fig. 6: VAST (top) has the highest number of citations to data mining venues, but citations from SciVis and InfoVis are also rising.

We are particularly interested in VAST, the youngest conference in the VIS family. VAST's topic is Visual Analytics, a field that combines visualization research with techniques from data mining to support human sensemaking for the analysis of large datasets. We are particularly interested in the communities that VAST publications cite, and start exploring them in the year 2011 using the citation flow panel (Fig. 1e). We can see that the largest clusters contains citations to expected venues from the visual analytics and information visualization community, e.g. *VAST*, *InfoVIS*, and *Information Visualization Journal*. Interestingly, there are also a significant number of data mining venues, e.g. *KDD (Knowledge Discovery in Databases)* and *Journal of Machine Learning Research*, and other data mining oriented conferences such as *Bioinformatics* and *ACL (Annual Meeting of the Association for Computational Linguistics)* in the same cluster. These attract our attention, and to analyze them further, we separate them into their own cluster by slightly increasing granularity (from five to six clusters) using the categories slider (Fig. 1f). Selecting this cluster gives us an overview of the citations to these conferences in the dataset, as shown in Fig. 6.

Although the data mining community has been cited before by InfoVis and SciVis, VAST has a comparably high absolute number of these citations. The difference is even larger when considering relative numbers, given the fact that the number of VAST publications is lower compared to InfoVis and SciVis, from 2006 throughout 2011. We further notice that the citation numbers to data mining venues from InfoVis and SciVis also seem to increase in later years, and wonder whether this is caused by mutual influence between the nascent Visual Analytics community and InfoVis and SciVis. An effect of this exchange might be the growing number of citations to VAST from InfoVis and SciVis, as shown in Fig. 7. We discover a second hint by exploring the authors of VAST and their "publication paths" with the author panel (Fig. 1d). It shows the publications of a selected author for each block as an overlay on the stream panel (Fig. 1a). Fig. 8 depicts a typical "publication path" of a scientist in the VIS community that starts with SciVis or InfoVis, and eventually also includes VAST publications. By exploring further authors, we learn that many of the highly prolific authors of VAST have made a similar transition. We view this as an indication of an avid exchange of ideas between the VIS conferences.

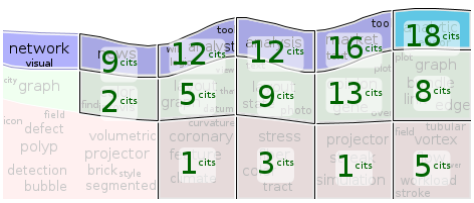

Fig. 7: Rising numbers of citations to VAST from VAST itself (top), but also from InfoVis (middle) and SciVis (bottom).

During our analysis session, we discover another interesting finding about the citing behavior within VAST. As can be seen in Fig. 5a, the average age of the cited publications in its first year is higher, with a dent visible in 2007 and 2008 followed by a stable level until 2011. Comparing this with the citation numbers from Fig. 7 suggests that the first value is higher because authors, not being able to resort to related work from previous years of the same conference, cite older material they find relevant for their work. Then, having access to fresh material from the same conference, prefer to cite these very young publications. The reason for the slight increase of the average age of within-conference cites is that the average age of the cited material grows, as the first VAST publications get older. We can see that after 2008 the average citation age levels out and stays roughly constant until 2011.

Focusing on citations so far, we now shift our attention to the thematic dynamics of the new field. We start with the word clouds that give us an impression of the topics specific for each installment between 2006 and 2011. In the first year of VAST, the term *network* is quite prominent and attracts our attention. The citation flow panel provides background information, as it includes one clusters that exclusively contains computer network conferences (Fig. 5b). In addition, the *Journal of Social Structure* in a different cluster (Fig. 5c) catches our attention. Based on these findings, we hypothesize that an unusual number of computer network analysis and social network analysis publications cause the prominence of the term *network*. Highlighting the publications that cite the respective conferences in the document trend plot (Fig. 1g) reveals four papers citing network conferences, and one citing the *Journal of Social Structure* twice. One of the computer network publication has the highest trendiness score in that year, while the other three have mid to low values. A high trendiness score indicates follow-up work influenced by the paper in later years. Selecting the cluster of network conferences shows the blocks in the streamgraph that contain publications that cite them. We see that VAST has further citations to these conferences in 2010, and we select the corresponding block for further analysis. In the updated citation flow panel, the network community cluster has grown by further venues. We select the whole cluster, and six publications are highlighted for further investigation. Along the same lines we can follow and analyze the development and evolution of other topics, financial analysis and text analysis being two further examples.

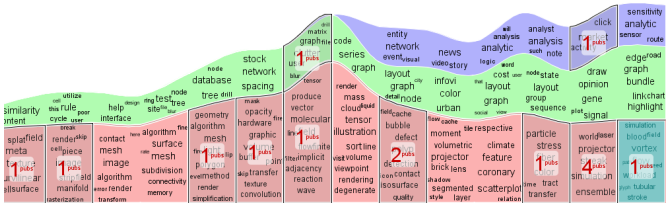

Fig. 8: Typical "publication path" of an author moving from SciVis and InfoVis to VAST.

7 Expert Feedback and Discussion

The discussion of the effectiveness of the approach is based on feedback from six experts, three active members of the natural language processing (NLP) community, and three active members of the visualization community. We prepared a different dataset for each group. The VIS dataset (see Section 6 for details) for the latter, and a dataset based on the ACL anthology[2] for the former, using the same techniques and steps as for the VIS dataset. The NLP dataset includes publications of three major venues from the NLP community: the ACL conference, the Journal for Computational Linguistics, and the EMNLP conference from 2000 to 2013. The visualization experts were associated with our department, one with long-term experience

[2]https://aclweb.org/anthology/

in the field, one PhD, and one PhD student. The group of NLP experts consisted of two postdoctoral researchers and one senior member of the field. After a tutorial of the prototype, we invited each expert to start analyzing the dataset from their community, exploring whatever they find interesting with both the content clustering and the conference grouping. They voiced all of their thoughts and findings which we, adhering to the think-aloud scheme, recorded on paper. Finally, we asked for their opinions on possible application areas, about the interpretability of the visual representations, and the interactions. We also asked for feedback on any missing aspect of the document data that would have supported their analysis.

All experts appreciated the CiteRivers approach as an effective top-down method for the analysis of scientific communities. They considered it useful for grant reviewers or conference planners to assess the dynamics of a scientific discipline. It could also help researchers new to a discipline, such as new PhD students or researchers looking for new publications possibilities. One of the experts also mentioned science journalists who could use CiteRivers for their journalistic inquiry. The experts stated that the prototype is fun to work with, and that they like the high degree of interactivity. One expert remarked that the approach demands some initial user training in order to grasp the data abstractions and their meaning. We agree that some initial training is necessary, but learned from our feedback sessions that the visual encodings and interactions can be learned quite fast, and experts were able to use the prototype after a couple of minutes of training. This is also corresponds with our experience from multiple demos we gave to researchers of various backgrounds, including humanities and social sciences, who quickly learned to read and interpret the visual representations.

An interesting finding from the feedback sessions is the difference in citing behavior between the NLP and the visualization community. While the latter has a broader citation behavior and tends to cite work outside of its community, citations from the former are much more narrowly focused on NLP and data mining venues. For better orientation among the frequently cited venues, the NLP experts suggested to highlight uncommon or otherwise interesting venues or clusters to steer the users' attention to them. One of the NLP experts was exploring the topics very closely, stating that thematic communities such as machine translation, parsing, or sentiment analysis can be followed quite well and that being able to adapt cluster granularity is important to find the right clustering for this. The two other NLP experts also mentioned that the streamgraph for the content clustering depicts the topics within the community quite nicely. Although there are hardly citations outside of the community, different foci of the aggregated venues helped to better understand and disambiguate topic streams. The expert focusing on the topic streams found that the machine translation cluster includes methodologically similar techniques such as information extraction, capturing the evolution of this topic. He further mentioned that the authors shown for the topics fit his perception of group structure of the field.

Although the experts found the depicted information about the publications quite comprehensive, one aspect that was missing, as remarked by four of them, was information about co-authorship relations, author affiliations, and citation relationships between authors. For the author panel (Fig. 1d), co-authorship, or affiliation information could be used to group authors within a column. In addition, dynamic graphs that depict changing collaborations between different authors can expose a lot about the developments in a field. Citation relationships between authors, on the other hand, could help to model the thematic relations between authors and groups of authors. A great challenge for the analysis of these relationships is the limited availability of large datasets that contain a comprehensive account of citation relations. There were additional comments concerning single elements of our approach. The trendiness score attracted two experts' attention, who extensively contrasted its results to other metadata aspects of the documents. This led to insights into the significance of a publication, and thus into the structure of the field. While the experts found that for some publications the results of the score are reasonable, for some others the score differed from what they had expected. This was cause for some criticism directed at the opaqueness of the score due to its level of abstraction. Although the document trend plot (Fig. 1g) offers lists of similar publication in past and future, the reason why these publications are similar remains unclear. We agree with the criticism and plan to add a suitable representation of common terms that two publications share as an explanation for their similarity rating.

In addition to the points discussed so far, there was also very distinctive feedback about the functionality of a specific element of the prototype. One visualization expert mentioned that it would be nice to zoom in onto a single stream in the streamgraph (Fig. 1a) for close inspection, hiding all others. The same expert also asked for labels describing the content of a whole stream. Two of the NLP experts stated that depending on the clustering granularity, topic streams can get quite small and thus hard to read. Being able to zoom in on a group of clusters would increase the flexibility of the approach, and we thus intend to include this feature in the future. Another feature that was mentioned by two experts and that we plan to include is linking the keywords in the streamgraph to the papers that contain them in the document trend plot (Fig. 1g) as an explanation for the labels of the block.

Overall, the feedback was positive, and all experts were able to discover new findings and gain new knowledge about their community. These included facts about authors and the prominence of publications, topical developments, and their citations into the same and other communities. In addition to the feedback reported above, the discussions with the experts also yielded analysis questions that are out of scope of the presented approach, constituting interesting research questions on their own.

8 Conclusion and Future Work

In this work, we have introduced a new Visual Analytics approach for the analysis of scientific literature. It comprises a novel technique to visually link grouped article sets with a user-steered abstraction of cited venues. We demonstrated the capabilities of our approach with a use case, for which we created a corpus of visualization publications. To assess the usefulness and applicability of the approach, we asked six experts for their feedback. We find that our approach is effective for gaining insight on the thematic dynamics of a scientific field, and their relation to other communities through their citations.

For future work, we plan to extend our approach in three ways. First, we intend to include information about author collaborations and networks into the approach. An interesting research question in this respect is the connection between topic dynamics and author collaborations. In particular, we want to support users in answering questions such as: How do authors collaborate across field boundaries, and how do collaborators of single authors and groups change depending on the topic they collaborate on? Second, we are aiming to extend the current top-down approach with bottom-up analytical capabilities that let users integrate and organize previous knowledge and hypotheses into an analysis sessions, and iteratively enrich and revise this knowledge with new insights gained during the analysis. Third, we are working on adapting the approach to a new type of scientific documents, namely patents. While patents have many common properties with scientific publications, they also differ in important aspects. Two of the challenges that we face with patents are their specific type of language, as well as their huge quantities, demanding an effective means to interactively and visually identify interesting sets for analysis.

References

[1] A. Ahmed, T. Dwyer, C. Murray, L. Song, and Y. X. Wu. Wilmascope graph visualisation. In *Proc IEEE S Info Vis*. IEEE, 2004.

[2] F. Beck, S. Koch, and D. Weiskopf. Pivotpaths: Strolling through faceted information spaces. *IEEE Trans Vis Comput Graph*, 2015. to appear.

[3] K. Börner. *Atlas of Science: Visualizing What We Know*. MIT Press, Cambridge, Mass., 2010.

[4] T. Bühler and M. Hein. Spectral clustering based on the graph p-laplacian. In *Proc ICML*, pages 81–88, New York, NY, USA, 2009. ACM.

[5] L. Byron and M. Wattenberg. Stacked graphs; geometry and aesthetics. *IEEE Trans Vis Comput Graph*, 14(6):1245–1252, Nov 2008.

[6] J. Cheeger. *A lower bound for the smallest eigenvalue of the Laplacian*, pages 195–199. Princeton University Press, June 1970.

[7] C. Chen. Citespace ii: Detecting and visualizing emerging trends and transient patterns in scientific literature. *J Am Soc Inf Sci Technol*, 57(3):359–377, 2006.

[8] C. Chen, J. Zhang, W. Zhu, and M. Vogeley. Delineating the citation impact of scientific discoveries. In *Proc ACM/IEEE Joint Conf Digit Libr*, pages 19–28. ACM, 2007.

[9] J. Choo, C. Lee, C. Reddy, and H. Park. Utopian: User-driven topic modeling based on interactive nonnegative matrix factorization. *IEEE Trans Vis Comput Graph*, 19(12):1992–2001, Dec 2013.

[10] J. Chuang, C. D. Manning, and J. Heer. "Without the Clutter of Unimportant Words": Descriptive keyphrases for text visualization. *ACM Trans Comput Hum Interact*, 19(3):19:1–19:29, Oct. 2012.

[11] C. Collins, F. B. Viégas, and M. Wattenberg. Parallel tag clouds to explore and analyze faceted text corpora. In *Proc IEEE S VAST*, 2009.

[12] I. G. Councill, C. L. Giles, and M.-Y. Kan. Parscit: An open-source crf reference string parsing package. In *Proc LREC Int Conf Lang Resour Eval*. European Language Resources Association, 2008.

[13] W. Cui, S. Liu, L. Tan, C. Shi, Y. Song, Z. Gao, H. Qu, and X. Tong. Textflow: Towards better understanding of evolving topics in text. *IEEE Trans Vis Comput Graph*, 17(12):2412–2421, 2011.

[14] M. Doerk, N. H. Riche, G. Ramos, and S. Dumais. Pivotpaths: Strolling through faceted information spaces. *IEEE Trans Vis Comput Graph*, 18(12):2709–2718, 2012.

[15] W. Dou, L. Yu, X. Wang, Z. Ma, and W. Ribarsky. Hierarchicaltopics: Visually exploring large text collections using topic hierarchies. *IEEE Trans Vis Comput Graph*, 19(12):2002–2011, Dec 2013.

[16] C. Dunne, B. Shneiderman, R. Gove, J. Klavans, and B. Dorr. Rapid understanding of scientific paper collections: integrating statistics, text analytics, and visualization. *J Am Soc Inf Sci Technol*, 2012.

[17] K. El-Arini and C. Guestrin. Beyond keyword search: Discovering relevant scientific literature. In *Proc KDD*, pages 439–447, New York, NY, USA, 2011. ACM.

[18] J.-D. Fekete, G. Grinstein, and C. Plaisant. Ieee infovis 2004 contest: The history of infovis.

[19] J.-D. Fekete and C. Plaisant. Excentric labeling: Dynamic neighborhood labeling for data visualization. In *Proc SIGCHI Conf Hum Factor Comput Syst*, pages 512–519, New York, NY, USA, 1999. ACM.

[20] A. A. Ferreira, M. A. Gonçalves, and A. H. Laender. A brief survey of automatic methods for author name disambiguation. *SIGMOD Rec*, 41(2):15–26, Aug. 2012.

[21] D. Fried and S. Kobourov. Maps of computer science. In *Proc IEEE S Pac Vis*, pages 113–120, March 2014.

[22] E. Garfield. Research fronts. *Current Contents*, oct 1994.

[23] C. Gorg, Z. Liu, J. Kihm, J. Choo, H. Park, and J. Stasko. Combining computational analyses and interactive visualization for document exploration and sensemaking in jigsaw. *IEEE Trans Vis Comput Graph*, 19(10):1646–1663, 2013.

[24] M. Harrower and C. A. Brewer. Colorbrewer. org: An online tool for selecting colour schemes for maps. *Cartogr J*, 40(1):27–37, 2003.

[25] S. Havre, E. Hetzler, P. Whitney, and L. Nowell. Themeriver: visualizing thematic changes in large document collections. *IEEE Trans Vis Comput Graph*, 8(1):9–20, Jan 2002.

[26] F. Heimerl, S. Koch, H. Bosch, and T. Ertl. Visual classifier training for text document retrieval. *IEEE Trans Vis Comput Graph*, 18(12):2839–2848, 2012.

[27] J. E. Hirsch. An index to quantify an individual's scientific research output. *Proc Natl Acad Sci U S A*, 102(46):16569–16572, Nov. 2005.

[28] W. Ke, K. Börner, and L. Viswanath. Major information visualization authors, papers and topics in the acm library. In *Proc IEEE S Info Vis*, page 216. IEEE, 2004.

[29] S. Koch, H. Bosch, M. Giereth, and T. Ertl. Iterative integration of visual insights during patent search and analysis. In *Proc IEEE S VAST*, pages 203–210, Oct 2009.

[30] B. Lee, M. Czerwinski, G. G. Robertson, and B. B. Bederson. Understanding eight years of infovis conferences using paperlens. In *Proc IEEE S Info Vis*, page 216, 2004.

[31] V. Levenshtein. Binary Codes Capable of Correcting Deletions, Insertions and Reversals. *Soviet Physics Doklady*, 10:707, 1966.

[32] M. Ley. The dblp computer science bibliography: Evolution, research issues, perspectives. In A. Laender and A. Oliveira, editors, *String Processing and Information Retrieval*, volume 2476 of *LNCS*, pages 1–10. Springer, 2002.

[33] C. D. Manning, P. Raghavan, and H. Schütze. *Introduction to Information Retrieval*. Cambridge University Press, New York, NY, USA, 2008.

[34] C. D. Manning, M. Surdeanu, J. Bauer, J. Finkel, S. J. Bethard, and D. McClosky. The Stanford CoreNLP natural language processing toolkit. In *Proc ACL*, pages 55–60, 2014.

[35] M. E. J. Newman. Coauthorship networks and patterns of scientific collaboration. In *Proc Natl Acad Sci U S A*, pages 5200–5205, 2004.

[36] D. Phan, L. Xiao, R. B. Yeh, P. Hanrahan, and T. Winograd. Flow map layout. In *Proc IEEE S Info Vis*, page 29, 2005.

[37] P. Rayson and R. Garside. Comparing corpora using frequency profiling. In *Proc Workshop Comparing Corpora*, pages 1–6, 2000.

[38] G. Salton and C. Buckley. Term-weighting approaches in automatic text retrieval. *Inf Process Manag*, 24(5):513–523, 1988.

[39] D. Shahaf, C. Guestrin, and E. Horvitz. Metro maps of science. In *Proc KDD*, pages 1122–1130, New York, NY, USA, 2012. ACM.

[40] B. Shaparenko, R. Caruana, J. Gehrke, and T. Joachims. Identifying Temporal Patterns and Key Players in Document Collections. In *IEEE ICDM Workshop on Temporal Data Mining: Algorithms, Theory and Applications*, pages 165–174, 2005.

[41] J. Stasko, J. Choo, Y. Han, M. Hu, H. Pileggi, R. Sadana, and C. D. Stolper. Citevis: Exploring conference paper citation data visually. *IEEE Conf Inf Vis*, 2013.

[42] J. Tang, J. Zhang, L. Yao, J. Li, L. Zhang, and Z. Su. Arnetminer: Extraction and mining of academic social networks. In *Proc KDD*, pages 990–998, New York, NY, USA, 2008. ACM.

[43] M. Tobiasz, P. Isenberg, and M. S. T. Carpendale. Lark: Coordinating co-located collaboration with information visualization. *IEEE Trans Vis Comput Graph*, 15(6):1065–1072, 2009.

[44] F. B. Viégas, S. Golder, and J. Donath. Visualizing email content: Portraying relationships from conversational histories. In *Proc SIGCHI Conf Hum Factor Comput Syst*, pages 979–988, New York, NY, USA, 2006. ACM.

[45] F. B. Viegas, M. Wattenberg, and J. Feinberg. Participatory visualization with wordle. *IEEE Trans Vis Comput Graph*, 15(6):1137–1144, 2009.

[46] U. Von Luxburg. A tutorial on spectral clustering. *Stat Comput*, 17(4):395–416, 2007.

[47] M. Wattenberg. Baby names, visualization, and social data analysis. In *Proc IEEE S Info Vis*, pages 1–7, Oct 2005.

[48] F. Wei, S. Liu, Y. Song, S. Pan, M. X. Zhou, W. Qian, L. Shi, L. Tan, and Q. Zhang. Tiara: a visual exploratory text analytic system. In *Proc KDD*, pages 153–162. ACM, 2010.

[49] J. D. West. *Eigenfactor: ranking and mapping scientific knowledge*. PhD thesis, University of Washington, 2010.

[50] J. A. Wise, J. J. Thomas, K. Pennock, D. Lantrip, M. Pottier, A. Schur, and V. Crow. Visualizing the non-visual: Spatial analysis and interaction with information from text documents. In *Proc IEEE S Info Vis*, Washington, DC, USA, 1995. IEEE Computer Society.

[51] P. C. Wong, E. G. Hetzler, C. Posse, M. A. Whiting, S. Havre, N. Cramer, A. R. Shah, M. Singhal, A. Turner, and J. Thomas. In-spire infovis 2004 contest entry. In *Proc IEEE S Info Vis*, volume 4, pages 51–52, 2004.

[52] Y. Wu, S. Liu, K. Yan, M. Liu, and F. Wu. Opinionflow: Visual analysis of opinion diffusion on social media. *IEEE Trans Vis Comput Graph*, 20(12):1763–1772, Dec 2014.

[53] J. Zhang, C. Chen, and J. Li. Visualizing the intellectual structure with paper-reference matrices. *IEEE Trans Vis Comput Graph*, 15(6):1153–1160, Nov 2009.