



Yang Wang · Minzhu Yu · Guihua Shan · Han-Wei Shen · Zhonghua Lu

VISPubComPAS: a comparative analytical system for visualization publication data

Received: 7 July 2019 / Accepted: 18 July 2019 / Published online: 10 August 2019
© The Visualization Society of Japan 2019

Abstract For an unfamiliar field, researchers who are looking for interdisciplinary collaboration or students who are going to start their research career often need to look for top research affiliations and domain experts according to the publication of top conferences or journals in this field. Further comparative analysis of affiliations or experts with similar achievements is also needed in order to find suitable collaborators or supervisors. In this work, we provide comprehensive visual analysis of research affiliations and domain experts based on papers accepted by the IEEE VIS from 1990 to 2018. First, we extract multi-word keywords from title and abstract automatically and then extract topics using LDA model based on these keywords. Second, we extract relationship between authors and affiliations based on co-author analysis. Third, we design and implement VISPubComPAS, a requirement-driven analysis system to (1) help users discover top affiliations and experts of required keywords; (2) visualize the relationships and statistics of these affiliations and experts; (3) compare two selected affiliations or experts of interest in detail by visualization. Finally, we conduct use cases and user reviews to demonstrate the effectiveness of VISPubComPAS.

Keywords Comparison · Scientific literature · Topic extraction · Visual analytics system · Publication analysis · Author analysis

1 Introduction

At present, more and more scientific research collaborations take place in different fields. An expert often knows his main research field very well, but he may not know much about other fields. For example, an astronomer is very good at studying the evolution of the universe, but he needs visualization experts or institutions to visualize his research results. However, he does not know much about visualization, nor do he know any experts or institutions of visualization. Faced with this situation, there are generally two ways: recommended by friends or searched on the Internet. Both methods have disadvantages: the first is limited to the social circle of his friends while the experts or institutions sought by the second way are often with more commercial advertising and their professionalism may not be good as well.

Electronic supplementary material The online version of this article (<https://doi.org/10.1007/s12650-019-00585-2>) contains supplementary material, which is available to authorized users.

Y. Wang · M. Yu · G. Shan (✉) · Z. Lu
Computer Network Information Center, Chinese Academy of Sciences, Beijing, China
E-mail: sgh@sccas.cn

H.-W. Shen
The Ohio State University, Columbus, USA

A scientific way to find collaboration experts or institutions is to analyze top experts and institutions from publications of top conferences or journals in this field, and then find the most suitable potential collaboration institutions or experts through analysis and comparison. The same situation occurs to students who are planning to start their careers in an unfamiliar field. They need to choose suitable universities and supervisors for their research directions. They need to know which universities are in the leading position in this field, which experts are outstanding in these universities, and then properly make their choice.

One problem in seeking domain experts and institutions of unfamiliar fields is that seekers may not know the proper keywords for searching. Thus, a set of keywords and topics should be extracted from the publications and recommended to seekers. A topic can be interpreted with a group of keywords. To extract topics in a field, we need to get keywords first. Obviously, author-supplied keyword in a paper is a convenient source. However, a paper doesn't always have author-supplied keywords, especially in early publications (Isenberg et al. 2017). And some authors point out that sometimes the author-supplied keywords do not represent the topics of their paper very well. In order to solve these problems, one effective way is to extract keywords from the title, abstract or even the full text of a paper. Meanwhile, single-word keywords often have ambiguity. For instance, the word network may refer to social network or neural network. Therefore, it is also necessary to extract phrases as keywords instead of individual word. Then, topics can be extracted through keyword clustering automatically, although most existing works extract them manually.

Another problem that are most concerned in seeking top experts for collaboration or instruction is that both researchers and students are prefer to choose the experts who are still active in research of the field. Some domain experts may have retired or changed their research interests. Some may have only begun their research of this field in recent years. So the details of the variation of research interests of experts are also required by the seekers.

There are several systems or tools to search domain experts or institutions through publications, such as csranks.org. However, these systems basically have the same shortcomings: firstly, most of the search results are described in texts, and lack of graphical visualization, which the users cannot see at a glance. Secondly, users can only browse the detail information of single experts or institutions one by one through the search result. Different experts or institutions cannot be compared together. Thirdly, the research interests of authors are always provided as a whole, and no detailed variations of research interests are provided.

From above, we can see that seeking institutes and experts for collaboration or choosing universities and supervisors for studying through publications have three similar requirements:

R1 Recommending a set of topics and keywords to users. To achieve the first step, we need to extract the key phrases and topics from titles, abstracts, author-provided keywords, etc. Then we mark authors and affiliations with these topics and key phrases as their research interests. Thus the users can search the corresponding authors and institutions more accurate.

R2 Displaying the variation of research interests of authors. Based on keyword and topic extraction, we gather statistics of publications for authors and their affiliations in each research interest, and thus we can display the variation of research interests of authors. The historical publication data and the recently published data must be counted separately in order to emphasis the recent activeness of authors.

R3 Comparing similar counterparts. Similar counterparts mean experts or institutes of similar publication statistics. These counterparts should be put together for comparison to help users make proper choice.

Our work is focused on the three requirements above. We extract key phrases as keywords from titles, abstracts, author-supplied keywords, IEEE keywords and indexing of INSPEC using natural language processing methods. Based on these keywords, topics are extracted by LDA model (Blei et al. 2003; Sievert and Shirley 2014). We have counted the history publications of authors and their affiliations as well as publications in the past 5 years for each topic and keyword. We also extract the relationships between authors and affiliations. We have designed an exploratory author-affiliation diagram to display these statistics and relationships. A comparative and expendable research interest variation graph is also designed to enable users to compare different experts and institutions by their annual publications on topics or keywords. Based on these work, we develop a visual analytic system VISPubComPAS to provide insights into experts and institutions in visualization based on IEEE VIS publications. The system has been evaluated by real user cases. We summarize our contributions as follows:

- Key phrases extracted from titles, abstracts and author-supplied keyword and topics generated with LDA.
- An author-affiliation diagram to show the information of top authors and affiliations of IEEE VIS conference.
- An expendable research interest variation graph to compare authors or affiliations.
- A comparative visual analytical system for visualization publication data.

The paper is organized as follows. Section 2 discusses related works. Section 3 presents methods for extracting keywords and clustering topics as well as the indicators we use for statistics. Section 4 shows our visualization and interaction designs. Section 5 describes the studies of user cases by using our visual analytics system VISPubComPAS. Real user feedbacks are also discussed in this section. We conclude by discussing design implications and open challenges in Sect. 6.

2 Related work

In this section, we discuss related work on analyzing scientific literature.

Chuang et al.'s (2012, 2013) machine learning tool for topic model diagnostics (Görg et al. 2013), visual text analysis using Jigsaw, and the CiteVis tool (Stasko et al. 2013) based on a dataset of visualization publications (Isenberg 2015). Latif and Beck (2018) developed an interactive visual system of publication records using template-based text generation and graph embedding to generate author profiles based on VIS publications. Guo and Laidlaw (2018) analyzed topic-based sense making frameworks and experiments with an iteratively designed visual analysis tool to identify design implications for facilitating research idea generation using visualizations. Isenberg et al. (2017, Stasko et al. 2013 and Isenberg et al. 2017) presented the results of a comprehensive multi-pass analysis of visualization paper keywords supplied by authors for their papers accepted by the IEEE Visualization conference series (now called IEEE VIS) between 2000 and 2015. Wu et al. (2015) developed egoSlider for researchers to explore the ego-network evolutionary histories of prestigious researchers in the fields of computer graphics, computer vision and visualization based on the DBLP collaboration network data, but there is no information of affiliations. Matejka et al. (2012) studied the relationships between research publications through their use of citations and developed an interactive visualization system Citeology. Wang et al. (2018) developed ImpactVis to show the research impact of authors based on their publications and citations. Maguire et al. (2016) addressed the challenge of visualizing publication impact through a visualization design that can exist in three states: impact graph, impact glyph and impact overview. Liu et al. (2018) studied co-occurrence relationships between the concepts of visualization and text mining using hierarchical visualization. None of the above works have visually discussed the relationships between authors and affiliations, and users cannot conveniently compare similar counterparts to find the difference between them. In our work, we bend our efforts for these two problems. We design author-affiliation diagram to visualize the relationships between authors and affiliations and research interest variation graph to compare research counterparts.

Apart from these efforts made by visualization community, there exist two accessible Web sites that offer information about top experts and institutions in different areas. Clarivate has reported highly cited researchers (HCR) 2018 (<https://hcr.clarivate.com>), demonstrated by production of multiple highly cited papers that rank in the top 1% by citations for field and year in Web of Science. It lists HCRs with information including last name, first name, primary institution, primary region and secondary affiliation. CSRankings (<http://csranks.org>) lists top computer science institutions in different areas around the world, followed by researchers working in each institution. It offers links to researchers' homepage, Google Scholar page and publication profile. They provide valuable information to help find collaboration experts and institutions, but it's not easy to explore the relationships between experts and institutions, let alone analyze and compare their research interests and research influence. In our work, we improve textual information of experts and institutions to visualization graphs and focus on the comparison of similar entities, which greatly improves the efficiency for users exploring the information of experts and institutions or analyzing the differences between similar entities.

3 Data processing

Our statistics and analysis are based on topics, and each topic includes several keywords. We extract key phrases as keywords from title, abstract, author-supplied keywords, IEEE keywords and indexing of INSPEC. It is because not all papers have keywords. Then, we extract topics from these extracted keywords using LDA model. Finally, we filter target affiliations and authors by statistics of keywords and topics and do further analysis.

3.1 Data source

We collect 3067 full papers accepted by IEEE VIS from 1990 to 2018. Each paper contains title, authors, published year, conference, abstract, citations, etc. Part of these papers have author-supplied keywords; some have IEEE keywords and indexing of INSPEC. These metadata is integrated from vispubdata (Isenberg et al. 2017), IEEE VIS official Web site, IEEE Xplore and Microsoft Academic (Sinha et al. 2015).

3.2 Keywords extraction and topic extraction

We design an approach to automatically extract keywords. First, we combine each paper's title, abstract, author-supplied keywords, IEEE keywords and indexing of INSPEC together as a document; clean them by unifying words to lowercase, removing special characters and lemmatization. Second, we use a set of natural language processing methods such as tokenization and tagging part of speech of words to extract phrases based on n-gram language model. Then we filter key phrases by document frequency and co-occurrence frequency. We drop phrases with less than 5 document frequency and make sure each selected key phrase has more than 2 co-occurrence frequency with any other selected key phrase. Finally, we get 1799 key phrases as our keywords. Then we delete words except keywords in each document, and feed these documents to LDA model, which is provided by genism library. We specify 6 topics to generate. Finally, we obtain each paper's topic distribution on 6 topics and each topic's keyword distribution on 1799 keywords.

3.3 Data selection

3.3.1 Topic selection

After having each paper's topic distribution, we need to assign most probable topics for each paper. We rank each paper's topics from high probability to low probability and delete topics that satisfy the following conditions:

$$(1) \quad p_i < \frac{1}{N}, i \in \{1, 2, 3, 4, 5, 6\}$$

$$(2) \quad p_i < \frac{p_{i-1}}{2}, i \in \{2, 3, 4, 5, 6\},$$

N is the number of topics and p_i denotes the probability of i_{th} topic of each paper.

3.3.2 Keyword selection

With each topic's keyword distribution, we select 30 most salient keywords for each topic according to the following formula:

$$s(w, k) = \log \left(\frac{\varphi_{kw}}{p_w} \right)$$

$s(w, k)$ is the saliency of keyword $w \in \{1, \dots, V\}$ to topic $k \in \{1, \dots, K\}$, where V denotes the number of keywords in document ($V = 1799$ in this paper), K denotes the number of topics ($K = 6$ in this paper), φ_{kw} denotes the probability of keyword w for topic k , and p_w denotes the marginal probability of keyword w in the corpus. The top five keywords for each topic are shown in Table 1.

Table 1 Top five most salient keywords for each topic

Topic	Keywords
1	Visual analysis, time series, decision making, interactive visual analysis, text visualization...
2	Social network, parallel coordinate, visualization design, graph visualization, graph layout...
3	Flow visualization, vector field, computational fluid dynamic, surface extraction, flow field...
4	Volume rendering, transfer function, graphic hardware, direct volume rendering, ray tracing...
5	Medical visualization, surface model, computerized tomography, polygonal model...
6	Finite element, geographic information system, wavelet transform, bar chart, information space...

3.3.3 Affiliation and author selection

There are hundreds of affiliations and thousands of authors extracted from IEEE VIS papers. Visualizing all these in a screen is impossible and unnecessary. According to requirement analysis, we only analyze affiliations and authors that satisfy following conditions:

$$(1) \quad R_p \leq T_1 \times N, R_{p_latest} \leq T_2 \times N; T_1, T_2 \in (0, 1]$$

$$(2) \quad R_c \leq T_3 \times N, R_{c_latest} \leq T_4 \times N; T_3, T_4 \in (0, 1]$$

R_p denotes affiliation or author's rank of history paper counts; R_{p_latest} denotes affiliation or author's rank of last 5 year's paper counts; R_c denotes affiliation or author's rank of history paper citations; R_{c_latest} denotes affiliation or author's rank of last 5 year's paper citations; N is the number of total affiliations or authors; and T_1, T_2, T_3, T_4 are threshold. When selecting affiliations, we set $T_1 = T_2 = T_3 = T_4 = 0.1$, while selecting authors we set $T_1 = T_2 = T_3 = T_4 = 0.03$. Finally, we obtain 112 affiliations and 136 authors.

4 Visualization design

We design a visualization system named VISPubComPAS to fulfill the requirements mentioned in Sect. 1. The system overview is shown in Fig. 1. It mainly consists of five panels (Fig. 1a–e). We introduce the detailed design of each panel in the following.

4.1 Keyword search selector

The keyword search Selector, shown in Fig. 1a, allows users to select keywords they want to search. The options of the selector are the keywords which are extracted by the methods of Sect. 3. Considering that users may not be familiar with a field, we provide users with the function of fuzzy matching. When inputting a keyword, keywords similar to the input are listed (Fig. 2a). Users can thus select multiple similar keywords to search. The selected keywords are listed in the text box below the selector.

4.2 Topic selector

Topic selector in Fig. 1b shows the topics which are extracted from LDA model, also shown in Fig. 2b. We list top four keywords of most frequency in each topic as topic interpretation. Each topic is coded with a unique color and functions as legend of Fig. 1c. Users can quickly and conveniently select keywords of one topic by clicking each legend, rather than inputting keywords through keyword search selectors. When users have selected keywords, the authors and their affiliations associated with the keywords will be visualized in the author-affiliation diagram.

4.3 Author-affiliation diagram

Figure 1c shows the author-affiliation diagram which shows the authors and affiliations related to the selected keywords. The diagram consists of two main columns connected with links. The left column is the affiliation column which lists the top institutes, and the right column is the author column which lists the top authors. The ranks of these two lists are based on selected option in Fig. 3a. There are four sort options for users to choose: by number of papers, by number of citations, by number of papers in last 5 years, by number of citations in last 5 years.

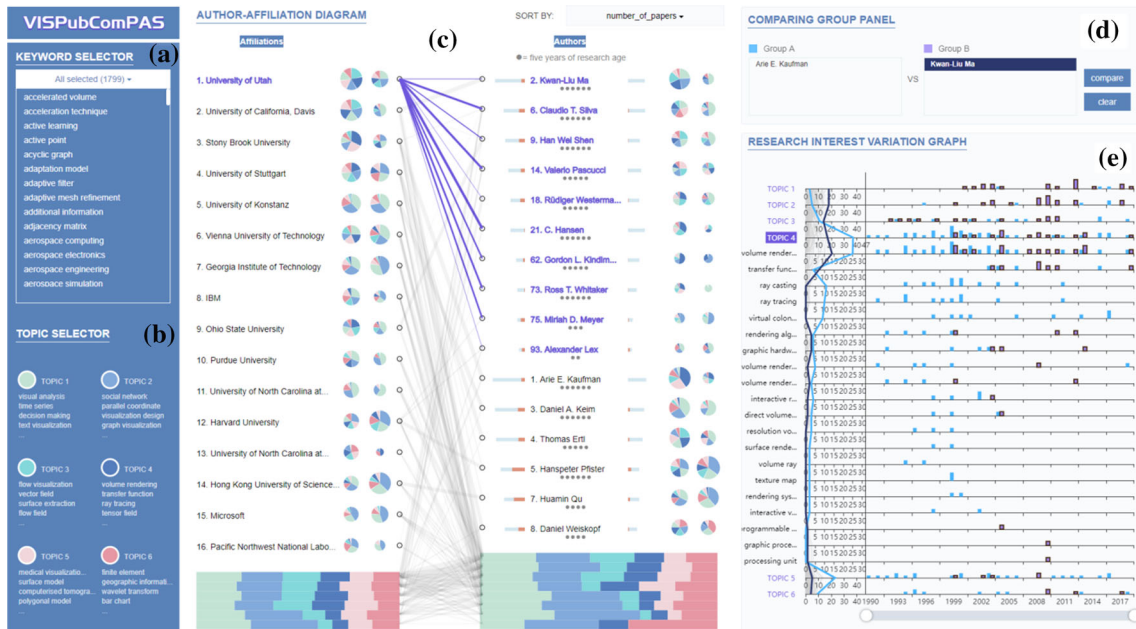


Fig. 1 Interface of the VISPubComPAS system for analyzing and comparing the authors and affiliations of publications in IEEE visualization conference:—**a** keyword search selector, for users to input and select keywords for search; **b** topic legends, showing the topics we extracted from LDA and their main representative keywords; **c** author-affiliation diagram, showing the statistical information and relations of top authors and affiliations for the selected topic/keyword; **d** comparing group panel, showing the authors or affiliations selected by users for further comparison; **e** research interest variation graph, showing the detailed research topic variation of the selected authors or affiliations by their annual publications

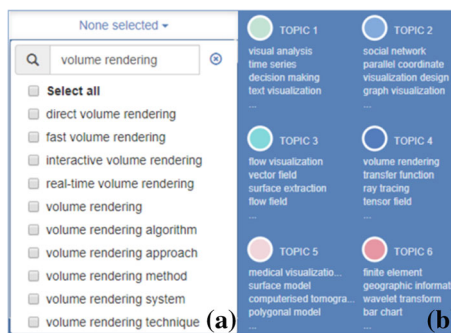


Fig. 2 **a** keyword search selector with fuzzy matching and **b** selectable topic legends with representative keywords

In the affiliation column, each institute item is designed as Fig. 3b. The ordinal number in front of institute name indicates the rank of the institute under current sort option. The first pie chart behind the name depicts the topic distribution of all publications of the institute, and the second one depicts the topic distribution of publications of the institute during last 5 years. Color encoding is consistent with topic legend. The encoding of pie chart size depends on selected sort option. If current sort option is “by number of papers” or “by number of papers in last 5 years,” the sizes are encoded with publication numbers. Otherwise, the sizes represent citation numbers. The hollow ring at the end of an item is the anchor point for links.

In the author column, each author item is designed as Fig. 3c. The ordinal number in front of author name indicates the rank of the author under current sort option. The bar on the left of an item indicates the number of all publications of the author, and the orange area in it indicates the number of papers published in last 5 years by the author. The bar on the right of an item indicates the number of citations gained by all publications of the author, and the orange area in it indicates the number of citations gained by papers published in last 5 years by the author. Small gray dots represent the length of an author’s research career

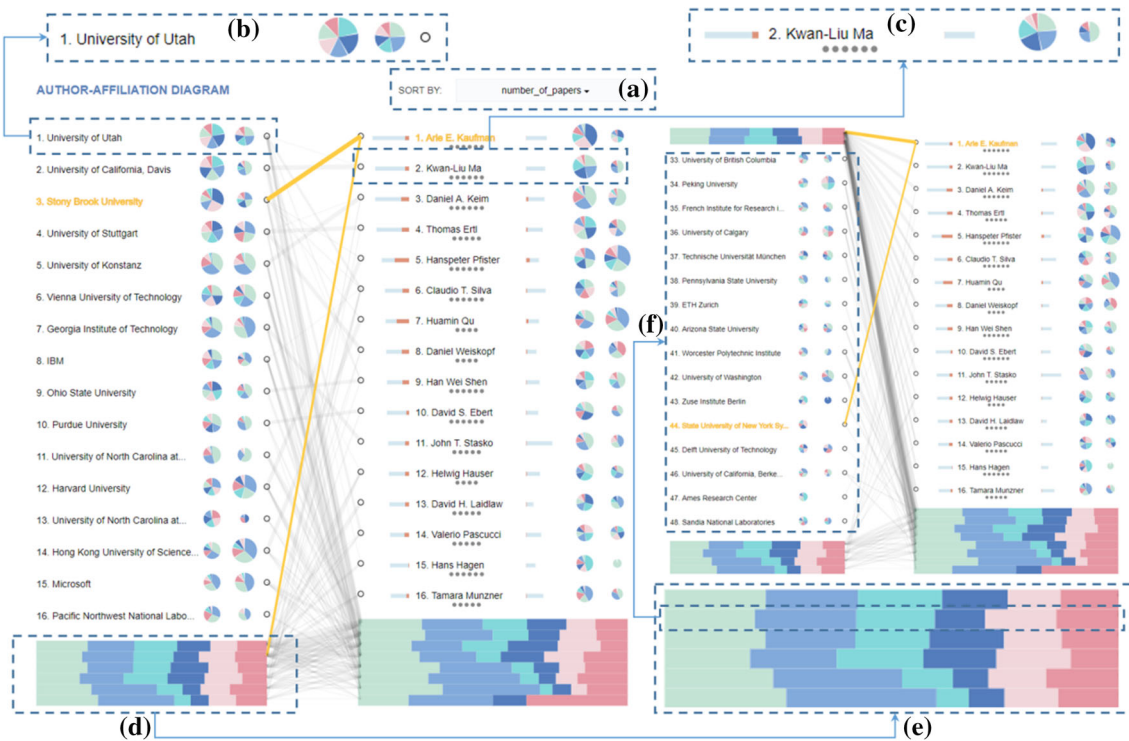


Fig. 3 Author-affiliation diagram. **a** Sort option selector, selecting the factor for ranking authors and affiliations. **b** An affiliation item, consisting of an ordinal number, affiliation name, 2 pie charts and a link anchor point. **c** An author item, consisting of ordinal number, author name, 2 pie charts, 2 bars, dots of research years, and a link anchor point. **d** Expendable color folders: hiding extra items to avoid overcrowding. Each folder (e) hides 20 items and can be expanded to (f)

(from the first publication year to the present). One dot stands for 5 years. Therefore, authors with 6 dots mean that they started their research 25–30 years ago. Author with only one dot indicates that they have just begun their research in the last 5 years. The couple of pie charts in the item are encoded the same as those in the affiliation column. The hollow ring at the front of an item is the anchor point for links.

Because of over a hundred of authors and affiliations, it would be too crowd to list all of them in the columns. Therefore, the number of listed items (N_{items}) is designed to be able to dynamically change according to the size of screen. In Fig. 3, only 16 are listed at one time for each column. Besides, we design expendable color folders to hide the rest; each folder stands for N_{items} items and can be expanded when clicked. As shown in Fig. 3e–f, each folder stands for 16 items and the folded 16 items can be expanded when clicked. The area occupied by each color represents the topic proportion of the folded N_{items} authors or institutions.

A link from an author item to an affiliation item indicates the author has publications during his work in this affiliation. The line width refers to the number of papers published by the author in this affiliation.

When users hover on an item (source item) in columns, the related links and items on the other column are highlighted. Sometimes, some related items may be folded into color folders. By clicking on the source item, all related items are repositioned to the top to facilitate reading information, as shown in Fig. 4. Users can drag any item of interest into the comparing group panel for further comparison.

4.4 Comparing group panel

Figure 1d shows the comparing group panel with two group box into which we can drop interested authors and affiliations for further comparison. When we drop an affiliation into either group box, the name of target box will be replaced by the affiliation name and the number of authors attached to this affiliation, as shown in Fig. 7a. Also, the target box will list all names of authors who worked or are working in this affiliation. When we click the “compare” button, the results of comparison will be displayed in research interest variation graph. Unique colors are also assigned to both groups.

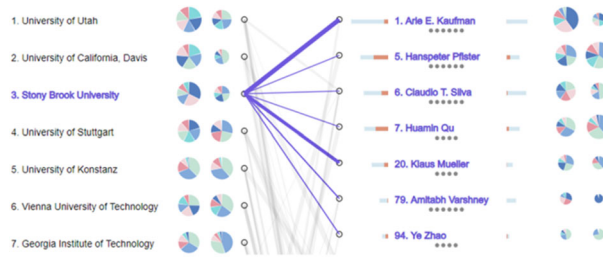


Fig. 4 By clicking on an item of a column, the items linked to it on the other column will be repositioned to the top and highlighted

4.5 Research interest variation graph

Figure 1e shows the research interest variation graph which shows the detail variation of research interests over the years of both groups selected by users in the comparing group panel, also shown in Fig. 5.

The research interest variation graph consists of two parts (Fig. 5a, b). In Fig. 5a, we design an expendable parallel coordinate which shows the overall topic distribution of an author’s publications. Each axis corresponds to a topic. Each line represents an author. The blue lines represent authors of group A, and the purple lines represent authors of group B. The gray lines behind are authors outside these two groups. Thus users can easily know the positions of authors in groups A and B among all authors. The right part, Fig. 5b, contains multiple bar charts, each corresponding to a topic. The x axis of each bar chart is the publication year ranging from 1990 to 2018, and the y axis is the number of publications. Each bar represents the annual publication number for a topic of the corresponding group. We add a time ranger for users to select time range which they are interested in by dragging the handlers at the head or end of the ranger, shown in Fig. 5c. By double-clicking on the topic label, the variation graphs of detailed keywords are expended from the topic, as shown in Fig. 5d. Thus, users can easily see what specific research spots a group of authors has contributed to and how their research interests transfer over the years. Furthermore,

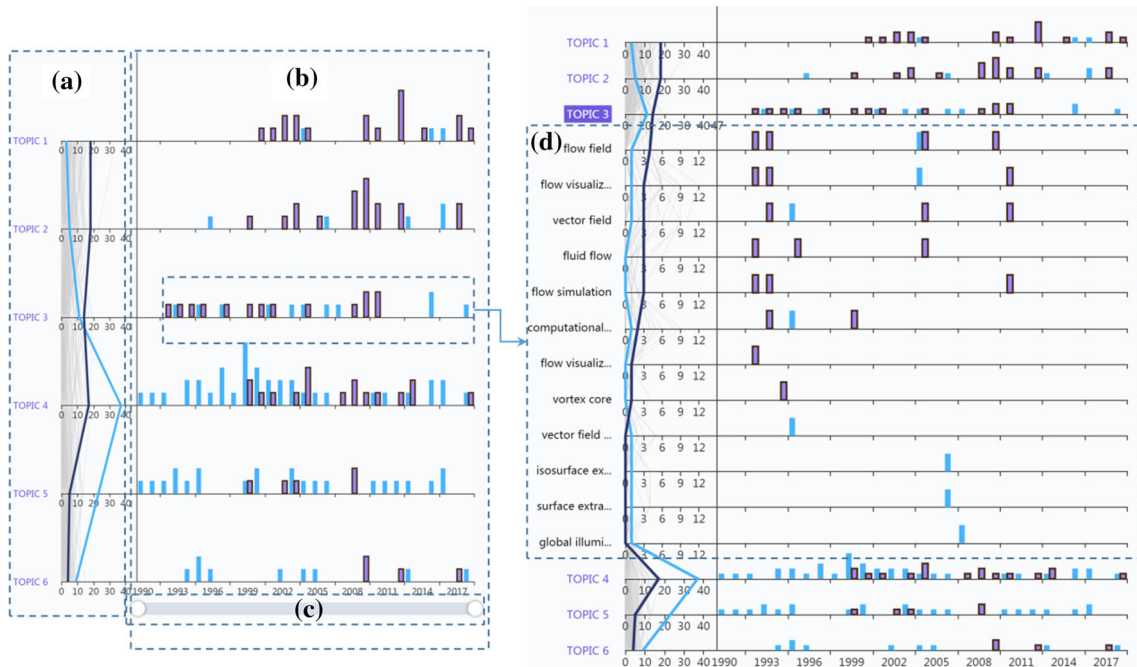


Fig. 5 Research interest variation graph, showing the detail variation of research interests over the years of both groups selected by users in the comparing group panel. **a** Expendable parallel coordinate, showing the overall topic distribution of an author’s publications. **b** Bar charts showing the topics involved in annual publications of each group. **c** Time ranger, selecting time range which users interested in. Each axis of (a) can be expanded to (d) to reveal the variation of detail keywords of the topic corresponding to the axis

when users hover on an author in comparing group panel, the corresponding line and bars of the author are highlighted by filling or stroking in black.

5 Case study

In our experiment, we have introduced our system to several researchers and students who are unfamiliar with visualization. They use our system to search and compare the experts and institutions of visualization techniques they are interested in and finally help them to seek out the proper ones. Here we take “flow visualization” as an example to demonstrate how our system works.

5.1 Searching “flow visualization”

We type the word “flow visualization” in the keyword search selector and select the related keywords. Here we select 5 keywords: flow visualization, flow visualization system, flow visualization techniques, flow volume and unsteady flow visualization. The top authors and affiliations in flow visualization are listed in the author-affiliation diagram, sorted by number of papers as default, shown in Fig. 6a.

From Fig. 6a, we can see that the second pie charts of Ohio State University (OSU), Otto-von-Guericke University Magdeburg (OVGU) and Peking University (PKU) have larger size than those of other institutions in the top 10, which means that these three universities are still very active in visualization research. So we choose these three universities for further comparison.

5.2 Comparison of institutes

We conduct two comparisons. One is OSU and OVGU because they are close in rank of total publication numbers. The other is OVGU and PKU because they are close in rank of recent publication numbers.

We drag OSU and OVGU to the comparing group panel, as shown in Fig. 7a. The results are shown in the research interest variation graph (Fig. 7b). From Fig. 7b, we can see that OSU has started its publications on IEEE VIS since 1995, while OVGU published its first paper on IEEE VIS late at 2006. OSU’s early research topics focused on topics 3, 4, 5, and since 2005 their focus has shifted gradually to topics 1, 2. OVGU’s research mainly focused on topics 3 and 5, and its researches on these topics have been very active in recent years. Since flow visualization is the main component of topic 3, we expend topic 3, as shown in Fig. 7c.

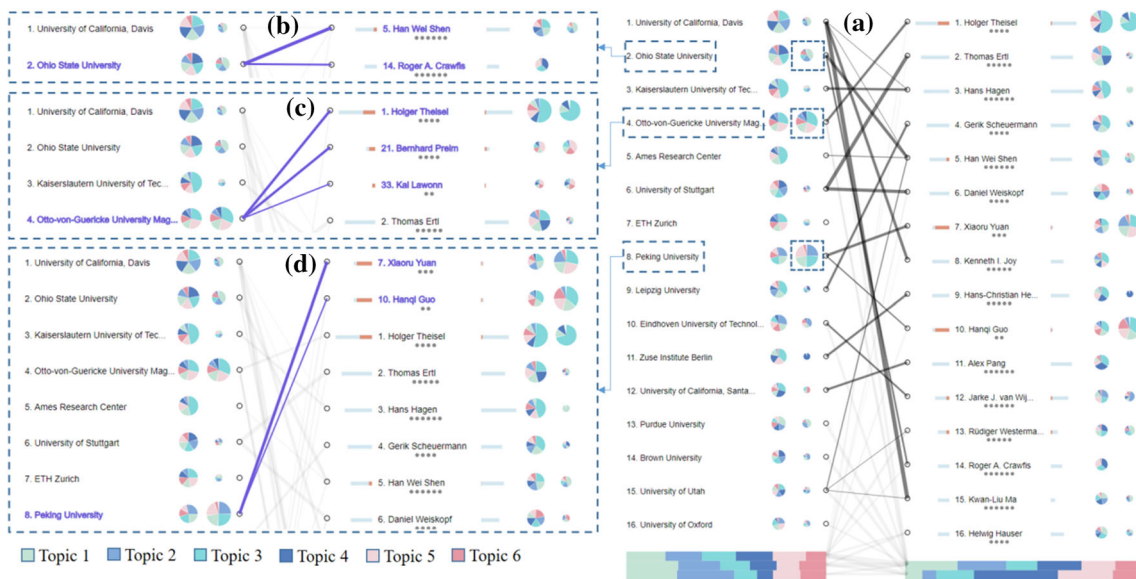


Fig. 6 The author-affiliation diagram of flow visualization. **a** Ranking authors and affiliations by number of papers, OSU, OVGU and PKU are chosen for further analysis. **b** Author items of OSU. **c** Author items of OVGU. **d** Author items of PKU

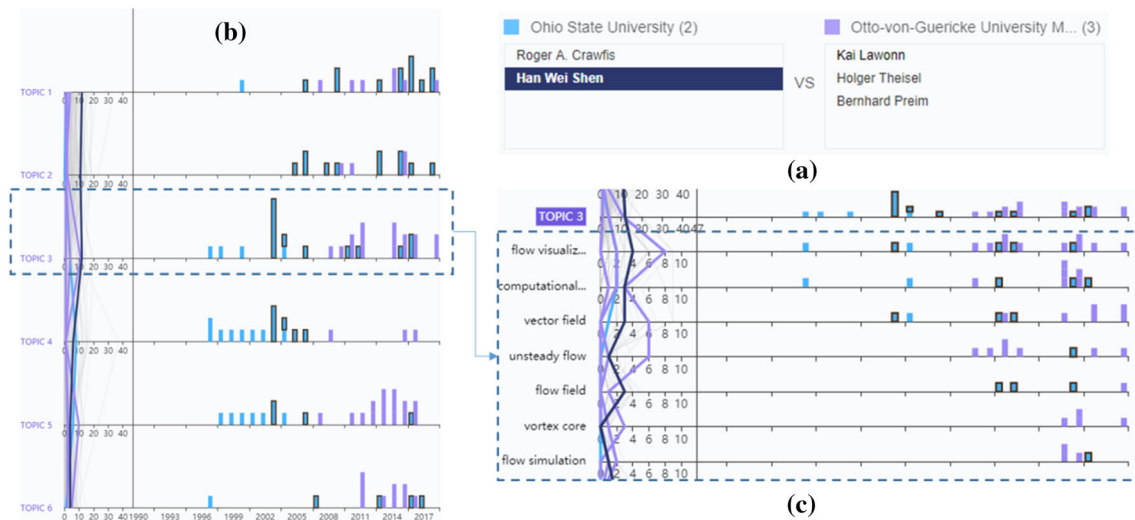


Fig. 7 Comparison of Ohio State University and Otto-von-Guericke University Magdeburg. **a** Listing the authors of two universities with an author hovered. **b** The variation of topics in annual publications of each university, black line in parallel coordinate and bars with black borders are data of the hovered author. **c** The variation of main keywords in topic 3 of the two universities

From Fig. 7c, we can see that the main research spots on topic 3 of OSU and OVGU include flow visualization, computational fluid dynamic, vector field, unsteady flow, flow field, vortex core, flow simulation, etc. From the parallel coordinate shown in Fig. 7b, we can see that Han Wei Shen (Highlighted with black) of OSU is the most productive expert in topic 3 among the authors of these two universities and he has also the most publications in topics 1, 2.

The same analytical method can be used to compare OVGU and PKU. We drag OVGU and PKU to the comparing group panel, as shown in Fig. 8a. The results are shown in the research interest variation graph (Fig. 8b). From Fig. 8b, we can see that both OVGU and PKU began publishing papers on IEEE VIS around 2006. OVGU's research mainly focused on topic 3 and topic 5, and PKU mainly on topics 1, 2, 3. Both universities have been very active on researches in recent years. We expand topic 3, shown in Fig. 8c, and we can see that the main research spots on topic 3 of OVGU and PKU are almost the same as those of OSU and OVGU. From the parallel coordinate shown in Fig. 8b, we can see that Holger Theisel (highlighted with black) of OVGU is the most productive expert in topic 3 among the authors of these two universities and he is very concentrated on the topic because he has few publications on other topics. In Fig. 8d, we can see that Xiaoru Yuan of PKU is the second most productive expert in topic 3 among the authors of these two universities and he also has the most publications in topic 1 and topic 2.

5.3 Comparison of authors

In the author-affiliation chart, as shown in Fig. 6a, we click on each university, and the authors attached to it are highlighted and repositioned at the top of the author list, shown in Fig. 6b–d.

In Fig. 6b, we can see that OSU has two main researchers working on flow visualization, Han Wei Shen and Roger A. Crawfis and Han Wei Shen contributes more publications and citations because both his bars are longer than his colleague. Han Wei Shen still have several publications in recent years for the orange area in left bar can be observed. While his recent work haven't gain much citations yet for the orange area in right bar is small. Roger A. Crawfis doesn't publish any papers in recent years, and the orange areas in both bars are invisible. Both researchers have been publishing papers at the IEEE VIS conference since at least 25 years ago, because they both have 6 dots. From the pie charts, we can see that Han Wei Shen has published papers on various subjects because the colors of his first pie chart are evenly distributed. In recent years, most of his publications are focused on topic 1 because the corresponding color area is much larger than the others. The main research topics of Roger A. Crawfis are topics 3, 4, 5 and he doesn't have a second pie due to no publication in recent years. Based on the analysis above, we can see that Han Wei Shen is more suitable as a potential collaborator or supervisor in OSU on flow visualization.

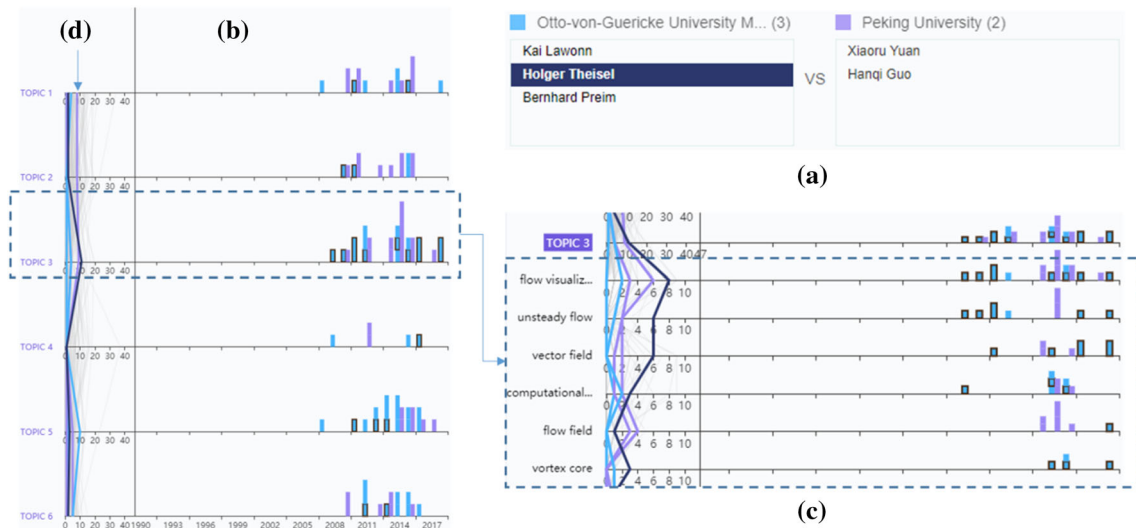


Fig. 8 Comparison of Otto-von-Guericke University Magdeburg and Peking University. **a** Listing the authors of two universities with an author hovered. **b** The variation of topics in annual publications of each university, black line in parallel coordinate and bars with black borders are data of the hovered author. **c** The variation of main keywords in topic 3 of the two universities

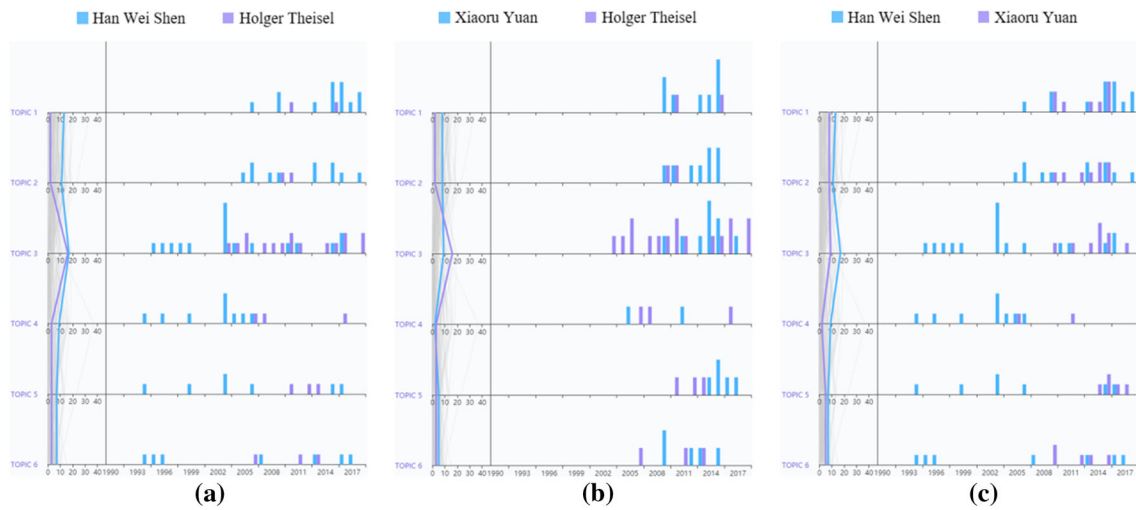


Fig. 9 Comparison of three authors **a** showing comparison of Han Wei Shen and Holger Theisel. **b** Comparison of Xiaoru Yuan and Holger Theisel. **c** Comparison of Han Wei Shen and Xiaoru Yuan

Similarly, we can analyze the authors of OVGU and PKU in Fig. 6c, d and may find that Holger Theisel and Xiaoru Yuan are the most suitable potential collaborators or supervisors in their own universities on flow visualization. The differences among them can be illustrated in Fig. 9.

From Fig. 9, we can see that Han Wei Shen has a much longer research age than the other two peers, and he has gradually transferred research interests from topics 3, 4, 5 to topics 1, 2, 3 since 2003. Xiaoru Yuan’s research interests mainly focus on topics 1, 2, 3 and recently he began to devote more interest to topic 5. Holger Theisel is much more concentrated on topic 3 than Han Wei Shen and Xiaoru Yuan. His publications on topic 3 are much more than the other two authors.

5.4 User feedback

In order to evaluate the usefulness and effectiveness of our system VISPubComPAS, we invited several users whose main research field is not visualization to test our system, including students, teachers and professors. Here we list two of these feedbacks.

Feedback 1 “The system divides the research direction of visualization into thousands of keywords. By combining these keywords with fuzzy search and dynamic combination, users can easily select the direction of interest. Through different ranking methods, users can analyze the research trends and hotspots of institutions and authors. Expendable color folders can retain basic distribution information and hide insignificant information. By dragging and grouping, users can compare and analyze the research trends of the two groups. The ability to expand views and display domain sub-topics facilitates fine-grained trend analysis. The overall design of the system is concise, efficient and functional. After clustering, the topic classification is reasonable and representative. Author-affiliation diagram can analyze research hotspots from two dimensions: overall and detailed. The comparison view can further help users to analyze the variation of research interests and sub-directions of the two groups in detail, and help users find collaborators or track research hotspots in the field.”

Feedback 2 “Through author-affiliation diagram, it’s easy to discover the most productive experts, the most influential experts and the most active experts. The same to institutions. By interaction, I can gain insights about working relationship and schoolmate relationship between experts. Even more, I can discover who has been working in a university after graduation while who has work experiences in both university and enterprise. After quick exploration, I can compare interested institutions and experts by dragging them into compare boxes. It’s really convenient. The research interest variation graph provides details about their strengths and revolution of interest. However, the system doesn’t provide enough information in other fields which may help me gain more comprehensive understanding about interested institutions and experts.”

Feedback 3 “The system is convenient and efficient. Also, it has a more beautiful and friendlier user interface than CSRankings. I have been looking for such a system that combine comparing with searching for many years and finally I get it here. I hope this system can be launched to serve more researchers and can extend to more research fields. Besides, it will be perfect if provides more information such as author’s publications, collaboration relationships of experts and institutions.”

6 Conclusion

Our work is aimed to meet the requirements of researchers seeking interdisciplinary collaborations or students who are going to start their research on a new field seeking supervisor based on the information of publications. We have studied the publications of IEEE VIS conference from 1990 to 2018. We have extracted keywords from the titles, abstracts and author-provided keywords and clustering them into topics by LDA. We have gathered statistics of authors and their affiliations. Based on these work, we have designed author-affiliation diagram and expendable research interest variation graph to support the visualization of the statistics, relationships and comparisons for authors and affiliations. Finally, we have developed an interactive visual analytic system and apply it on real user cases. The feedbacks of users have proved the usefulness and effectiveness of our system.

In the future, there is still much work to be done to improve our system. First, lots of information of authors and affiliations still remain to be mined, such as the collaboration relationships between affiliations or authors based on co-authorship. Second, the topic extraction algorithm has to be improved. The usage of LDA to extract topics needs to be improved since it is possible that the extracted topics are not the topics of users’ interest and some of the extracted topic may not be 100% correct. One solution is to refine our LDA model, adding interactive function to fine-tune the topic extraction. Another way might be using other NLP model, for example, Word2vec or BERT. Also, in this paper, we only extracted 6 topics through LDA, which can only summarize visualization papers at a coarse level. Users are expecting more detailed extraction since there are many different topics in the domain of visualization. Hierarchical topic models such as hLDA and hierarchical visualization are need. Third, our system only supports comparison of two groups, we shall expand the number of groups in the future, and thus more rational visualization designs are needed for multiple groups. Last but not least, the comparison of authors and affiliations is far more complex than what we have done. The evaluation results vary due to different practical requirements. More reasonable evaluation indexes need to be established. Combining with bibliometrics will be a good research direction for the problem.

Acknowledgements This work was supported by the Strategic Priority Research Program of the Chinese Academy of Sciences, Grant No. XDA19080102, the 13th Five-year Informatization Plan of Chinese Academy of Sciences, Grant No. XXH13504 and the Key Research Program of Frontier Sciences, CAS, Grant No. QYZDB-SSW-SMC004-02.

References

- Blei DM, Ng AY, Jordan MI (2003) Latent dirichlet allocation. *J Mach Learn Res* 3(Jan):993–1022
- Chuang J, Manning CD, Heer J (2012) Termite: visualization techniques for assessing textual topic models. In: *Proceedings of the international working conference on advanced visual interfaces*. ACM
- Chuang J et al. (2013) Topic model diagnostics: assessing domain relevance via topical alignment. In: *International conference on machine learning*
- Görg C, Liu Z, Kihm J et al (2013) Combining computational analyses and interactive visualization for document exploration and sensemaking in jigsaw. *IEEE Trans Visual Comput Graph* 19(10):1646–1663
- Guo H, Laidlaw DH (2018) Topic-based exploration and embedded visualizations for research idea generation. In: *IEEE transactions on visualization and computer graphics*
- Isenberg P et al. (2015) Visualization publication dataset. <http://vispubdata.org/>. Published June 2015
- Isenberg P, Isenberg T, Sedlmair M et al (2017a) Visualization as seen through its research paper keywords[J]. *IEEE Trans Visual Comput Graph* 23(1):771–780
- Isenberg P, Heimerl F, Koch S et al (2017b) vispubdata. org: a metadata collection about iee visualization (vis) publications. *IEEE Trans Vis Comput Graph* 23(9):2199–2206
- Latif S, Beck F (2018) VIS author profiles: interactive descriptions of publication records combining text and visualization. In: *IEEE transactions on visualization and computer graphics*, pp. 1–1
- Liu S et al. (2018) Bridging text visualization and mining: a task-driven survey. In: *IEEE Trans Vis Comput Graph*
- Maguire E, Montull JM, Louppe G (2016) Visualization of publication impact. arXiv preprint: [arXiv:1605.06242](https://arxiv.org/abs/1605.06242)
- Matejka J, Grossman T, Fitzmaurice G (2012) Citeology: visualizing paper genealogy. In: *CHI'12 extended abstracts on human factors in computing systems*. ACM, pp. 181–190
- Sievert C, Shirley K (2014) LDAvis: a method for visualizing and interpreting topics. In: *Proceedings of the workshop on interactive language learning, visualization, and interfaces*, pp. 63–70
- Sinha A, Shen Z, Song Y, Ma H, Eide D, Hsu BJ, Wang K (2015) An overview of microsoft academic service (MA) and applications. In: *Proceedings of the 24th international conference on world wide web (WWW '15 Companion)*. ACM, New York, NY, USA, pp. 243–246
- Stasko J, Choo J, Han Y et al. (2013) Citevis: exploring conference paper citation data visually. In: *Posters of IEEE InfoVis, 2013*
- van Raan AFJ (2004) Sleeping beauties in science. *Scientometrics* 59(3):467–472
- Wang Y, Shi C, Li L, Tong H, Qu H (2018) Visualizing research impact through citation data. *ACM Trans Interact Intell Syst (TiIS)* 8(1):5
- Wu Y, Pitipornvivat N, Zhao J, Yang S, Huang G, Qu H (2015) egoSlider: visual analysis of egocentric network evolution. *IEEE Trans Vis Comput Graph* 22(1):260–269