An Analysis and Visualization Tool for DBLP Data

Michael Burch, Daniel Pompe, and Daniel Weiskopf
VISUS, University of Stuttgart
Stuttgart, Germany
Email: {michael.burch, daniel.pompe, daniel.weiskopf}@visus.uni-stuttgart.de

***Abstract*—The Digital Bibliography and Library Project (DBLP) is a popular computer science bibliography website hosted at the University of Trier in Germany. It currently contains 2,722,212 computer science publications with additional information about the authors and conferences, journals, or books in which these are published. Although the database covers the majority of papers published in this field of research, it is still hard to browse the vast amount of textual data manually to find insights and correlations in it, in particular time-varying ones. This is also problematic if someone is merely interested in all papers of a specific topic and possible correlated scientific words which may hint at related papers. To close this gap, we propose an interactive tool which consists of two separate components, namely data analysis and data visualization. We show the benefits of our tool and explain how it might be used in a scenario where someone is confronted with the task of writing a state-of-the-art report on a specific topic. We illustrate how data analysis, data visualization, and the human user supported by interaction features can work together to find insights which makes typical literature search tasks faster.**

I. Introduction

We live in a data-driven era, making a data exploration task more and more challenging. This holds for nearly any discipline that has to deal with data, for example social networking, software engineering, or bioinformatics.

Also in digital bibliographies, we can see this trend towards bigger and bigger amounts of textual data. A steady increase of paper titles recorded in the DBLP can be recognized, published in the years 1930 until 2014, currently reaching the number of 2,722,212 publications.

Not only the storage of the data is a challenge but also its exploration in order to answer user tasks with it. By the growing number of research papers and the growing number of subfields of research, such tasks get more and more time-consuming. What is oftentimes missing, is an overview of such large data as a starting point for further explorations.

If, for example, a young scientist comes into a specific field of research, he is oftentimes lost in the large number of related scientific research papers and articles in the field. To get a first overview about previous and relevant work, he might use a digital bibliography such as the DBLP. Also a more experienced researcher is typically not able to remember or to be aware of all papers related to his fields of research. Consequently, a data analysis and visualization tool is required which supports at quickly getting an overview about related work. Interesting tasks for an analyst might be:

- What are all the papers with certain words or substrings in their titles?
- When were those papers published and by whom?
- What are the most frequently used words in a specific time interval/by a list of authors/occurring together with other words (hot topics)?
- What is the frequency distribution of several words or substrings over time?
- Are there correlations among a list of words or substrings and what are the extents of these correlations? Can we derive time-varying patterns?
- What are the time-dependent frequencies of word correlations?
- Are there correlations between several authors?
- What are the extents of those author correlations and are they changing over time?

Answering such tasks reliably and fast can be of great help for researchers of different levels of background knowledge in a specific field. It may be noted that this is just an excerpt of a larger list of possible tasks which we are able to answer algorithmically or visually.

In this work, we present a tool which first preprocesses the DBLP data and stores it in an efficient format in a local file system. In this step, the data is split into several data sources with additional tags and identifiers. This strategy allows us to later interactively and quickly explore the data, i.e. although the preprocessing step might take some time, it only has to be done once, except the DBLP data is updated.

The visualization component of the tool provides several visualization techniques to further visually explore the data. To this end, we support word clouds [1], [2], [3] with additional spark lines, also known as spark clouds [4]. Line graphs, time-dependent and static bar charts [5], correlation matrices for paper title words and authors, as well as a list-based overview representation are integrated in the tool.

We illustrate the usefulness of our tool by demonstrating how a researcher starts writing a state-of-the-art report on a specific topic, here on 'dynamic graph visualization' for illustrative purposes. In this scenario, we demonstrate how the single components interact and how the human user is also integrated in decision making.

II. Related Work

In 2009, Michael Ley [6] illustrated the DBLP service and described its evolution since it has been hosted online for the very first time. Although the web service is already very useful to gain insights in research articles and to browse and navigate in them, it might be more supported by visualization

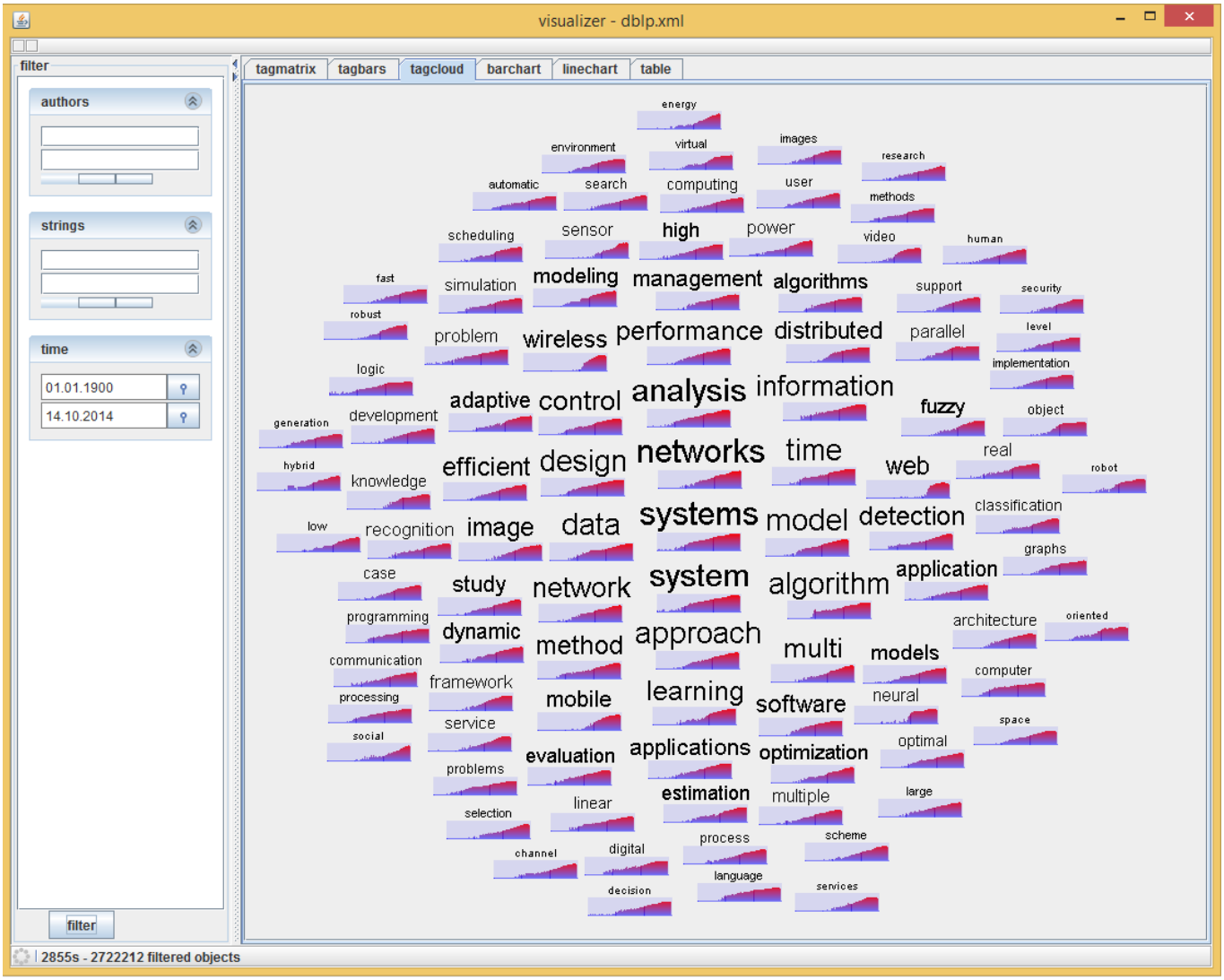


Fig. 1. The graphical user interface of our visualization tool: On the left hand side the user can make parameter settings, apply filter functions, and change color codings for example. On the right hand side the user is able to visually analyze the filtered data, select and highlight several objects, and get details on demand about the filtered scientific papers. All 2,722,212 paper titles are shown in a spark cloud for time-varying word frequencies.

techniques to accelerate the detection of useful insights in the data base which would allow to dig deeper in the data. Some support is already given by interactive visualization for this vast amount of textual data. For example, Reitz [7] presented a visualization framework for showing dynamic graphs by subdividing each link into color coded segments. His ego-centered network is useful for single entities but as the number of words, authors, and their correlations in the DBLP is large, the biggest issue of his approach is visual scalability.

Most of the works dealing with DBLP data rather prefer to extend or improve the functionality based on pure data analyses. Klink et al. [8], [9], [10], for example, analyzed social networks within bibliographical data, but they did not show the results of their analyses by sophisticated visualization techniques for such relational data. They introduce the DBL-Browser as a tool to navigate in the textual data but multiple coordinated views [11] are not integrated in the tool which are linked to show the data from different visual perspectives. In another work, Weber et al. [12] also used some visualization approaches to support the multi-layered browsing in those digital libraries, but also in this approach, visual scalability is the biggest issue.

However, visual scalability is a big challenge for today's visualization techniques since the amounts of recorded data are still growing due to the progress in hardware technology making data generation processes and data storage faster and more efficient. But the storage of the data is typically not that problematic anymore but the extraction of useful knowledge becomes a big issue which is supported by means of so-phisticated visualization techniques, their interplay, interaction techniques, algorithmic concepts such as data mining [13], and the human user as typically applied in Visual Analytics [14]. In many scenarios we still need an overview about a large part of the data which is then followed by zoom and filter interactions

and details on demand as illustrated by Shneiderman [15].

In our current work we are more interested in exploring time-varying behavior of word occurrence frequencies in paper titles as well as weighted correlations among them over time, as already researched by Burch et al. [16] using an edge splatting concept for dynamic node-link diagrams for graphs. In their approach time-dependent graphs [17] are shown in a side-by-side bipartite layout in a splatted node-link diagram but unfortunately, single word trends and correlation trends cannot be derived that easily. In our technique we use a more word-based technique such as word clouds [1], [2], [3] and adjacency matrices with attached word labels to directly show the correlations in context to individual words.

Burch et al. [18] also investigated the problem of visually representing word occurrences in form of a word cloud [1], [2], [3] in which they first computed common word prefixes to reduce the space required to display all the words. Although such a prefix word cloud seems to be a suitable concept, we cannot explore the words for weighted correlations between them. Lohmann et al. [19] described a word cloud approach investigating interactive co-occurrence highlighting of words from a microblogging context which might also be useful for DBLP data.

Also the idea of spark clouds [4] was introduced to see the most relevant words in a single overview while at the same time showing the evolution of the occurrence frequencies over time in a sparkline attached to each single word. But unfortunately, spark clouds have not been applied to bibliography data as a means to see the word evolution over time. Such time-based visualizations [20] help to uncover hot topics in specific time periods. Moreover, they can show time-varying visual patterns such as trends, countertrends, periodicities, temporal shifts, or temporal anomalies as described in the work of Burch and Weiskopf [21] applying time-series visualization to water level data.

III. DATA ANALYSIS AND TAGGING

The input data for our data analysis and visualization tool is given in an XML format. A corresponding DTD file defines the legal building blocks of the XML file which is used by our tool to transform the data into tool-specific data formats. In this paper we demonstrate how DBLP data can be transformed, analyzed, and visualized, but in general, any kind of textual data similar to DBLP data (Twitter messages, log information from software development and the like) can serve as input data for our tool.

A. Removal of Stop Words

The DBLP web page provides its raw data via an XML file which currently reached the size of several megabytes. It contains for example the paper titles with a list of authors, conference and journal titles, years, pages and the like.

Before working with the data we apply a stop word removal algorithm [22], [23], [24] to clean the data from irrelevant words and consequently, reduce the amount of textual data. Each relevant word is tagged by attaching a unique identifier number to it. We also attach a unique identifier to each of the authors in the XML file.

In our approach, we first filter the raw XML DBLP input file for all stored papers with entries for the XML tags *title*, *author*, and *timestamp*. Moreover, at least one of the tags related to relevant BibTeX information (*article, inproceedings, proceedings, book, incollection, phdthesis, masterthesis,* etc.) should be set in a data record. The corresponding paper title for each entry is transformed into a set of words by first removing special characters. We use word separation by empty spaces as indicator for splitting into single words. Stop words are filtered out from the obtained words because they do not provide extra information, but instead, make the data more noisy and need additional computing resources.

B. Computing Frequencies and Correlations

From the cleaned and tagged list of words we count the occurrence frequencies and all existing weighted correlations among the words, the authors, as well as word-author relations. The algorithm is also able to compute time-varying correlations which yield a dynamic weighted graph structure.

This can be modeled as a sequence of $n \in \mathbb{N}$ graphs

$$\Gamma := (G_1, \ldots, G_n)$$

where each individual

$$G_i := (V, E_i)$$

is an undirected graph. V expresses the set of vertices consisting of paper title words and authors. The set $E_i \subseteq V \times V$ models the weighted edges (correlations) where a weight function w is attached to each edge $e \in E_i$, i.e., a function

$$w : E_i \longrightarrow \mathbb{R}^+.$$

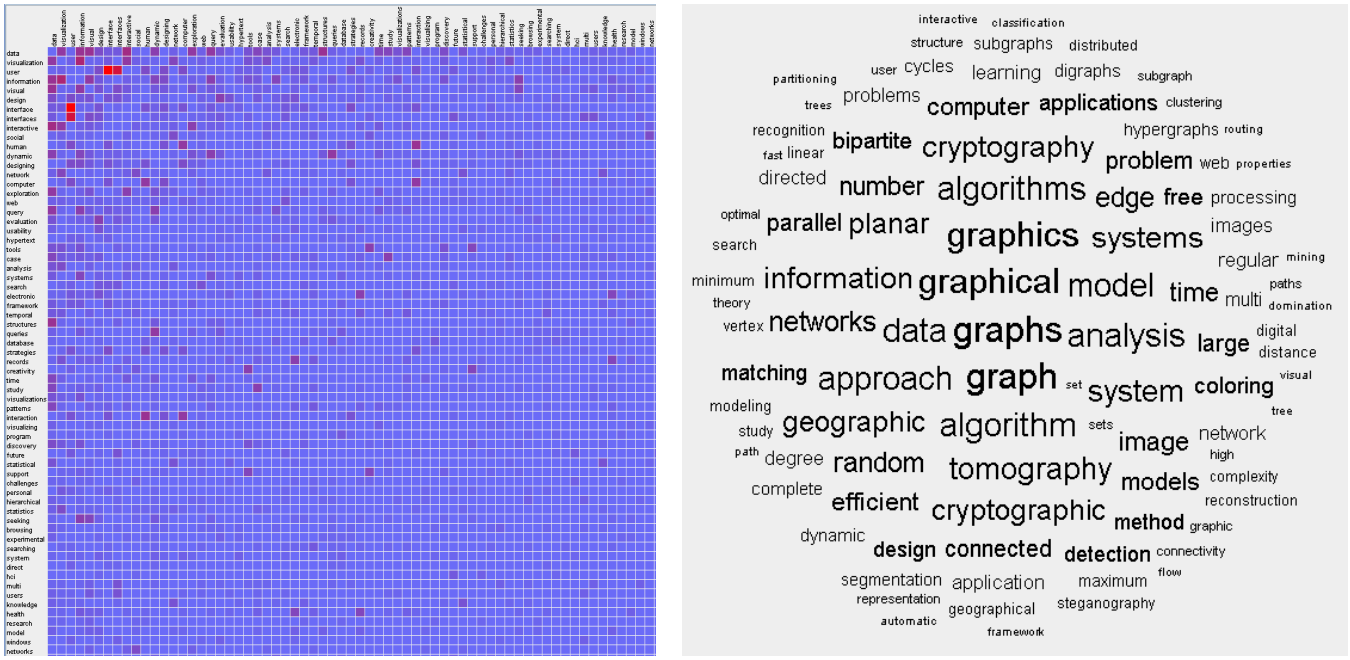
The dynamic weighted graph data is then internally stored by a list of correlation matrices allowing a fast access during user interactions with the tool. Switching from support values to confidence values for the edge weights can simply be achieved by dividing a support value by its corresponding value on the matrix diagonal.

The weighted correlations are computed by modeling each paper as a transaction, i.e., composed of the paper title and title words with the corresponding authors and co-authors. This strategy can be used to generate occurrence frequencies for the word, author, and word-author correlations, i.e., support and confidence metrics can be computed [25], [26].

IV. VISUALIZATION TECHNIQUES

After the preprocessing step, our tool is able to graphically depict the generated and tagged data (see Figure 1 for a spark cloud of the most frequently occurring words in all of the recorded paper titles). This can be done in several ways which we will describe in more detail below, see Figures 2 (a) and (b) or Figures 3 (a) and (b).

Apart from using textual descriptions of a list of selected papers, visualization is particularly useful when it comes to large amounts of textual data which can only be explored by



(a) (b)

Fig. 2. (a) Correlation matrix showing weighted word correlations as color coded matrix cells. (b) A word cloud visually depicts the occurrence frequencies of a list of words by visual features, important ones are word positions and font sizes.

browsing interaction. If the computed frequencies and correlations (static and dynamic) are visually encoded in an effective, understandable, and intuitive way, the provided visualization techniques can be used to rapidly see visual patterns, which consequently show data patterns. Finally, we get insights in the textual DBLP data.

To this end, we support an analyst by visualization techniques for showing word and author correlations as well as for showing frequency distributions of papers, words, and authors for explorations of static (summed up weights) and dynamic (time-dependent) data. To reach this goal, we use prominent visualization techniques such as a color coded correlation matrix, word clouds, time-dependent bar charts (also line graphs), and tag bars. In the following, we explain the details of each visualization technique.

A. Correlation Matrix

In this technique we visually represent correlations between a list of objects. These objects can be either authors or words. To compute the weights used to color code the matrix cells, we treat each paper as a single transaction, i.e. each correlation entry is incremented by one in the corresponding correlation matrix. Doing this process for all papers in the database leads to the unfiltered correlation matrix which contains all binary weighted correlations of words and authors together. Such a matrix is too large to be displayed completely, even in a pixel-based representation. Consequently, filtering or aggregation can be applied to decrease the data size.

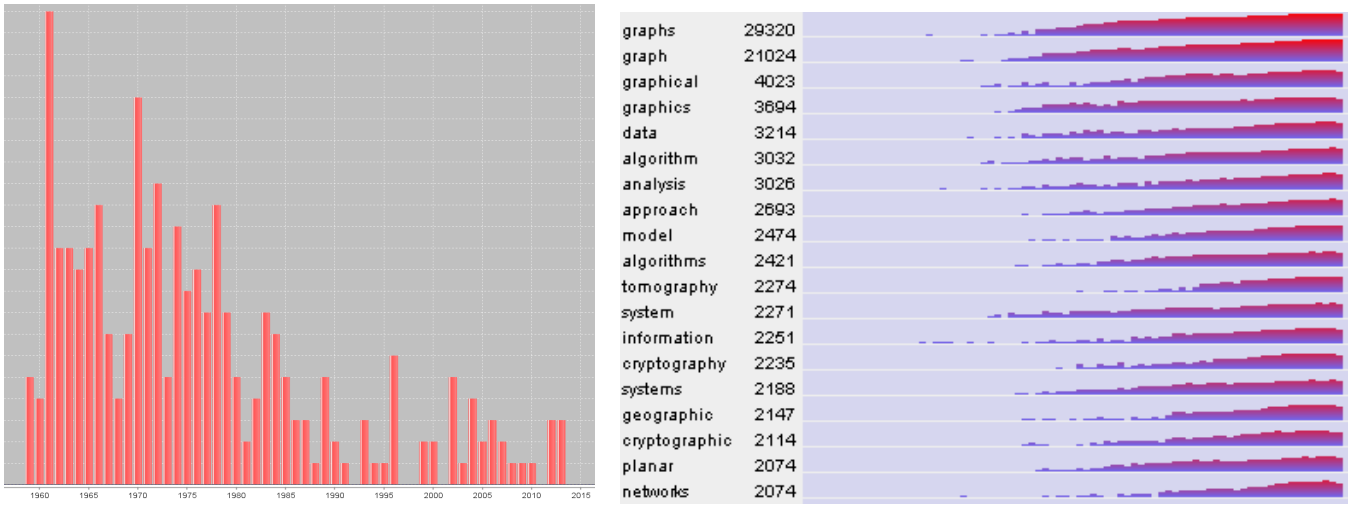
Each matrix entry is, when visualized, selectable and details on demand can be shown. Moreover, we can switch between the metrics *total number of word or author correlations* (which we refer to as support metric) and *this number divided by the total number of either one of the words or authors* (which we refer to as confidence metric). While the support matrix is symmetric, the corresponding confidence matrix does not have this property in general.

Figure 2 (a) illustrates such a correlation matrix with color coded entries showing the occurrence frequencies. We can easily see that there are many red colored matrix cells indicating strong correlations between the corresponding words. This insight can be used to see other possibly relevant words in this context which can then again be used for further analyses.

B. Word Cloud

Word clouds are very prominent representations for quickly understanding the word frequencies in a text document. Typically, they do not provide any semantic meaning between the single words, but they just serve as an overview visualization for starting further analyses. In our visualization tool we also make use of this concept and show the word occurrence frequencies by applying a spiral word cloud layout, i.e. the most frequently occurring words are located in the center of the view.

Figure 2 (b) shows a word cloud visualization for the most frequently occurring words in the analyzed DBLP data source. To generate this word cloud layout we used a spiral layout algorithm which first places the most frequent words in the center of the display. The occurrence frequencies are visually encoded in the font sizes, but additional visual variables might be added to the standard word cloud such as color coding or orientation of the single words. As demonstrated in Figure 1 also additional time-varying quantitative data can be attached



(a) (b)

Fig. 3. (a) A time-dependent bar chart shows the distribution of occurrence frequencies over time. (b) The tag bars are used as small sparkline-like graphics which can easily be attached to words or author names to display the evolving weights.

to each word showing the occurrence frequencies over time, a visualization which we refer to as spark cloud [4].

In Figure 2 (b), we can directly observe that the words 'graph', 'graphs', 'graphics', and 'graphical' occur most frequently in this data source. The reason for that is that we already filtered the complete DBLP data for the word 'graph', yielding a word cloud for all word occurrences related to the word 'graph'.

C. Time-Dependent Bar Charts

If someone is merely interested in visually exploring time-varying quantities for both single words, single authors, word correlations, or word-author correlations, he can use a standard time-dependent bar chart. Hovering the mouse over a bar shows details on demand and shows the most related words and authors to the selected word.

Figure 3 (a) shows the time-varying occurrence frequency of the word 'Algol', a popular programming language. We can see that there is a decreasing trend in papers with this word in the title. Such a plot helps to identify the global time-dependent distribution of words or authors. The bars can also be used in a stacked manner or as a ThemeRiver [27] representation.

D. Tag Bars

The tag bars, instead, are smaller graphical primitives than the traditional bar charts. Consequently, they allow to graphically represent many more timelines attached with a tag, i.e., a word or an author. They can be used as small multiples representations aligned by the time axis which supports easy time-based comparisons. If such direct aligned time-based comparisons are not required they can also be used as spark lines attached to the words in a word cloud. The tag bars are very useful to see, for example, hot topics in a certain time period, trends and countertrends, or outliers and anomalies.

Figure 3 (b) shows the time-varying correlation frequencies of words related to the word 'graph' in decreasing order. We can directly see that not all of the related words start at the same point in time. For example, the word 'tomography' starts much later but seems to be correlated. On the other hand, the word 'information' starts at the same point in time.

Our visualization tool is designed in a way that all views are linked with each other, i.e., selecting or highlighting one or more elements in one view keeps them selected or highlighted in each other view, i.e., brushing and linking interaction features are supported. It may be noted that it is easy to extend the tool by additional views since those are implemented as separate components working with the same kind of data.

V. INTERACTION TECHNIQUES

The visualization component of the tool supports several interaction techniques. Those are responding quite fast since we preprocessed and tagged the data beforehand, i.e., while interacting with the data, the tool does not have to access the raw XML data provided by the DBLP web page, but instead, the correlation data can be looked up in the corresponding weighted correlation matrices. In the following we classify the provided and most important interaction techniques:

- **Filtering:** To reduce the number of words, authors, or time steps, our tool provides filtering techniques for each of them separately and in combination.
- **Brushing and linking:** All views are linked to each other, i.e. selecting one or a group of graphical elements (brushing) leads to a highlight of all those elements which are then highlighted in all of the views simultaneously (linking).
- **Logarithmic values:** If the gap between small and large quantities is that large, a traditional linear representation by the visual variables position, length, or color coding leads to a pop-out effect of the larger quantities, but on the

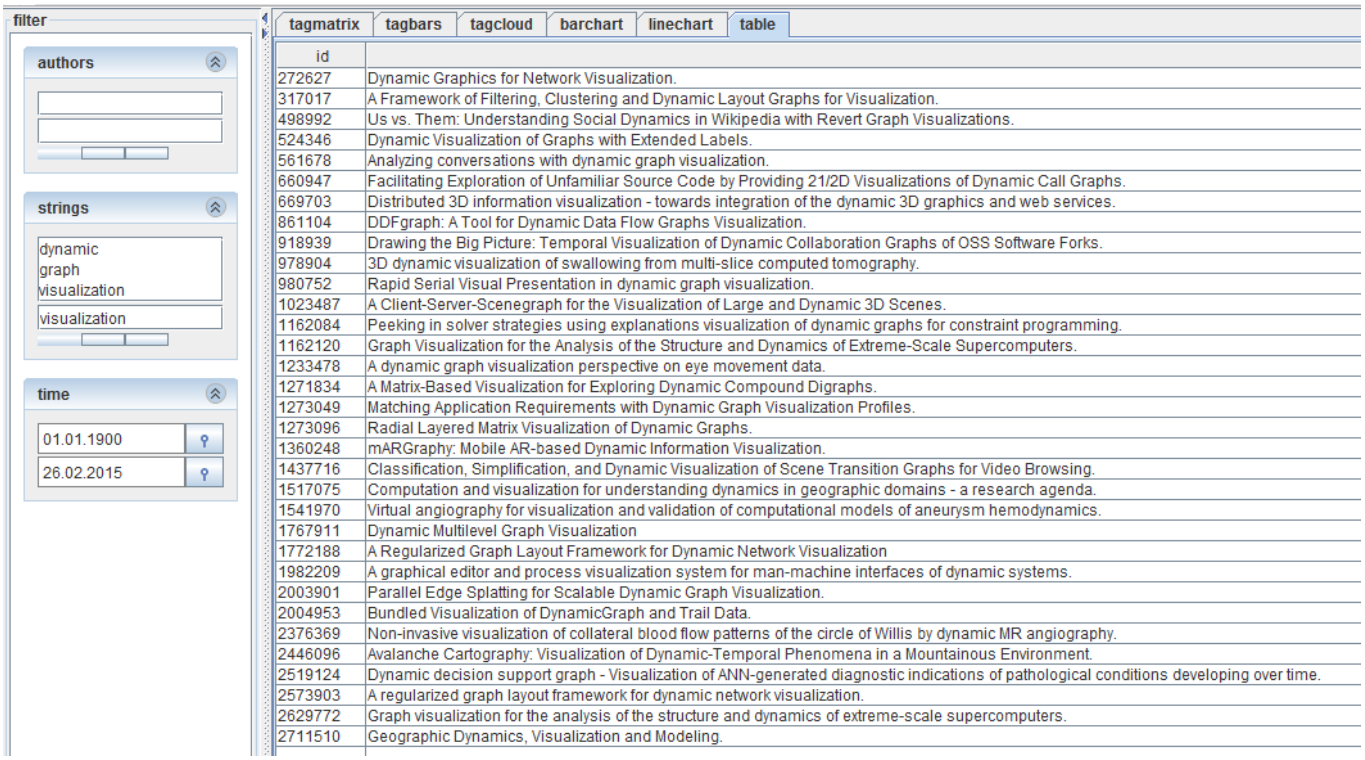


Fig. 4. A view on the filtered paper titles for the words ‘dynamic’, ‘graph’, and ‘visualization’.

negative side, also hides the small ones. Applying a logarithm function to the values before visually representing them can uncover interesting patterns.

- **Data export:** When a user has found relevant papers he can use the data export feature to store those papers. This filtered list can later be reloaded into the tool and the exploration process can be restarted from this smaller amount of textual data again.
- **Color coding:** The weighted correlations and the support and confidence metrics are of quantitative nature allowing to use different color codings for visually encoding the distribution of the values.
- **Details on demand:** Hovering the mouse cursor to a graphical primitive gives additional information, either in a separate panel or as a tool tip.

VI. APPLICATION SCENARIO

We demonstrate our DBLP data analysis and visualization tool in a scenario where we are confronted by the task of writing a state-of-the-art report on ‘dynamic graph visualization’. The very first step is to start the visualization tool which is illustrated in Figure 1. Here we can see a word cloud showing all words occurring in the complete DBLP dataset (more than 2 million paper titles). Sparklines are attached to see the evolution of the word occurrence frequencies. From this aggregated default view, we can start to do further data explorations.

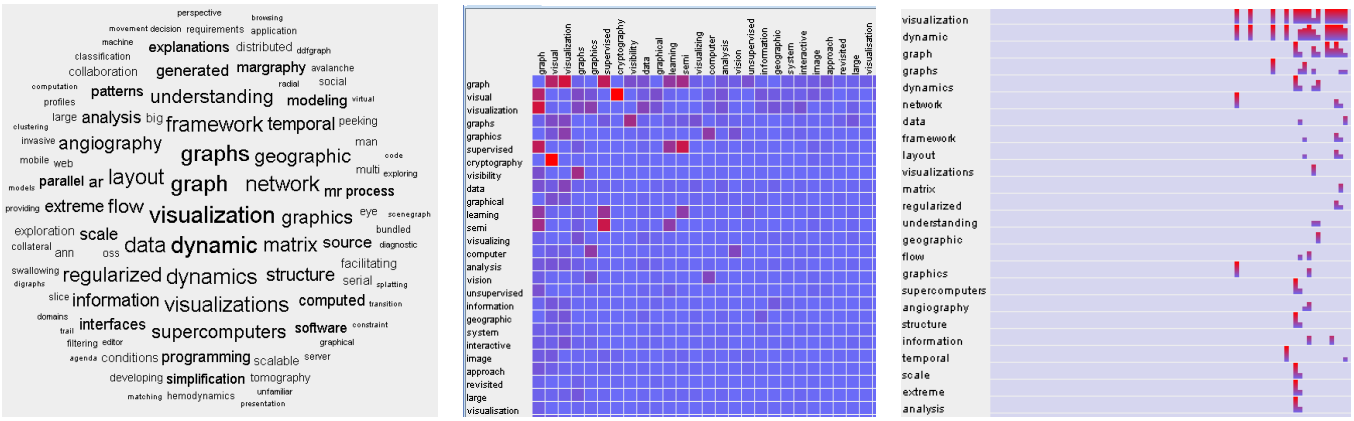
On the left hand side, we decide to filter for the words ‘dynamic’, ‘graph’, and ‘visualization’ which returns a smaller word cloud in the center view, see Figure 5 (a), but also a list of paper titles with their corresponding id numbers (see Figure 4). The data might also be filtered for specific authors or time periods as indicated in the left column in Figure 1. We can see that there is a small list of relevant papers, but we expect that this list does not contain all existing papers for this topic. Consequently, we have a look at the word correlation matrix to find possible words occurring in the same context, see Figure 5 (b). The tag bars view gives additional insights in the evolution of the word correlations over time, see Figure 5 (c). After having inspected all views we can start to extend our relevant word list by the words ‘temporal’, ‘network’, or ‘matrix’ which should be further investigated.

Another way of finding more relevant papers is to look for the authors involved in the filtered list of papers. Typically, a researcher writes several articles on a similar topic. This step gives us many more papers, but as a negative consequence, also those which are not relevant for us. To find this out, we have to scan through the paper titles, and finally, we have to decide about the relevance by reading the papers.

VII. DISCUSSION AND LIMITATIONS

Although we provide an analysis and visualization technique which is capable of dealing with more than 2 million paper titles, we are aware of the fact that there are various limitations, algorithmic and visual scalability issues, as well as possible further extensions of the current state of our tool.

- **Word semantics:** Our correlation computation approach is just based on the authors and word occurrences in



(a) (b) (c)

Fig. 5. The most related words to ‘dynamic’, ‘graph’, and ‘visualization’ as a word cloud in a spiral layout (a), as a weighted correlation matrix (b), and as a time-varying tag bar visualization (c).

the same paper title. This process does not count word correlations by involving semantics or by using larger weights for words which are closer together than others for example. Consequently, our correlation matrix can only be taken as an indication for equally weighted co-occurrences of words or authors based on those paper title occurrences which is a naive, but also helpful strategy.

- **Word synonyms:** A problematic issue in this analysis process comes from the fact that words have typically synonyms which are rarely written in the same paper title together (such as ‘graph’ and ‘network’). Consequently, words meaning similar or the same aspect do not show large correlation weights in many cases.
- **Additional text sources:** In this work, we only take into account the words contained in the paper titles. We might extend our approach to also analyze the abstract sections or also the complete text of each paper. This might help to reduce the problem with the synonyms and might give better correlation values.
- **Paper relevance:** Although our tool supports a scientist at literature search by showing weighted correlations between the words and authors, papers still have to be read to check for their suitability to a certain topic. As a positive effect, our tool can reduce the number of existing papers and serves as a starting point for further explorations.
- **Algorithmic scalability:** To enhance the interactive response of our visualization component we first preprocess the data and store it internally in specific data formats. The preprocessing step can take several minutes to generate the required data. Moreover, if the number of DBLP entries grows in future, the running time of the data preprocessing step also increases. If the tool is applied to different domains (such as Twitter messages) algorithmic scalability might be more problematic than in the DBLP scenario.
- **Visual scalability:** In many views we only show a limited number of words such as in the word clouds where only the most relevant ones are represented. Consequently, visual scalability only occurs when the viewer is interested in unaggregated data, less relevant paper titles, or correlations among a larger number of words and authors. Also the time dimension is not a problem for visual scalability since the DBLP history only covers up to a hundred of years, i.e. time steps.

VIII. CONCLUSION AND FUTURE WORK

In this paper we illustrated how visualization techniques can be used to support scientists when browsing and exploring DBLP data. To reach this goal, we first preprocessed the raw XML data by cleaning it from stop words and by tagging the words and author names with unique identifiers. As a second step, we computed weighted correlations among the words and authors by analyzing their co-occurrences in the same paper, i.e., paper title with corresponding authors. The generated correlation matrices attached with support and confidence metrics serve as input for the provided visualization techniques. We demonstrated the usefulness of the DBLP analysis and visualization tool in an application showing how a search for related work for a state-of-the-art report on a specific topic such as ‘dynamic graph visualization’ can be supported.

For future work we plan an online version of our visualization tool which may also be applied to twitter messages with the goal to find correlations among words occurring in word transactions from a different domain. The already given repertoire of visualization techniques should be extended by further techniques which should be linked. Due to the fact that our visualization component is separated from the data analysis component we can easily add source code for those added views without having to change the functionality of the data analyses. Also a user study might be of interest which can give helpful insights in how users are working with the tool, if those are really faster to explore the DBLP data, and if they do a faster literature search in the end. Tracking people’s eyes while using the tool might be important to understand where visual attention is paid over time [28].

REFERENCES

[1] F. Viégas and M. Wattenberg, "Timelines - Tag clouds and the case for vernacular visualization," *Interactions*, vol. 15, no. 4, pp. 49–52, 2008.

[2] S. Lohmann, J. Ziegler, and L. Tetzlaff, "Comparison of tag cloud layouts: Task-related performance and visual exploration," in *Proceedings of the International Conference on Human-Computer Interaction*, 2009, pp. 392 – 404.

[3] J. Sinclair and M. Cardew-Hall, "The folksonomy tag cloud: when is it useful?" *Information Science*, vol. 34, no. 1, pp. 15–29, 2008.

[4] B. Lee, N. H. Riche, A. K. Karlson, and S. Carpendale, "Sparkclouds: Visualizing trends in tag clouds," *IEEE Transactions on Visualization and Computer Graphics*, vol. 16, no. 6, pp. 1182–1189, 2010.

[5] M. Hlawatsch, F. Sadlo, M. Burch, and D. Weiskopf, "Scale-stack bar charts," *Computer Graphics Forum*, vol. 32, no. 3, pp. 181–190, 2013.

[6] M. Ley, "DBLP - some lessons learned," *Very Large Data Bases*, vol. 2, no. 2, pp. 1493–1500, 2009.

[7] F. Reitz, "A framework for an ego-centered and time-aware visualization of relations in arbitrary data repositories," *CoRR*, 2010.

[8] S. Klink, M. Ley, E. Rabbidge, P. Reuther, B. Walter, and A. Weber, "Visualising and mining digital bibliographic data," in *Proceedings of INFORMATIK*, 2004, pp. 193–197.

[9] ——, "Browsing and visualizing digital bibliographic data," in *Proceedings of the Symposium on Visualization (VisSym)*, 2004, pp. 237–242.

[10] S. Klink, P. Reuther, A. Weber, B. Walter, and M. Ley, "Analysing social networks within bibliographical data," in *Proceedings of 17th International Conference on Database and Expert Systems Applications*, 2006, pp. 234–243.

[11] J. Roberts, "State of the art: Coordinated & multiple views in exploratory visualization," in *Proceedings of Fifth International Conference on Coordinated and Multiple Views in Exploratory Visualization*, 2007, pp. 61–71.

[12] A. Weber, P. Reuther, B. Walter, M. Ley, and S. Klink, "Multi-layered browsing and visualisation for digital libraries," in *Proceedings of the 10th European Conference on Research and Advanced Technology for Digital Libraries*, 2006, pp. 520–523.

[13] D. A. Keim, F. Mansmann, D. Oelke, and H. Ziegler, "Visual analytics: Combining automated discovery with interactive visualizations," in *Proceedings of 11th International Conference of Discovery Science*, 2008, pp. 2–14.

[14] D. A. Keim, F. Mansmann, J. Schneidewind, J. Thomas, and H. Ziegler, "Visual analytics: Scope and challenges," in *Visual Data Mining - Theory, Techniques and Tools for Visual Analytics*, 2008, pp. 76–90.

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

[16] M. Burch, C. Vehlow, F. Beck, S. Diehl, and D. Weiskopf, "Parallel edge splatting for scalable dynamic graph visualization," *IEEE Transactions on Visualization and Computer Graphics*, vol. 17, no. 12, pp. 2344–2353, 2011.

[17] F. Beck, M. Burch, S. Diehl, and D. Weiskopf, "The state of the art in visualizing dynamic graphs," in *Proceedings of State-of-the-Art Reports of EuroVis*, 2014.

[18] M. Burch, S. Lohmann, D. Pompe, and D. Weiskopf, "Prefix tag clouds," in *Proceedings of 17th International Conference on Information Visualisation*, 2013, pp. 45–50.

[19] S. Lohmann, M. Burch, H. Schmauder, and D. Weiskopf, "Visual analysis of microblog content using time-varying co-occurrence highlighting in tag clouds," in *Proceedings of the International Working Conference on Advanced Visual Interfaces*, 2012, pp. 753–756.

[20] W. Aigner, S. Miksch, H. Schumann, and C. Tominski, *Visualization of Time-Oriented Data*, ser. Human-Computer Interaction. Springer, 2011.

[21] M. Burch and D. Weiskopf, "Visualizing dynamic quantitative data in hierarchies - TimeEdgeTrees: Attaching dynamic weights to tree edges," in *Proceedings of the International Conference on Information Visualization Theory and Applications*, 2011, pp. 177–186.

[22] J. Savoy, "englishst. website. url http://members.unine.ch/jacques-.savoy/clef/englishst.txt," 2013.

[23] ——, "germanst, website, url http://members.unine.ch/jacques-.savoy/clef/germanst.txt," 2013.

[24] ——, "IR multilingual resources at unine. website. url http://members-.unine.ch/jacques.savoy/clef," 2013.

[25] M. Burch, S. Diehl, and P. Weißgerber, "EPOSee - A tool for visualizing software evolution," in *Proceedings of the 3rd International Workshop on Visualizing Software for Understanding and Analysis*, 2005, pp. 127–128.

[26] ——, "Visual data mining in software archives," in *Proceedings of the ACM 2005 Symposium on Software Visualization*, 2005, pp. 37–46.

[27] S. Havre, E. G. Hetzler, and L. T. Nowell, "Themeriver: Visualizing theme changes over time," in *Proceedings of the IEEE Symposium on Information Visualization*, 2000, pp. 115–123.

[28] K. Kurzhals, B. D. Fisher, M. Burch, and D. Weiskopf, "Evaluating visual analytics with eye tracking," in *Proceedings of the Fifth Workshop on Beyond Time and Errors: Novel Evaluation Methods for Visualization, BELIV*, 2014, pp. 61–69.