Galex: Exploring the Evolution and Intersection of Disciplines

Zeyu Li, Changhong Zhang, Shichao Jia, and Jiawan Zhang, *Senior Member, IEEE*

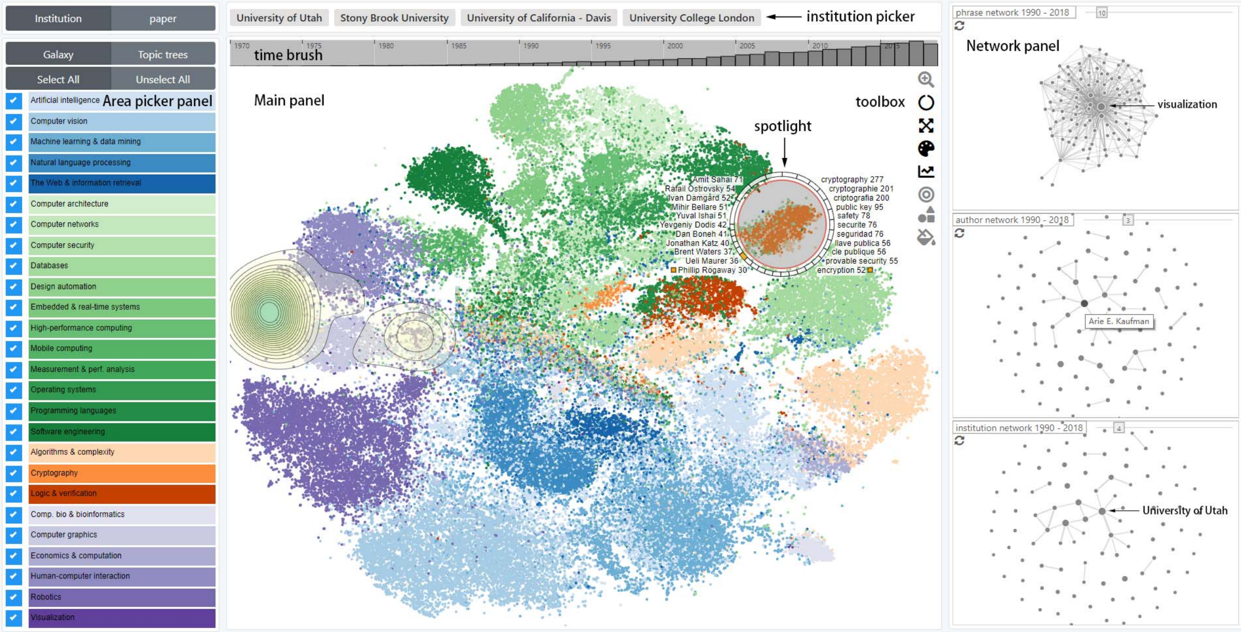


Fig. 1: Interface of Galex consists of three panels: area picker panel (left), main panel (middle), and network panel (right). The main panel includes three layers: discipline, area, and institution layers. To date, all papers (over 86,000) from all 26 areas are shown in the discipline layer, forming an overview (galaxy) of computer science. The area picker controls the areas that are displayed. A contour line indicates the paper distribution of one area (here, visualization). The spotlight reveals the semantics of the regions of interest. The time brush acts as a time filter. The toolbox provides useful components, such as pallet, spotlight, and time slice selector. The network panel includes three kinds of networks to uncover the structure of one area.

Abstract—Revealing the evolution of science and the intersections among its sub-fields is extremely important to understand the characteristics of disciplines, discover new topics, and predict the future. The current work focuses on either building the skeleton of science, lacking interaction, detailed exploration and interpretation or on the lower topic level, missing high-level macro-perspective. To fill this gap, we design and implement Galaxy Evolution Explorer (Galex), a hierarchical visual analysis system, in combination with advanced text mining technologies, that could help analysts to comprehend the evolution and intersection of one discipline rapidly. We divide Galex into three progressively fine-grained levels: discipline, area, and institution levels. The combination of interactions enables analysts to explore an arbitrary piece of history and an arbitrary part of the knowledge space of one discipline. Using a flexible spotlight component, analysts could freely select and quickly understand an exploration region. A tree metaphor allows analysts to perceive the expansion, decline, and intersection of topics intuitively. A synchronous spotlight interaction aids in comparing research contents among institutions easily. Three cases demonstrate the effectiveness of our system.

Index Terms—Science evolution, science mapping, interdisciplinary, knowledge domain visualization, visual analysis

1 Introduction

Given the decreased difficulty of collecting vast scientific literature data, mining the information contained in literature metadata has become an important research topic [21]. Researchers have implemented numerous interesting applications, such as identifying emerging topics [55], determining whether to recruit someone [35], predicting article's cited times [62], and investigating the factors that contribute to academic success [52].

As a powerful tool to gain insights from abstract data, visualization is widely used in various fields to explore science from multiple aspects. Researchers in bibliometrics and physics revealed the skeleton and composition of science by creating enormous citation networks or collaboration networks with different levels of data aggregation [3, 4, 33, 49]. However, the created static graphs lack text information and interactivity, causing difficulty in comprehending and obtaining valuable details. Several works only described either the evolution of topics [17, 37, 41] or semantic relevance among topics [38, 40, 46]. Several research combined both [8, 31, 39] but failed to illustrate the structure of science or to integrate multiple entities. Insufficient exploration perspective, simplex data description hierarchy, and incomplete key entities limit

- *Z. Li, C. Zhang, S. Jia and J. Zhang are with College of Intelligence and Computing, Tianjin University. E-mail: {lzytianda, ch_zhang, jsc, jwzhang}@tju.edu.cn.*
- *Corresponding author J. Zhang is with Tianjin cultural heritage conservation and inheritance engineering technology center and Key Research Center for Surface Monitoring and Analysis of Relics, State Administration of Cultural Heritage, China.*

Manuscript received 31 Mar. 2019; accepted 1 Aug. 2019.
Date of publication 16 Aug. 2019; date of current version 20 Oct. 2019.
For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org, and reference the Digital Object Identifier below.
Digital Object Identifier no. 10.1109/TVCG.2019.2934667

the usage of these works in a wide range of real-world scenarios.

Considering the selection of promising scientific fields, investors might not only focus on the evolution of hotspots from the viewpoint of time but also assess the influence and potential of the field of interest by inspecting its relationship to other fields from the perspective of relevance. Ph.D. candidates who want to select appropriate research topics need to investigate from a macro-background perspective at the discipline or area level and focus on details at the topic level. Different analysts pay attention to various entities. College freshmen may want to select supervisors by looking for top scholars in a particular field, whereas the deans may pay attention to seeking potential partner institutions to promote the internationalization of their colleges. A comprehensive, well-designed, and easy-to-use system is required to support the above potential applications.

In this paper, we design and implement an effective and integrated visual analysis system called Galaxy Evolution Explorer (Galex), the same name as a space telescope launched by NASA which aims at studying the formation history of galaxies and stars[1]. In a similar sense, with its rich interactions, our Galex is dedicated to enabling analysts, such as decision-makers, professors, and students, to rapidly explore multiple facets of one discipline. We offer three exploration perspectives: detecting the evolution of one discipline from the perspective of time, identifying the intersection of discipline sub-fields from the perspective of textual relevance, and investigating the skeleton of a discipline from the perspective of entity co-occurrence. The data description hierarchy contains three progressively refined levels: discipline level describes an overview of the discipline and the intersections among its sub-fields, area level reveals the origin, growth, fusion, separation, and decline of topics, and institution level allows analysts to compare the academic performance and research topics among institutions as well as checking the popularity of phrases. A variety of entities are involved in Galex, including phrases, institutions, and authors. We also design a variety of interactive components, such as spotlight, topic tree, time slice creator, and pallet, which allow analysts to apply their own experience to excavate hidden knowledge.

In this paper, we use computer science as an example disciplinary to demonstrate the effectiveness of our method. The data (authors, institutions, and title for each paper) originate from CSRankings[2], which includes about 86 thousand papers published in 75 well-selected top conferences in computer science from 1970 to 2018. CSRankings classifies these conferences into 4 categories and 26 areas (Fig. 3). In addition, we collect missing but important text information, namely, abstract and keywords, from Scopus[3]. In fact, our method is not limited to a specific discipline, it can be applied to any discipline.

Our main contributions in this paper are summarized as follows:

- Design and implementation of a hierarchical and integrated visual analysis framework that allows analysts to understand a scientific field from macro- to micro-perspective;
- Design of a flexible time brush and a context-aware spotlight, they empower analysts to quickly grasp the semantics of region of interest in any time slice. We also design a synchronous spotlight which enables the comparison of sets of document collections;
- Design of a topic tree metaphor which enables analysts to review topic evolution intuitively and to find representative papers of each topic and potential inter-topical papers[4] expediently;
- Provision of a set of networks revealing the backbone of a field, as a complementary to the structure captured by text data.

2 Related Work

2.1 Science mapping

The idea of making science visible has attracted the attention of scholars from various fields. These individuals mainly aim to uncover the science structures and sub-field relationships. Scholars from scientometrics are interested in creating science maps by drawing vast citation networks of papers/journals, cooperation networks of scholars/institutions and co-occurrence networks of words [2, 43, 60]. Rosvall et al. [49] constructed a two-level structure of science via a random-walk community detection algorithm, which revealed knowledge flows between scientific communities. However, these efforts were over-reliant on citation data and neglected intuitive and interpretative text data. Skupin et al. [53] and Fried et al. [22] produced science maps that look almost like real-world maps, where countries represent clusters. The galaxy view from IN-SPIRE system [34] used the same metaphor as ours to project documents into a 2D space. ThemeView [64] and VxInsight [15] created 3D landscapes of topic spaces, in which the height of mountains is proportional to the word/document quantity of topics. Untangle [38] provided a connected triangle web which roughly describes an overview of computer sciencewhere by locating relevant computer science conferences nearby. ContexTour [39] superimposed a contour-map over entity networks which enhances community representation compared to node-link graph. It offered the same context for different entities to prevent from losing context. Similar to our discipline layer, all the above works can provide an overview of vast document collections or entities. The difference is that these studies pay more attention on how to create an overview by algorithms, whereas we are more interested in how to interpret the overview by various interactions.

2.2 Visual Exploration of the Intersections among Scientific Fields

Cross-domains hold infinite potential for innovative research. Utilizing the metaphor of coin, Oelke et al. [46] showed the common and discriminative topics of InfoVis, SciVis and Siggraph. Jiang et al. [31] explored the interactions among visualization, data mining, and computer graphics by detecting co-occurrence topical words, they also showed the evolution of topic relevance by Sankey diagrams. Isenberg et al. [29] created CiteVis2 and CiteMatrix, which are simple tools for analyzing the relationship between three main branches of visualization by showing citations between them. Heimerl et al. [27] drew the knowledge flows from other fields to visualization with user-steered citation aggregation. Chuang et al. [11] mapped the topic similarity between university departments in an egocentric visualization, where undistorted distance guarantees trustworthy similarity. Topicpanorama [40] packed hierarchical topic graphs in a radial icicle plot, offering a comprehensive visual summary that presents different aspects of relevant topics. PivotSlice [65] provided a flexible capability to discover entity relationships by constructing a series of dynamic queries on tabular views. Federico et al. [20] summarized various efforts in VIS community on relation-seeking on literature data. The works above focus on low-level analysis units, such as entity and topic; our system also provides a wider exploration space by enabling exploration at discipline level.

2.3 Visual Exploration of the Evolution of Scientific Fields

Small multiples and animation are widely used methods for revealing dynamic evolution of graphs and scatterplots. For example, Zhao et al. [66] generated multiple co-word networks of different periods to indicate the changes happended in knowledge domains. Alsakran et al. [1] used animation to trace theme variations in research projects. Chen et al. [7] progressively visualized the evolution of a knowledge domain by merging a series of time-registered co-networks. However, the effectiveness of small multiples and animation sharply declines when losing context due to drastic varieties between periods or extremely high number of nodes shown in each period. Aggregation is a common way to mitigate this problem. For example, Isenberg et al. [30] characterized topic clusters of VIS community by strategy diagrams, in which the changes in the position and size of clusters between periods reflect the evolution. Similarly, we create a series of snapshots to track the variation in hotspots which are detected by contour line.

Timeline is another broadly applied technology for reflecting topic evolution. Numerous timeline-based visualizations [8, 16, 17, 32, 37, 45] projected articles to horizontal time-axis, and denoted topics or clusters as vertical axis. These visualizations can reveal the development

[1] http://www.galex.caltech.edu/about/overview.html

[2] http://csrankings.org

[3] We collect the text data from Scopus in two ways: bulk export by querying the conference name; single export by querying the paper title.

[4] Inter-topical papers are those that cover multiple topics.

of topics intuitively. Tiara [41] and CiteRivers [27] extended ThemeRiver [25] by embedding topical words in multiple non-overlapping time slices of theme streams. However, non-overlapping time slices prevent analysts from checking the cumulative changes. Besides, the above systems do not allow adjusting the position and length of time slices. TextFlow [14] further extended the above works by introducing the splitting and merging of topics along with a timeline. Rosvall et al. [48] developed a Sankey-like alluvial diagram to illustrate the science development such as the emergence of a field. In our system, we map topics into a 2D plane because we believe that, compared to representing topics in only one dimension (rivers), 2D space is more intuitive and effective in capturing topic relationships. Also, we design a time brush to support creating arbitrary time slices.

2.4 Literature Visualization Software

Various existing literature visualization software achieve complex analysis tasks by integrating multiple views. Combining with visual analysis and data mining on networks, CiteSpace [8] assists analysts with literature review, scientific paradigms identification and topic evolution detection. VOSviewer [59] focuses on creating bibliometric networks where nodes can be journals, authors, papers, and phrases. It uses heatmap to represent hotspots. CitNetExplore [61] aims at analyzing the development of a research topic and identifying its original and influential papers, by visualizing a reference stream. Jigsaw [56] provides a series of views, such as a list view to help analysts explore entity associations, and a word tree view to display the main contents of a paper collection. Action Science Explorer [18] integrates reference management and network analysis; automatic document collection summarization is a practical tool of it. As a feature compared to the above systems, Galex provides a hierarchical understanding of one discipline from macro to micro.

3 Requirement Analysis

We invited four professors who specialize on visualization, computer networks, artificial intelligence (AI), and computer vision (CV) to discuss the possible usage scenarios of uncovering the evolution and intersection of a discipline. All the professors are doctoral supervisors, and two were once invited by school's disciplinary planning department to help make decisions on resource integration and allocation.

The professors proposed three typical usage scenarios. First, for disciplinary planning, decision-makers have to know the relationships between fields to decide whether to integrate resources for joint development. Second, for daily research, supervisors and students have to perceive the research trends by checking the dynamic of hotspots. Occasionally, they search for research opportunities by investigating the gap between topics. Finally, for career planning, researchers and students need to compare academic performances and topic distributions among institutions to decide which institution to join in.

Based on these underlying usage scenarios and the literature review, we identify the following specific requirements:

R1 - Comprehendsion of the overview of one field, inspection of the relationship between its sub-fields, and rapid determination of key entities, such as phrases, institutions, and authors, of any part of the field;

R2 - Perception of hotspots and their evolution, detection of the emergence, rise, decline, fusion, and separation of a field, and inspection of changes in key entities across time;

R3 - Checking the influence scope of a field and exploration of its intersection with other fields;

R4 - Comparison of the academic performances and topic distributions among institutions;

R5 - Detection of top scholars and institutions in a field and assessment of the cooperation among them.

4 Interface and Interactions

Fig. 1 shows the interface of Galex. The area picker allows analysts to select single or multiple areas of interest. The main panel is the core panel for hierarchical analysis of evolution and intersection and contains three layers with progressively refined data description granularity. The network panel includes three networks that provide different perspectives for interpreting the structure of an area. Below are details about interface, design consideration, and interaction of each layer, as well as an introduction to the network panel.

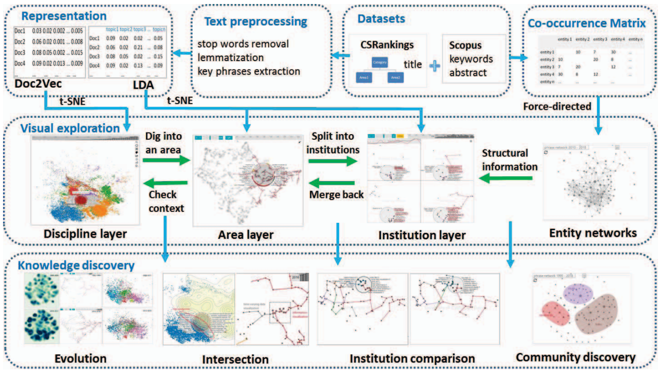


Fig. 2: Pipline of Galex.

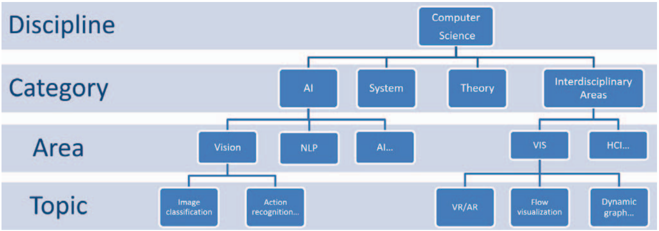


Fig. 3: Schematic of terminologies of science hierarchy. Categories and areas are determined by CSRankings. The AI category includes an area also called AI.

4.1 Discipline Layer

Following the design principle of "overview first, zoom and filter, then details-on-demand" [51], we provide an overview (galaxy) of computer science at the discipline level. As shown in Fig. 1, each of the four categories (AI●, systems●, theory●, and interdisciplinary areas●) corresponds to a color scheme in which the areas are differentiated by hues generated by ColorBrewer [24]. The layout process of papers consists of three steps: preprocessing of titles, keywords, and abstracts by stop word removal, lemmatization, and key phrase extraction by ToPMine [19]; obtaining the document vectors via Doc2Vec [36]; dimension reduction by t-SNE [42] (we abbreviate the last two steps as Doc2Vec + t-SNE). All these three steps are executed offline. In this manner, papers with similar semantics are placed nearby. The fuzzy boundaries between areas imply potential intersections.

To build a mental map of computer science from a hierarchical perspective (Fig. 3), we use plenty of colors in a single view. To alleviate the difficulties in distinguishing areas and locating a specific one, we propose two solutions: 1. providing a pallet that enables analysts to reassign colors for any group of areas; 2. implementing a contour line in which color depth is proportional to the density of papers, and line spacing is inversely proportional to the density gradient. Compared with the pallet, the contour line provides the following benefits on facing vast and highly overlapping data points:

1. Estimation of the scope of one area, identification of the core zone, and detection of the potential intersection zone between two areas;
2. Enabling of analysts to focus on the evolution of the core zone by hiding the details;
3. Relief from the common misleading notion that the greater scope covered by an area, the vaster is the quantity of its papers;
4. Reminder indicating that the same area may be compartmentalized into disconnected parts;

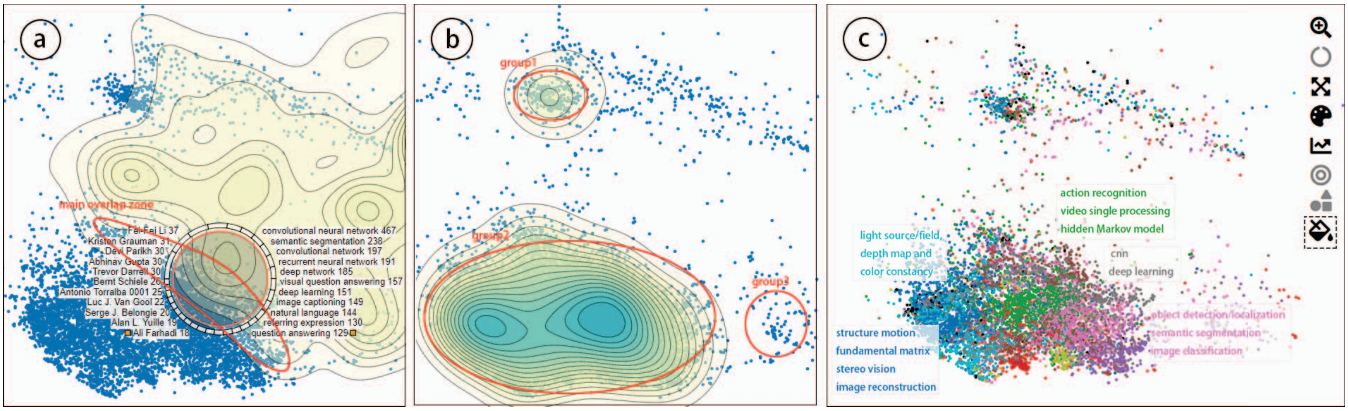


Fig. 4: The main overlap zone between the AI (contour line) and CV area (points) located at the upper right side of the latter. Spotlight indicates the top scholars, institutions (arcs), and phrases inside the covered region ⓐ. Contour line of CV indicates that most papers are distributed across three groups and Group 2 owns most of the papers, forming two cores ⓑ. Paper will be re-colored by its predominant topic after clicking 🪣 ⓒ.

5. Identification of one area by directly pointing out its location.

Fig. 4(a) shows the overlap zone between CV and AI by keeping only the former in the landscape and overlaying the latter's contour line. The articles in the overlap zone constitute the "passage through science" [54], which may include ideas and technologies that bridge one domain to another.

The toolbox on the right provides several effective interative components to support sense making. To elucidate the design decisions and roles of the main tools in the workflow of the discipline layer, we use the AI area as an example to show how to obtain an objective understanding of the AI area through our system in terms of its impact in computer science, its main topics, and evolution.

After activating the contour line of AI area by moving the mouse over its picker, analysts find that the whole AI category locates at the bottom of the galaxy overview and the papers of the AI area are widely distributed, resulting in intersections with other areas in the AI category. Then, the analysts hide the other areas and fit the AI area in the center of the screen by using 🔍 (zoom and pan). To understand the research contents in the AI area, the analysts use ○ to create a circular zone filter called spotlight. Similar to the lens in DocuCompass [28], our spotlight can be freely re-sized and moved, enabling analysts to explore document landscapes effectively. Additionally, our spotlight extends the DocuCompass in terms of correlation analysis of key entities and comparative analysis of multiple document collections. We will discuss in detail in the next two sections. The pivotal entities involved in the selected zone, such as discriminating phrases, top authors, and top institutions (sectors), are shown around the spotlight, revealing the semantic details (Fig. 4). Combining 🔍 and ○, analysts could freely explore any location in the galaxy of computer science and discover and interpret outliers, clusters, and intersections.

However, analysts may notice that occasionally, the key phrases provided by the spotlight are chaotic or extremely broad to make sense. The reasons are two-fold:

1. By default, the key phrases are the top frequent keywords in the covered zone. Many of them are index terms (e.g. "artificial intelligence") in Scopus and appear in almost all AI papers, resulting the loss of discrimination when zooming in the AI area. We solve this problem by introducing G^2 statistics [47] (enabled by ●▲) which is proven to be effective and better than TF-IDF [50] in detecting high-quality descriptive words [10, 13, 27]. Compared with the straightforward frequency statistics, G^2 requires more computation because it considers all the papers by default, including those uncovered ones that act as reference. However, this scalability issue can be relieved by sampling the uncovered ones.
2. When the spotlight covers more than one semantic zone, the meaning revealed by the key phrases is ambiguous. Analysts must be explicitly reminded of the distribution of semantic zones. By clicking 🪣, each paper of a given area will be re-colored according to its dominant topic. The number of topics is determined by analysts and is set to 8 by default. We call this function a topic assistant. It is implemented by pre-using LDA results which are heavily used in the next layer. In this way, analysts could adjust the size and position of the spotlight reasonably.

After enabling the G^2 statistics and topic assistant, analysts are able to quickly grasp the main topics in the AI area with the spotlight. Next, analysts would like to examine the evolution of the AI area, and this goal can be achieved in two manners:

1. Using the time brush at the top of the main panel. The time brush allows analysts to rapidly build arbitrary time filters, including non-overlapping time slices and cumulative time slices, by zooming and moving a time window. The contour line will be updated automatically after changing the time brush.
2. Creating small snapshots by time slice creator. The time slice creator (Fig. 5) enables the analyst to produce consecutive, non-overlapping, and equal-length time slices expediently. In small snapshots (Fig. 6), an enduring core can be considered as the knowledge base, whereas the transient trends act as research frontiers. The key entities of each time slice can be ranked by frequency statistics or G^2 statistics. Similar to Parallel Tag Clouds [13], analysts can immediately check the ranking changes of an entity between periods.

These two manners complement each other. Compared with the time brush, small snapshots reduce repetitive operations of splitting time and free the analyst from remembering the key entities of multiple periods. Nevertheless, we emphasize the arbitrariness of the time brush which facilitates analysts to examine closely when a hotspot started and how long it lasted. A typical workflow is that analysts first use small snapshots to obtain a macroscopic overview of hotspots and pick a period of interest, then explore a more precise time point with the time brush, next return to the small snapshots and apply the learned time slice to obtain an exact and easy-to-understand presentation.

After creating a series of small snapshots with a five-year interval, the analysts notice that the AI area has expanded rapidly to other areas in recent years and learn about current trending technologies and their application fields (Fig. 6). At this point, all information needs are

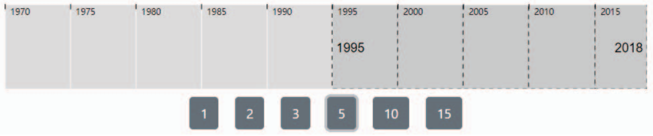
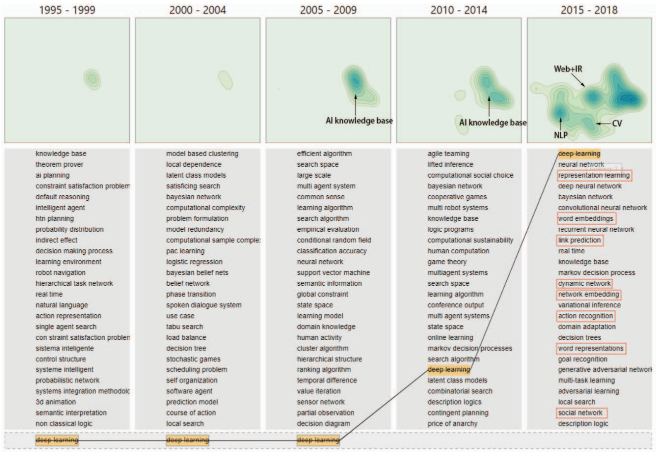


Fig. 5: Time slice creator allows analysts to create small snapshots in two steps: selecting a time interval by brushing and then clicking one bottom button to segment the interval into equal-length (here, five years) time slices.



1995 - 1999	2000 - 2004	2005 - 2009	2010 - 2014	2015 - 2018
knowledge base	model based clustering	efficient algorithm	agile teaming	deep learning
theorem prover	local dependence	search space	lifted inference	neural network
ai planning	latent class models	large scale	computational social choice	representation learning
constraint satisfaction problem	satisficing search	multi agent system	bayesian network	deep neural network
default reasoning	bayesian network	common sense	cooperative games	bayesian network
intelligent agent	computational complexity	learning algorithm	multi robot systems	convolutional neural network
htn planning	problem formulation	search algorithm	knowledge base	word embeddings
probability distribution	model redundancy	empirical evaluation	logic programs	recurrent neural network
indirect effect	computational sample comple:	conditional random field	computational sustainability	link prediction
decision making process	pac learning	classification accuracy	human computation	real time
learning environment	logistic regression	neural network	game theory	knowledge base
robot navigation	bayesian belief nets	support vector machine	multiagent systems	markov decision process
hierarchical task network	belief network	semantic information	search space	dynamic network
real time	phase transition	global constraint	learning algorithm	network embedding
natural language	spoken dialogue system	state space	conference output	variational inference
action representation	use case	learning model	multi agent systems	action recognition
single agent search	tabu search	domain knowledge	state space	domain adaptation
con straint satisfaction problen	load balance	human activity	online learning	decision trees
sistema inteligente	decision tree	cluster algorithm	markov decision processes	word representations
control structure	stochastic games	hierarchical structure	search algorithm	goal recognition
systeme intelligent	scheduling problem	ranking algorithm	deep learning	generative adversarial networl
probabilistic network	self organization	temporal difference	latent class models	multi-task learning
systems integration methodolc	software agent	value iteration	combinatorial search	adversarial learning
3d animation	prediction model	sensor network	description logics	local search
semantic interpretation	course of action	partial observation	contingent planning	social network
non classical logic	local search	decision diagram	price of anarchy	description logic
deep learning	deep learning	deep learning		

Fig. 6: Small snapshots and entity lists of the AI area. In recent years, AI techniques have rapidly applied to other areas of the AI category which significantly expanded the original knowledge base of the AI area. For example, in 2015-2017, deep-learning-related technologies, such as convolutional neural network, recurrent neural network, and generative adversarial network, were widely used in the following fields: representation learning (word embedding and network embedding), social network (link prediction and dynamic network), and CV (action recognition). Currently, phrases are sorted by G^2 statistics.

met. If the analysts need more details about topics, they should further explore the area layer.

4.2 Area Layer

In the area layer, analysts zoom into a specific area to track the history of its topics and find representative papers of each topic and inter-topical papers. The points in the area layer still represents the papers; the layout algorithm is still t-SNE (Fig. 7). However, the document vectors we used for layout not originate from Doc2Vec but the topic probability distributions generated by topic model PhraseLDA [19]. We call this process flow as LDA + t-SNE. We abandon Doc2Vec because its vectors trained by considering all computer science papers will emphasize the diversity between areas but weaken the difference among topics of an individual area. Thus, we cannot obtain a fine-grained and well-defined semantic structure of one area by directly using its Doc2Vec vectors. Fig. 8 shows the results of applying Doc2Vec + t-SNE workflow (a) and LDA + t-SNE workflow (b) on computer networks area. The prominent topic or cluster of each paper determines its color. We use HDBSCAN [5], an outstanding soft cluster algorithm, to determine the clusters for Doc2Vec + t-SNE workflow. In this case, the numbers of topics that are manually set in LDA model and automatically determined by HDBSCAN are 8 and 2, respectively. The points in (b) are nearly uniformly distributed, and the size of clusters are notably uneven. On the contrary, we can quickly identify individual topics in (a), and the meaning of each topic is verified as distinct by the expert of computer networks. Similar consequences are observed in several other areas.

For each topic, we regard those papers whose membership with the current topic is larger than a user-set threshold 0.5 as the papers belong to this topic. Then, we generate a minimum spanning tree (MST) for each topic using the 2D coordinates of papers calculated by t-SNE. We opt not to compute the MST in the original high-dimensional space because such action will create a mess of line crossings in the local zone due to the information loss brought by dimensionality reduction and these visual confusions will cause poor user experience. We call these MSTs as topic trees and assign a color to each of them. The higher the threshold, the fewer papers each topic tree contains. The papers left by a high-threshold filter can be regarded as representative papers of each topic. Papers that are connected by multiple colored lines are inter-topical papers. We use topic trees to express topic distributions rather than simply coloring papers according to its predominant topic because the former overcomes the following two notable shortcomings of the latter: 1. the distortion caused by neglecting other topics may mislead analysts; 2. inconvenient recognition of inter-topical papers.

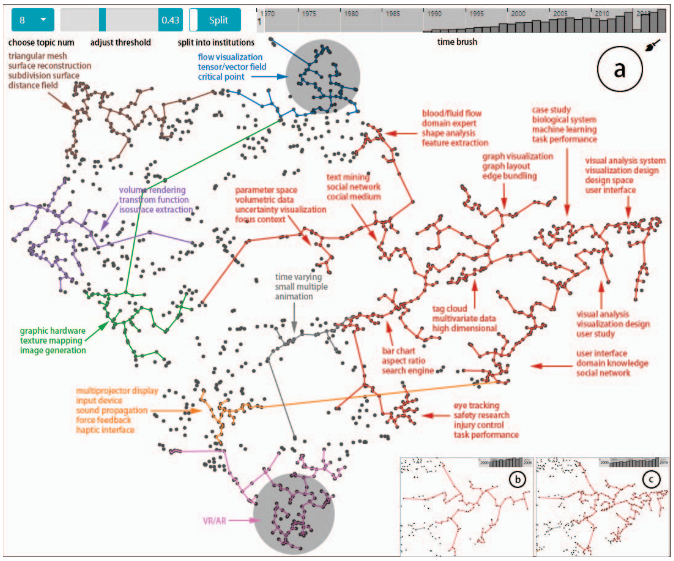


Fig. 7: Topic trees reveal eight topics in the visualization area ⓐ. Using the cumulative time slice, we can determine that the topic tree of VA&InfoVis grew readily during 2010-2014 (ⓑ, ⓒ).

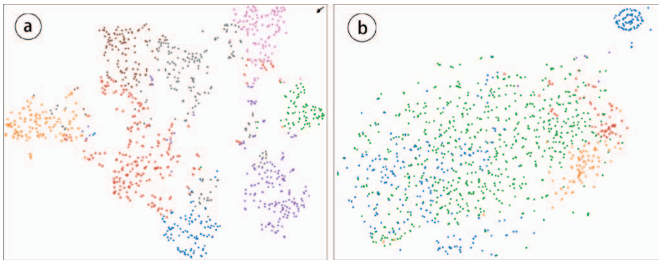


Fig. 8: Layout results of the LDA + t-SNE workflow ⓐ and the Doc2Vec + t-SNE workflow ⓑ on the computer networks area. The color of paper is determined by its predominant topic. The clusters (topics) in ⓐ are visually significant and semantically explicit, as verified by an expert in computer networks.

Same as the time brush at the discipline layer, the time brush at the area layer supports freely creating time slices. Cumulative time slice allows analysts to view the papers being added to the topic tree gradually, thus providing a historical playback view of the growth of tree branches (Fig. 7 (b), (c)). As the topic expands, topic trees begin to connect and inter-topical papers act as bridges between topics. By creating non-overlapping time slices, analysts could inspect the prosperity or decline of topics (Fig. 15).

We enhance the interaction of spotlight at the area layer by combining context awareness and detail exploration. When the mouse moves to an institution (author), its main collaborators found in the covered papers will be connected to it by a curve (Fig. 9 (a), (b)). At the same time, all papers published by the focused institution (author) in the whole area are highlighted, whereas the others are blurred to offer a

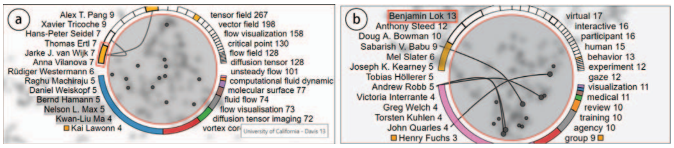


Fig. 9: Enhanced spotlights corresponding to the two dark grey circles in Fig. 7. UC Davis owns the most papers on the topic of flow visualization. Three scholars at UC Davis published thirteen papers and collaborated with two other institutions on this topic ⓐ. Benjamin Lok published the most papers on VR&AR topic. He has three main cooperators on this topic ⓑ.

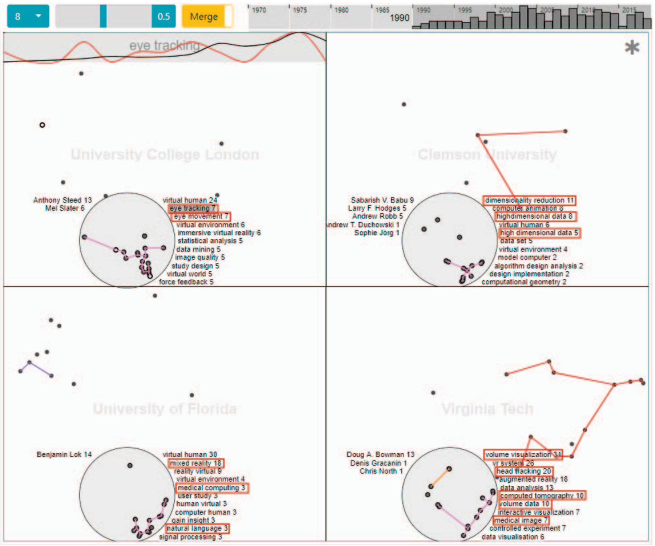


Fig. 10: Interface of the institution layer. The synchronous spotlights allow analysts to compare the research contents of institutions. A line chart indicates the heat of a focused phrase.

context. The bottom half of the ring represents the topic probability distribution of the covered papers. When the mouse hovers over an arc, its corresponding topic tree will be highlighted. The above interactions are cue-based focus + context techniques [12] that support context-awareness and navigation.

The enhanced spotlight also enables analysts to dig into details. When the mouse hovers over an author for more than 1 s, the keywords on the right will update to the main keywords in the publications of that author (Fig. 9 (b)). Analogously, when the mouse hovers over a keyword for more than 1 s, the authors on the left will update to the top authors who focus on that keyword.

Evidently, our spotlight not only concentrates on characterizing focused documents by revealing their feature phrases but also on integrating entity correlation analysis by exploring collaborations among institutions and scholars. This is the first core difference between our spotlight and many other interactive lenses [58], especially the lens in DocuCompass [28].

4.3 Institution Layer

At the institution level (Fig. 10), Galex allows analysts to identify the differences between institutions in terms of academic performance, topic distribution, research content, and evolutionary history.

Switching from the area layer to the institution layer is achieved by animation. The area layer will be drawn in n copies (n = the number of concerned institutions) after clicking the split button Split. Then each copy is translated to an institution grid with the same size. Subsequently, for each grid, papers that are not published by the corresponding institution are faded out, meanwhile its topic tree is constructed. Smooth transitions help analysts perceive associations between layers without losing context.

To reveal the specific differences under the same topic, we design an interaction technology called synchronous spotlight. The spotlight created in any institution grid will be replicated to the same location in other institution grids. These spotlights can move synchronously and display key entities of each institution extracted by G^2 statistics. Synchronous interaction is the second core difference between our spotlight and the lens in DocuCompass [28]. When the mouse points at a key phrase in spotlight, a line chart will appear at the bottom or top of the focused grid, showing the changes over time in terms of its frequency and the total cited times of papers containing it. These indicators reveal potential research fronts, with which analysts could determine which institutions are more focused on the research fronts. The differences in topic distribution are evident by comparing the topic trees between institutions. Combining the time brush, analysts could understand the differences in evolutionary history.

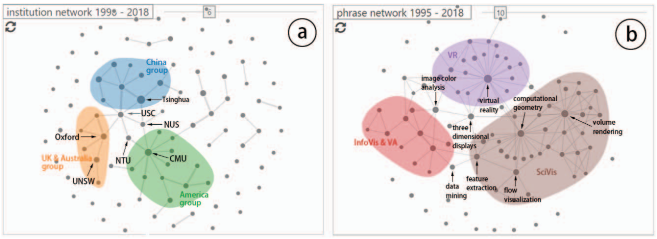


Fig. 11: Cooperation network of top 100 institutions of AI area ⓐ (all papers are from AAAI and IJCAI conference) and co-occurrence network of top 100 phrases of VIS area ⓑ (all papers are from VIS and VR conference) (colored bubbles are hand-painted). After filtering out the link with less than five times of cooperation, we detect three academic groups in the AI area, namely, America group, China group, and UK&Australia group. National University of Singapore (NUS), Nanyang Technological University (NTU), and University of Southern California (USC) act as pivots. After deleting the phrase (“visualization”, “data visualization”, and “computer graphics”) that links with almost all the other phrases, we identify three main topics in VIS area: SciVis, InfoVis&VA, and VR&AR.

The spotlight and time brush are two core interactive components of Galex, running through all three levels of the main panel. The balance between customization and consistency makes it easier for analysts to learn how to use our system.

4.4 Network Panel

We provide three entity networks in the network panel: phrase co-occurrence, author cooperation, and institution cooperation networks. These networks help analysts to deepen the understanding of the structure of research areas from different perspectives, as a complementary to the structure information captured by text data in the main panel.

Analysts could extract the skeleton of the network by filtering out less weighted links. For example, in the institution cooperation network, analysts can filter out links with less than a certain number of cooperations, and then identify major academic groups (Fig. 11 (a)). In each network, arbitrary nodes can be removed to simplify the network. For example, in the phrase co-occurrence network, after deleting the phrase that exist in almost all papers, we could identify the main topics of one area (Fig. 11 (b)). The delete interaction is proven to be effective by comparing with the initial meaningless phrase network shown in Fig. 1.

5 Case Studies

We consider three cases to demonstrate the effectiveness of Galex in uncovering the evolution and intersection of one field and the potential of Galex in real-world usage scenarios. All the insights into the cases have been verified by domain experts.

5.1 Case 1: Focus on Computer Vision

All the papers in this case are from the top three CV conferences, namely, CVPR, ICCV, and ECCV, with a time span from 1986 to 2018. We imagine an analyst David who is an expert in CV and is currently writing an article to retrospect the history of CV. He wants to verify or refute his subjective experience in a data-driven manner.

First, David intends to check whether the topics given by Galex are in line with his mental map. He checks the location of CV in the galaxy of computer science via a contour line. He observes that CV is surrounded by other areas of AI category, such as data mining&machine learning and AI. Focusing on CV, he hides all the other areas and notices that most papers are distributed in three main groups (Fig. 4 (b)). The largest group holds most of the papers whcih form two core zones with relatively high density. The interest in discovering the group meaning, especially the two dense zones, drives David to use the spotlight. Initially, he sets the size of the spotlight to cover the whole group when checking each group. In this case, only the meaning of Group 2 is identified as evident and straightforward. This group

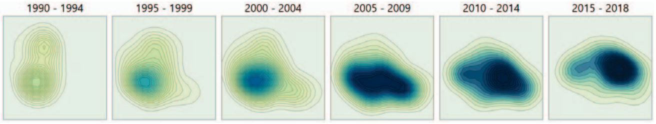


Fig. 12: Evolution of computer vision from 1990 to 2018. It shows a noticeable right shift of the hotspot from 3D vision and computational-geometry-related topics to 2D-image-processing-related topics which involve machine learning and deep learning techniques.

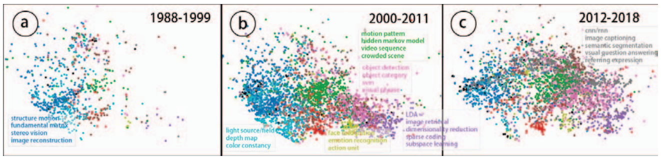


Fig. 13: Two shifts in the history of computer vision. The first shift is topic-driven, from ⓐ to ⓑ, with the emerging of several new topics. The second shift is technique-driven, from ⓑ to ⓒ, with the spread of deep learning techniques.

consists of papers that study face recognition, image segmentation, and image classification using data mining and machine learning methods, such as matrix decomposition, subspace clustering, and kernel learning.

The chaotic semantics of the other two groups makes David realize that he needs finer-grained semantic information. Thus, he enables the topic assistant by clicking . He adjusts the size of the spotlight and discovers that the left core consists of two topics: topic 1, which is about light source/field, depth map, and color constancy ●; topic 2, which is about structure motion, fundamental matrix, stereo vision, and image reconstruction ●. The right core consists of three topics: topic 1, which is about action recognition, video signal processing, and hidden Markov model ●; topic 2, which is about object detection/localization, semantic segmentation, and image classification ●; topic 3, which is about convolutional neural network and deep learning ● (Fig. 4 (c)). David confirms that these topics are almost the same with those in his mind. Then, David realizes that the two cores represent two important camps of CV: the left core represents the 3D vision in traditional vision research; the right core represents the learning-based 2D image processing that involves machine learning and deep learning. To further validate this assumption, David overlays the contour line of machine learning&data mining (ML) and AI, and he notices that they both overlap with CV on its right side (Fig. 4 (a)) which supports his conjecture.

David also wants to browse the evolution of CV. As we all know, deep learning has risen rapidly in CV in recent years. Thus, the right core is predictably rising. The questions that really interest David are whether the left core declines, and whether the hotspot directly transfers from the left core to the right one. To find out, he first uses the time slice creator to create six snapshots from 1990 to 2018 with a five-year interval (Fig. 12). He observes a remarkable right shift of hot topics (core zones). Preliminary conclusion are evident: more researchers have been attracted by the advantages of deep learning, whereas the attention to traditional 3D vision has dropped dramatically. However, David notices that the right shift was already significant in 2005-2009, whereas deep learning technologies broke out in academia since 2012, indicating that deep learning cannot fully explain the right shift. After reviewing the keywords in 2005-2009, re-checking the contents of papers located between the two core zones, and constructing time slices of different lengths by the time slice creator and the time brush, David gradually understands the complete reasons for the right shift.

The right shift occurred twice in history, dividing the whole history of CV into three stages: 1988-1999, 2000-2011, and 2012-2018 (Fig. 13). The first shift indicates a topic-driven hotspot transfer. Before 2000, scholars focused on 3D vision and computational geometry whose main contents were structured light, light field, fundamental matrix, computational geometry, and camera equipment. From 2000 to 2011, with the development of data mining and machine learning algorithms, researchers introduced new topics related to 2D image processing, such as semantic segmentation, object detection, face recognition, image classification, and motion recognition, etc. The techniques used in this period included hidden Markov, support vector machine, shallow network, sparse coding, linear discriminant analysis, and graph cut, etc. The second shift is a technique-driven paradigm shift. After 2012, deep learning developed rapidly, sweeping almost all the topics and technologies flourished in 2000-2011. Although the research questions that interest the academia were nearly unchanged, deep learning techniques significantly concentrated recent papers to the upper right corner of the CV, where the CV and AI area intersect. Meanwhile, the 3D vision was less affected, especially for topic 1 of the left core. It showed new opportunities in recent years, such as rain streak and rolling shutter. Overall, however, the heat of the 3D vision further declined. David thinks that the idea of two-stage shift with two driven modes is novel since he has never realized it before.

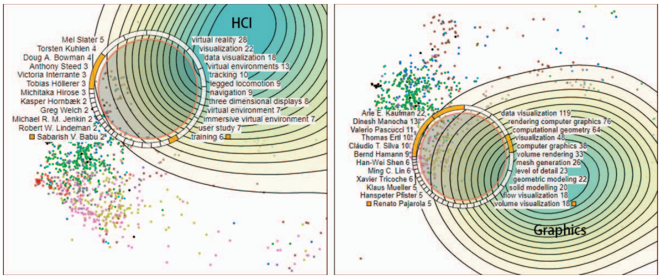


Fig. 14: Human-computer interaction and computer graphics overlap with the upper and lower parts of the visualization, respectively.

5.2 Case 2: Focus on Visualization

All papers in this case are from VIS and VR conferences from 1990 to 2018. We imagine an analyst Ella, who is an undergraduate student with a keen interest in visualization. She is taking a visualization course and aspires to gain a more systematic understanding of visualization at different levels of detail by our system. This information also help she find a school to pursue her master's degree with a major in visualization.

First, Ella intends to grasp a macroscopic understanding of visualization in the discipline layer. She turns off other areas, enables the topic assistant, and checks the semantics of each topic with the spotlight. She notices that papers in the upper part are about VR&AR, information visualization (InfoVis) and visual analysis (VA), while papers in the lower part are about scientific visualization (SciVis), including keywords, such as volume rendering, ray tracing, and computational geometry, etc. Then, Ella overlaps the contour line of Human-computer interaction (HCI) and Computer Graphics and observes that the intersections of visualization with these areas are located at the upper and lower of the visualization, respectively (Fig. 14). These findings are consistent with her perception of visualization learned from the course.

Then, Ella moves down to the area level and expects to obtain specific details about evolution. She checks all the pre-computed LDA results with different topic numbers and finally determines one with eight topics, because its topic clusters are well-separated (Fig. 7). Then, she carefully examines each topic using spotlight and believes that the semantics of most topics are clear and specific. The topic with the largest space ● can be interpreted as the combination of information visualization (mainly on the left side) and visual analysis (mainly on the right side).

Next, Ella creates a five-year time slice using the time brush and moves it to learn the evolution of visualization (Fig. 15). She observes that before 2005, the red topic tree was small and grew slowly; it consisted of the early classic works on information visualization. By contrast, the topic tree of flow/vector field visualization ●, triangular/polygonal mesh ●, transfer function, volume rendering ● and image generation, texture mapping, and graphics hardware ● grew considerably faster. However, in the next five years (2005-2009), the tree of VA and InfoVis presented a rapid growth. In the next decade (2010-2018), the red topic tree continually and rapidly expanded and intersected with others, whereas the others declined except for the VR&AR topic, which also grew rapidly, especially in the recent five years.

Next, Ella is interested in time-varying data and visual design, so she

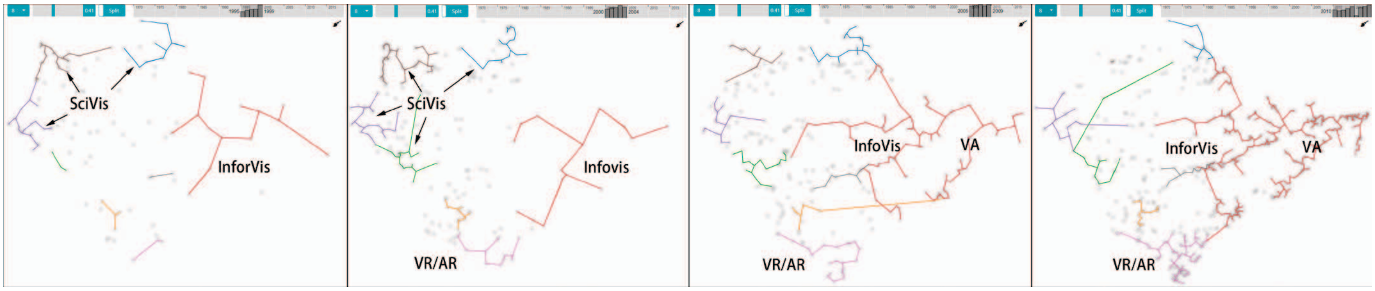
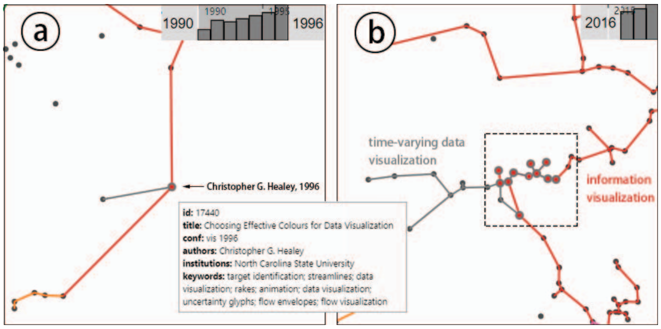


Fig. 15: Growth of topic tree revealing the evolution of visualization. In the early age of VIS (1990-2004), SciVis dominates the research contents. For the next 15 years, VA/InfoVis and VR&AR grew rapidly. These findings are consistent with the VIS history described in vispubdata.org [29].


a
1990
1990
1995
1996
Christopher G. Healey, 1996
b
2016
2015
time-varying data visualization
information visualization

id: 17440
title: Choosing Effective Colours for Data Visualization
conf: vis 1996
authors: Christopher G. Healey
institutions: North Carolina State University
keywords: target identification; streamlines; data visualization; rakes; animation; data visualization; uncertainty glyphs; flow envelopes; flow visualization

Fig. 16: Intersection between time-varying data visualization and information visualization with the threshold of 0.35. The first inter-topical paper was published in 1996. Eleven inter-topical papers were published in recent three years.

expects to find the work that covers time-varying data visualization● and information visualization●. In the area layer, Ella lowers the threshold of the topic tree and then creates a cumulative time slice. As the branches of the topic tree extend, she discovers the first paper that meets the requirement: Choosing Effective Colors for Data Visualization [26] (Fig. 16 (a)). Then, Ella focuses on the last three years and identifies 10 underlying satisfactory papers (Fig. 16 (b)). She examines them through keywords and abstract provided in tooltips and confirms that nine are reasonable and desired. Ella believes it is very convenient to find inter-topical papers, and it helps find possible research opportunities by investigating the place where no intersection has ever occurred.

Finally, Ella looks forward to understanding the current progress in VR&AR. By spotlight, she finds the following top four institutions in this field: University College London (UCL), Clemson University (CU), University of Florida (Florida), and Virginia Tech (VT). She sets these institutions as focus and zooms into the institution layer (Fig. 10). After comparing the paper distributions among these institutions and using the synchronous spotlight to view specific differences in research contents, Ella finds that only Florida and VT show interests to other topics (SciVis and VA, respectively) apart from VR/AR. In the VR&AR topic, UCL conducted several works on eye tracking and ray tracing; CU explored the possibility of using VR for high-dimensional data analysis; Florida applied mixed-reality technology on medical computing and natural language processing; VT focus on volume data analysis and medical image analysis.

5.3 Case 3: Focus on the Entire Computer Science

In this case, Ryan is an investor who has a keen eye for hotspots, he wants to first learn the evolution patterns of computer science and the areas of greatest change and then decide whether to invest.

First, Ryan divides the entire period from 1970 to 2017[5] into five segments with a 10-year interval (Fig. 17)[6]. He notices that developments in computer science between 1970 and 2000 focused on traditional areas, such as databases, programming languages (PL), computer architecture (Arch), and design automation (EDA). In the recent decade, the rate of development increased significantly. The cores of each area gradually expanded, and they are no longer isolated fields of studies. In particular, the five areas in the AI category are almost all linked. The leading force changed from traditional areas to application areas, such as HCI, AI, ML, and robotics. Then, Ryan divides 2010-2017 into three periods (Fig. 18) and notices that the most drastic changes in the recent decade occurred in CV, natural language processing (NLP), HCI, software engineering (SE) and computer security (security). More opportunities can be found in these areas based on Ryan's experience.

Ryan also identifies numerous evolution patterns on individual or pair of areas. For example, databases area is split into two parts from a compact core in the last decade. Robotics area goes from two cores (2000-2009) to multiple cores (2010-2017). Visualization has grown to be an independent area through separating from computer graphics in the last ten years[7].

6 Discussion

6.1 User Feedback

We presented Galex to the four professors who previously participated in the discussion on usage scenarios and five Ph.D. candidates with a major in different areas of computer science. We showed them all the three cases and then allowed them to interact with our system on the ground for at least 30 minutes. Finally, we collected their feedback informally. They all agreed that Galex is comprehensive and well-designed, enabling them to gain new insights into a discipline from multiple perspectives quickly and effectively. They favored the interaction components for their natural, flexible, and easy-to-use characteristics. They also pointed out some weaknesses, such as the weak logical association of the last layer with the first two, and the newly added papers in topic trees are not explicitly highlighted. They also put forward several suggestions, such as implementing panning and zooming in the area layer and replacing the uncommon interaction: triggering an action by hovering over the element for 1 s. Finally, all the professors are willing to use our system in their daily research.

6.2 Data Quality

Galex expects text data of adequate quality in terms of the completeness of data fields, quantity of paper, and quality of the text itself. As for the data fields, we use title, keywords, and abstract in current implementation. A total of 31.8%, 14.7%, and 11.3% papers miss keywords, abstract, and both, respectively. Papers with only the title lack sufficient information, forming a narrow elliptical chaotic center of the galaxy in the discipline layer (Fig. 1). As for the number of paper, the more the better. Massive data helps constructe a truthful and detailed overview on the premise that the data fields are sufficient. The embedding results of Doc2vec are poor when the amount of data is small. As for the quality of text itself, we use two kinds of keywords,

[5] Note that the text data of 2018 is incomplete for several areas.

[6] Abbreviations: EDA (Design automation), Mobile (Mobile computing), Web+IR (The Web&information retrieval), Embedded (Embedded&real-time systems), HPC (High-performance computing), Crypto (Cryptography), Theory (Algorithms&complexity), Logic (Logic&verification), Ecom (Economics&computation), Comp. bio (Comp. bio&bioinformatics).

[7] The evolution diagram of HCI, databases, and robotics are shown in supplemental materials.

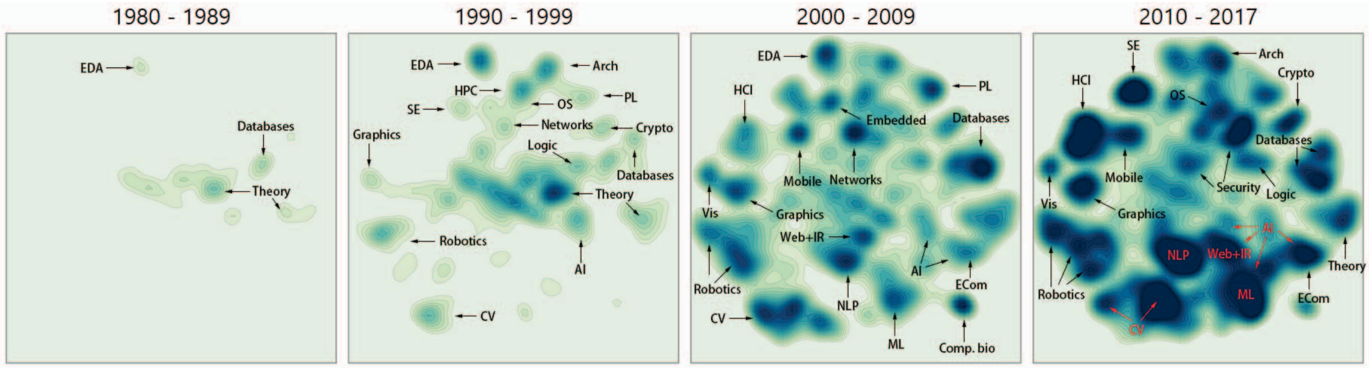


Fig. 17: Evolution of computer science from 1980 to 2017 with a 10-year interval. The core force driving the development of computer science has changed from the traditional fundamental areas to the rising application areas.

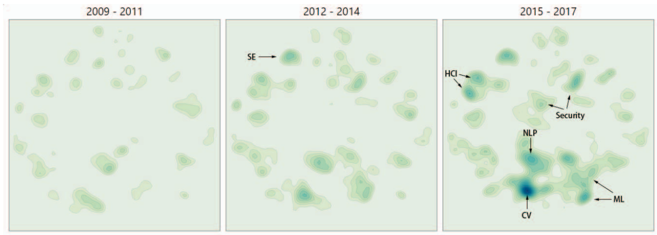


Fig. 18: Evolution of computer science from 2009 to 2017 with a three-year interval. HCI, computer security, and the areas in the AI category have grown rapidly in the last 3 years.

namely, author keywords and Scopus index keywords, with the latter enriching textual information. We recommend using phrases rather than single words to construct corpus.

6.3 Scalability

Overall, Galex exhibits good scalability. First, all t-SNE and LDA are calculated offline, in line with the real-world usage scenario where long wait is unacceptable and most analysts only expect to explore the ready-made landscape. For analysts who would like to upload their own data, in the future, we will provide a set of preprocessing tools for turning raw data into the expected inputting of Galex. In addition, offline computation prevents the instability issue of t-SNE. Additional efficient and scalable dimensionality reduction (DR) techniques, such as Umap [44], LargeVis [57], and t-SNE-CUDA [6], are worth trying. Second, MST is calculated online to support the dynamic user-set thresholds. However, compared with the data at the discipline layer, the data at the area layer are considerably reduced. Higher thresholds reduce the computation of MST because they produce smaller topic trees. Third, rendering and interaction issues faced by magnanimous data can be easily relieved by sampling in the discipline layer, because paper-level details are unnecessary in this layer. As for the area layer, in which analysts may want to find inter-topical papers, sampling is imperfect but is still a reasonable solution. Fourth, other online computations, such as G^2 statistics and contour line drawing, can also benefit from sampling without losing pivotal information. Fifth, we relieve the scalability issue of color using by offering a palette and a contour line design (detailed in 5.1); other techniques introduced by Wang et al. [63] and Gansner et al. [23] are effective in enhancing color discrimination in neighbors.

6.4 Limitations

Galex also has several inherent limitations. First, our spotlight is fixed as a circle, whereas occasionally, semantic regions that interest analyst are non-circular. Inaccurate region selection of the spotlight may cause ambiguous semantics. Second, a low threshold of the topic tree, for example, 0.25 for 8 topics, may create a mass of long edges with meaningless crossings which cause visual confusion. Third, current version of Galex lacks the capability to understand the reason for the formation of chaotic regions, such as the center of the galaxy in the discipline layer. Galex cannot affirm whether the chaotic region is really an intersection of multiple areas or an aggregation with papers that miss text information until we check the raw data of these papers and affirm the latter case.

Two exogenous considerations may limit the widespread use of Galex. First, current used DR algorithm t-SNE dose not support out-of-simple feature which prevents Galex from incrementally updating new data. Umap is a good alternative to overcome this issue. Second, we now use an existing conference-level classification offered by CSRankings for computer science. For other disciplines, a reasonable classification is required.

6.5 Generalizability

Galex, as an integrated analysis framework that focuses on analytical method and design, is not limited to a specific discipline as long as it is fed by quality text data. Becides databases, such as Scopus and Web of Science, several free, high-quality, enormous, multidisciplinary, and easy-to-export scientific literature data sources, such as Microsoft Academic graph, Aminer, Semantic Scholar, and arXiv, can be used for literature analysis. We will attempt to use these data sources in Galex in the future.

7 Conclusion

In this paper, we presented Galex, a visual analysis system that allows analysts to understand the history of a discipline and identify the intersections between its sub-fields by mapping, exploring, and interpreting the science map. The framework and rich interoperable interactions of Galex enable analysts to gain new insights from macro to micro. Three cases and feedback from potential users demonstrate the effectiveness of our system.

We emphasize that as stated in *Mapping Scientific Frontiers* [9], visual analysis systems aim at presenting evidence and suggestions, helping analysts in developing new hypotheses, but the verification of potential findings should be performed by domain experts based on facts and their own knowledge. Occasionally, acquiring a meaningful and coherent story or explanation requires knowledge on the macroscopic sociology and philosophy of science, especially when considering the whole science.

Acknowledgments

This work was supported by National NSF of China (No.61702359).

References

[1] J. Alsakran, Y. Chen, D. Luo, Y. Zhao, J. Yang, W. Dou, and S. Liu. Real-time visualization of streaming text with a force-based dynamic system. *IEEE Computer Graphics and Applications*, 32(1):34–45, 2012.

[2] K. Börner, C. Chen, and K. W. Boyack. Visualizing knowledge domains. *Annual review of information science and technology*, 37(1):179–255, 2003.

[3] K. W. Boyack and R. Klavans. Creation of a highly detailed, dynamic, global model and map of science. *Journal of the Association for Information Science and Technology*, 65(4):670–685, 2014.

[4] K. W. Boyack, R. Klavans, and K. Börner. Mapping the backbone of science. *Scientometrics*, 64(3):351–374, 2005.

[5] R. J. Campello, D. Moulavi, and J. Sander. Density-based clustering based on hierarchical density estimates. In *Pacific-Asia conference on knowledge discovery and data mining*, pp. 160–172. Springer, 2013.

[6] D. M. Chan, R. Rao, F. Huang, and J. F. Canny. t-sne-cuda: Gpu-accelerated t-sne and its applications to modern data. *arXiv preprint arXiv:1807.11824*, 2018.

[7] C. Chen. Searching for intellectual turning points: Progressive knowledge domain visualization. *Proceedings of the National Academy of Sciences*, 101(suppl 1):5303–5310, 2004.

[8] C. Chen. Citespace ii: Detecting and visualizing emerging trends and transient patterns in scientific literature. *Journal of the American Society for information Science and Technology*, 57(3):359–377, 2006.

[9] C. Chen. *Mapping Scientific Frontiers: The Quest for Knowledge Visualization*. Springer Science & Business Media, 2013.

[10] J. Chuang, C. D. Manning, and J. Heer. “without the clutter of unimportant words”: Descriptive keyphrases for text visualization. *ACM Transactions on Computer-Human Interaction (TOCHI)*, 19(3):19, 2012.

[11] J. Chuang, D. Ramage, C. Manning, and J. Heer. Interpretation and trust: Designing model-driven visualizations for text analysis. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 443–452. ACM, 2012.

[12] A. Cockburn, A. Karlson, and B. B. Bederson. A review of overview+ detail, zooming, and focus+ context interfaces. *ACM Computing Surveys (CSUR)*, 41(1):2, 2009.

[13] C. Collins, F. B. Viegas, and M. Wattenberg. Parallel tag clouds to explore and analyze faceted text corpora. In *2009 IEEE Symposium on Visual Analytics Science and Technology*, pp. 91–98. IEEE, 2009.

[14] 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 transactions on visualization and computer graphics*, 17(12):2412–2421, 2011.

[15] G. S. Davidson, B. Hendrickson, D. K. Johnson, C. E. Meyers, and B. N. Wylie. Knowledge mining with vxinsight: Discovery through interaction. *Journal of Intelligent Information Systems*, 11(3):259–285, 1998.

[16] W. Dou, X. Wang, D. Skau, W. Ribarsky, and M. X. Zhou. Leadline: Interactive visual analysis of text data through event identification and exploration. In *2012 IEEE Conference on Visual Analytics Science and Technology (VAST)*, pp. 93–102. IEEE, 2012.

[17] W. Dou, L. Yu, X. Wang, Z. Ma, and W. Ribarsky. Hierarchicaltopics: Visually exploring large text collections using topic hierarchies. *IEEE Transactions on Visualization and Computer Graphics*, 19(12):2002–2011, 2013.

[18] C. Dunne, B. Shneiderman, R. Gove, J. Klavans, and B. Dorr. Rapid understanding of scientific paper collections: Integrating statistics, text analytics, and visualization. *Journal of the American Society for Information Science and Technology*, 63(12):2351–2369, 2012.

[19] A. El-Kishky, Y. Song, C. Wang, C. R. Voss, and J. Han. Scalable topical phrase mining from text corpora. *Proceedings of the VLDB Endowment*, 8(3):305–316, 2014.

[20] P. Federico, F. Heimerl, S. Koch, and S. Miksch. A survey on visual approaches for analyzing scientific literature and patents. *IEEE Trans. Vis. Comput. Graph.*, 23(9):2179–2198, 2017. doi: 10.1109/TVCG.2016.2610422

[21] S. Fortunato, C. T. Bergstrom, K. Börner, J. A. Evans, D. Helbing, S. Milojević, A. M. Petersen, F. Radicchi, R. Sinatra, B. Uzzi, et al. Science of science. *Science*, 359(6379):eaao0185, 2018.

[22] D. Fried and S. G. Kobourov. Maps of computer science. *2014 IEEE Pacific Visualization Symposium*, Mar 2014. doi: 10.1109/pacificvis.2014.47

[23] E. R. Gansner, Y. Hu, and S. Kobourov. Gmap: Visualizing graphs and clusters as maps. In *2010 IEEE Pacific Visualization Symposium (PacificVis)*, pp. 201–208. IEEE, 2010.

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

[25] S. Havre, E. Hetzler, P. Whitney, and L. Nowell. Themeriver: Visualizing thematic changes in large document collections. *IEEE transactions on visualization and computer graphics*, 8(1):9–20, 2002.

[26] C. G. Healey. Choosing effective colours for data visualization. In *Proceedings of Seventh Annual IEEE Visualization’96*, pp. 263–270. IEEE, 1996.

[27] F. Heimerl, Q. Han, S. Koch, and T. Ertl. Citerivers: Visual analytics of citation patterns. *IEEE transactions on visualization and computer graphics*, 22(1):190–199, 2016.

[28] F. Heimerl, M. John, Q. Han, S. Koch, and T. Ertl. Docucompass: Effective exploration of document landscapes. In *2016 IEEE Conference on Visual Analytics Science and Technology (VAST)*, pp. 11–20. IEEE, 2016.

[29] P. Isenberg, F. Heimerl, S. Koch, T. Isenberg, P. Xu, C. D. Stolper, M. Sedlmair, J. Chen, T. Möller, and J. Stasko. vispubdata. org: A metadata collection about ieee visualization (vis) publications. *IEEE transactions on visualization and computer graphics*, 23(9):2199–2206, 2017.

[30] P. Isenberg, T. Isenberg, M. Sedlmair, J. Chen, and T. Möller. Visualization as seen through its research paper keywords. *IEEE Transactions on Visualization and Computer Graphics*, 23(1):771–780, 2017.

[31] X. Jiang and J. Zhang. A text visualization method for cross-domain research topic mining. *Journal of Visualization*, 19(3):561–576, 2016.

[32] D. A. Keim, H. Barro, C. Panse, J. Schneidewind, and M. Sips. Exploring and visualizing the history of infovis. In *IEEE Symposium on Information Visualization*, 2004.

[33] R. Klavans and K. W. Boyack. Toward a consensus map of science. *Journal of the American Society for information science and technology*, 60(3):455–476, 2009.

[34] P. N. N. Laboratory. in-spire. `https://in-spire.pnnl.gov/`.

[35] S. Latif and F. Beck. Vis author profiles: Interactive descriptions of publication records combining text and visualization. *IEEE transactions on visualization and computer graphics*, 25(1):152–161, 2019.

[36] Q. Le and T. Mikolov. Distributed representations of sentences and documents. In *International conference on machine learning*, pp. 1188–1196, 2014.

[37] B. Lee, M. Czerwinski, G. Robertson, and B. B. Bederson. Understanding research trends in conferences using paperlens. In *CHI’05 extended abstracts on Human factors in computing systems*, pp. 1969–1972. ACM, 2005.

[38] Y.-R. Lin, N. Cao, D. Gotz, and L. Lu. Untangle: Visual mining for data with uncertain multi-labels via triangle map. In *2014 IEEE International Conference on Data Mining*, pp. 340–349. IEEE, 2014.

[39] Y.-R. Lin, J. Sun, N. Cao, and S. Liu. Contextour: Contextual contour visual analysis on dynamic multi-relational clustering. In *Proceedings of the 2010 SIAM International Conference on Data Mining*, pp. 418–429. SIAM, 2010.

[40] S. Liu, X. Wang, J. Chen, J. Zhu, and B. Guo. Topicpanorama: A full picture of relevant topics. In *2014 IEEE Conference on Visual Analytics Science and Technology (VAST)*, pp. 183–192. IEEE, 2014.

[41] S. Liu, M. X. Zhou, S. Pan, Y. Song, W. Qian, W. Cai, and X. Lian. Tiara: Interactive, topic-based visual text summarization and analysis. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 3(2):25, 2012.

[42] L. v. d. Maaten and G. Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(Nov):2579–2605, 2008.

[43] K. K. Mane and K. Börner. Mapping topics and topic bursts in pnas. *Proceedings of the National Academy of Sciences*, 101(suppl 1):5287–5290, 2004.

[44] L. McInnes, J. Healy, and J. Melville. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018.

[45] S. A. Morris, G. Yen, Z. Wu, and B. Asnake. Time line visualization of research fronts. *Journal of the American society for information science and technology*, 54(5):413–422, 2003.

[46] D. Oelke, H. Strobelt, C. Rohrdantz, I. Gurevych, and O. Deussen. Comparative exploration of document collections: a visual analytics approach. In *Proceedings of the 16th Eurographics Conference on Visualization*, pp. 201–210. Eurographics Association, 2014.

[47] P. Rayson and R. Garside. Comparing corpora using frequency profiling. In *Proceedings of the workshop on Comparing corpora-Volume 9*, pp. 1–6. Association for Computational Linguistics, 2000.

[48] M. Rosvall and C. T. Bergstrom. Mapping change in large networks. *PloS one*, 5(1):e8694, 2010.

[49] M. Rosvall and C. T. Bergstrom. Multilevel compression of random walks on networks reveals hierarchical organization in large integrated systems. *PloS one*, 6(4):e18209, 2011.

[50] G. Salton and C. Buckley. Term-weighting approaches in automatic text retrieval. *Information processing & management*, 24(5):513–523, 1988.

[51] B. Shneiderman. The eyes have it: A task by data type taxonomy for information visualizations. In *The Craft of Information Visualization*, pp. 364–371. Elsevier, 2003.

[52] R. Sinatra, D. Wang, P. Deville, C. Song, and A.-L. Barabási. Quantifying the evolution of individual scientific impact. *Science*, 354(6312):aaf5239, 2016.

[53] A. Skupin. The world of geography: Visualizing a knowledge domain with cartographic means. *Proceedings of the National Academy of Sciences*, 101(suppl 1):5274–5278, 2004.

[54] H. Small. A passage through science: Crossing disciplinary boundaries. *Library Trends*, 48(1):72–108, 1999.

[55] H. Small, K. W. Boyack, and R. Klavans. Identifying emerging topics in science and technology. *Research Policy*, 43(8):1450–1467, 2014.

[56] J. Stasko, C. Görg, and Z. Liu. Jigsaw: supporting investigative analysis through interactive visualization. *Information visualization*, 7(2):118–132, 2008.

[57] J. Tang, J. Liu, M. Zhang, and Q. Mei. Visualizing large-scale and high-dimensional data. In *Proceedings of the 25th international conference on world wide web*, pp. 287–297. International World Wide Web Conferences Steering Committee, 2016.

[58] C. Tominski, S. Gladisch, U. Kister, R. Dachselt, and H. Schumann. Interactive lenses for visualization: An extended survey. In *Computer Graphics Forum*, vol. 36, pp. 173–200. The Eurographs Association & John Wiley & Sons, Ltd., 2017.

[59] N. van Eck and L. Waltman. Software survey: Vosviewer, a computer program for bibliometric mapping. *Scientometrics*, 84(2):523–538, 2009.

[60] N. J. Van Eck and L. Waltman. Text mining and visualization using vosviewer. *arXiv preprint arXiv:1109.2058*, 2011.

[61] N. J. Van Eck and L. Waltman. Citnetexplorer: A new software tool for analyzing and visualizing citation networks. *Journal of Informetrics*, 8(4):802–823, 2014.

[62] D. Wang, C. Song, and A.-L. Barabási. Quantifying long-term scientific impact. *Science*, 342(6154):127–132, 2013.

[63] Y. Wang, X. Chen, T. Ge, C. Bao, M. Sedlmair, C.-W. Fu, O. Deussen, and B. Chen. Optimizing color assignment for perception of class separability in multiclass scatterplots. *IEEE transactions on visualization and computer graphics*, 25(1):820–829, 2018.

[64] 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 *Proceedings of Visualization 1995 Conference*, pp. 51–58. IEEE, 1995.

[65] J. Zhao, C. Collins, F. Chevalier, and R. Balakrishnan. Interactive exploration of implicit and explicit relations in faceted datasets. *IEEE Transactions on Visualization and Computer Graphics*, 19(12):2080–2089, 2013.

[66] L. Zhao and Q. Zhang. Mapping knowledge domains of chinese digital library research output, 1994–2010. *Scientometrics*, 89(1):51–87, 2011.