A Survey of Scholarly Data Visualization

AA, *Member, IEEE,* BB, *Fellow, OSA,* CC, *Life Fellow, IEEE*

***Abstract*—Scholarly information usually contains millions of raw data such as authors, papers, citations, as well as scholarly networks. With the rapid growth of the digital publishing and harvesting, how to present the data efficiently becomes challenging. Nowadays, various visualization techniques can be easily applied on scholarly data visualization and visual analysis, which enable scientists have a better way to represent the structures of scholarly datasets and reveal hidden patterns in the data. In this paper, we first introduce the basic concept and the collection of scholarly data. Then we provide a comprehensive overview of related data visualization tools, and emphasize existing techniques as well as systems for analyzing volumes of diverse scholarly data. The open issues are discussed to pursue new solutions for abundant and complicated scholarly data presenting as well as techniques that support a multitude of facets.**

***Index Terms*—Scholarly Data, Scholarly Data Visualization, Visual Analysis, Scholarly Data Analysis**

I. Introduction

SCHOLARLY data contains abundant academic resources such as scholarly documents (i.e., papers, books, patents, and scientific reports) as well as associated data including information of authors, citations, figures, tables, etc. In addition, the concept of scholarly networks is central to the study on scholarly data [1], [2]. Scholarly data has become a vital part of scientific research with the appearance of various digital libraries and the rapid development of scholarly data analysis technologies. It enables researchers look into the science itself with a new angle by studying scholarly data [3], [4]. Beyond that, it also helps researchers learn better about acknowledge production processes with wide variety of scholarly data collection methods. Meanwhile, it is not only important for academia, but also promotes the understanding of human social activities, i.e., for sociologists to observe researcher interaction [5] and community formation [6], for countries to evaluate the impact of institutions or scientists and allocate resources [7].

Visualization can be described as ”make something visible”. To be more specific, it represents the process or ability to form a sensible mental picture in a person’s brain. It can also serve as a target to express the visualization results [8]. In this era of computer science, visualization means the technology which can enhance human’s perception by using perceptual competence to visualize the interaction of data [9]. It not only converts the raw data into intuitional graphics, symbols, colors, arts, etc., but also enhances data recognition efficiency and passes valid information. In a word, data visualization transforms the data into multiple easy-understand forms. Initially, two well-established branches of data visualization are that of scientific data visualization and abstract, unstructured information visualization. As data analysis is becoming more and more important, combining visualization with analytics, visual analytics becomes one of the major areas of interest within the field of data visualization. Based on the theory of computer graphics, scientific visualization aims to create a visual expression instead of numerical complex scientific concepts or results [10]. In comparison of scientific visualization, information visualization focuses on dealing with the unstructured high-dimensional data such as textual data, financial data, and multimedia data [11]. Visual analytics is on behalf of “the science of analytical reasoning facilitated by interactive visual interfaces” [12]. One of the most significant current discussions in visual analytics is the development of methods and tools in other areas which related to machine learning, geographic information science, and big data [13], [14], [15].

Scholarly data analysis tries to deal with the problems within the scope of Science of Science [16]. However, as more scholarly data are available to scientists, how to use the massive amounts of data becomes a critical problem to be solved [17]. Fortunately, with the development of visualization technologies, it is much easier to have a better understanding of scholarly data and put it to great use. The process of scholarly data visualization combines the theory of scientific visualization, information visualization and visual analytics. It can transform scholarly data sets into an appropriate representation. It also enables us to have a better understanding of the structure and dynamics of science through a visualization way. Visualization technologies have a strong applicability in scholarly data, whether tools with or without programming language are suitable for users to visualize the data in different ways. For instance, based on the minimum spanning tree, visualization of the author can reflect the collaboration network as a two-dimensional picture. At the same time, it provides a key insight into the direct influence on authors [18]. Scholarly data analysis plays a crucial role in science itself. For example, it can mine implicit relationships which are hidden in the citation networks especially in the co-citation networks. It can help to understand how scientists interact with each other [19]. CiteSpace [20] is well known by its strong ability of analyzing the co-citation network. Through the visualization of Institute for Scientific Information (ISI), Chinese Social Sciences Citation Index (CSSCI), China National Knowledge Infrastructure (CNKI), and other literature database analysis, the system provides the track research areas of hotspots and trends. It helps understand the research frontier and the evolution of critical pathways, important literature, authors and institutions as well.

并列

datasets?

并列

M. Shell was with the Department of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, GA, 30332 USA e-mail: (see http://www.michaelshell.org/contact.html).

J. Doe and J. Doe are with Anonymous University.

Manuscript received April 19, 2005; revised August 26, 2015.

The purpose of this paper is to review recent researches into the field of scholarly data visualization. To the best of our acknowledge, although scholarly data visualization is important, there is no study that provides a comprehensive review of it. Therefore, we summarize the overall research issues on scholarly data visualization. The main issues addressed in this paper include: details of scholarly data and visualization technologies, visualization of scholarly data including single attributes and heterogeneous networks in academia, and visual analytics of scholarly data. The scholarly data visualization section is concerned with the tools and systems used for scholarly data presentation and analysis. There are two primary aims of this study: 1. To provide a comprehensive understanding of development and challenges in the field of scholarly data visualization. 2. To find significant issues which can ascertain the future of this emerging discipline.

The remaining part of the paper proceeds as follows: Details of scholarly data and collection methods are presented in Section 2. Section 3 lays out theoretical dimensions of the visualization tools. Section 4 is concerned with the generic visualization tools and systems used for scholarly data. Section 5 shows how visualization can be combined with various analytical techniques in different visual analytical systems to enlarge the understanding of scholarly data. Finally, this survey is concluded and future works are highlighted in Section 6. The overall idea of exploring scholarly data visualization is summarized in Fig. 1.

II. Scholarly Data Collection

datasets?

Before the visualization of scholarly big data, it is crucial to collect and manage the relevant scholarly data sets. In the age of big data and open science, more and more scholarly documents can be freely accessed from the Internet. In this section, we introduce the entities in scholarly data set and how to collect them for data visualization.

A. Scholarly Data Extraction

Scholarly data is obvious heterogeneous with various entities. Sinha et al. [21] model the academic community as a heterogeneous graph consisting of six types of entities including author, paper, venue, institution, event, and field of study. Specifically, the venue entity means the journal and conference series, e.g. WWW, KDD, TKDE, etc. The event entity means the conference instances, e.g., KDD 2017. The goal of scholarly data visualization is to present the dynamic relationships vividly among these different entities. The relationships and size of these entities can be seen in Fig. 2 based on the dataset of Microsoft Academic Search. Most of these entities can be inferred from the raw data, e.g., author, and citation information. We describe how to collect these data entities from scholarly dataset, e.g., DBLP in the following subsections.

1) Raw Data Extraction: Raw data (or Metadata) is the first set of data extracted from the online digital libraries or academic search engines. It is the basis for academic searching, indexing, and visualization. Specifically, raw data contains authors, title, abstract, keywords, venue, publisher, page number, date of publication, DOI, etc. In order to collect raw data, rule-based metadata extraction has been proposed. For example, Guo and Jin [22] propose a rule-based framework for metadata extraction from scientific papers. The framework utilizes format information such as font size and position to guide the metadata extraction process. They use header information as rules which contain author, title, abstracts located in the first page of the paper. Such rule-based raw data extraction techniques can achieve high accuracy in header information extraction.

2) Author Information Extraction: Most scholarly data visualization requires author information for analysis. Usually, an article can be used to trace user information. Author entity can be well profiled with article raw data including authors' name, affiliation, research interest, etc. Furthermore, based on the author information, the most important academic relationship, coauthorship, can be inferred, where two authors are considered to be connected if they have coauthored a common paper. Author entity is usually the basis of academic search engines and digital libraries. CiteseerX [23] proposes to infer author information including name, institution, affiliation, and email address from the PDF-converted documents. However, many aspects of author information are neglected if we merely focus on the article itself. Yao et al. [24] propose to build a semantic profile for an academic researcher by identifying and annotation author information on the Web including academic search engines and author homepage. Their approach has been proven to achieve a significant improve in identifying expert. Meanwhile, a scholar-centric academic search engine Aminer [25] is built based on this approach.

3) Citation Information Extraction: Another important relationship in scholarly big data is the citation relationship. If paper A appears in the reference list of paper B, it means that paper B cites the paper A. The citation relationship is a direct relationship which has been extensively used to quantify academic impact as well as to trace the origin of new knowledge. Since the citation information is located in the ”Reference” section of a paper, citation information extraction requires accurately locating a section of a given paper with indicator ”References”, ”Bibliography”, or ”Sources”. Usually, these section can be found at the end of a given paper. ParsCit [26], FLUX-CiM [27], and CRF-based system [28] are the three widely-used tools to extract citation information.

4) Other Information Extraction: There are various of other entities in the scholarly big data including the institution, venues, as well as the content information like algorithms and figures. The Microsoft Academic Search [21] proposes to collect venue entities from a few semi-structured websites from Bing which are used by conference organizers to post conference information. They further conflate the venue event including conference instances and series across different websites with various signals inferred from the semi-structured data including full name, year, location, etc. Meanwhile, the journal and institution entities are mainly extracted from the in-house knowledge base, for example, ACM digital library.

Apart from these mentioned entities, there are various other entities and knowledge which need to be further investigated. For example, figures, table, algorithms, and acknowledge are

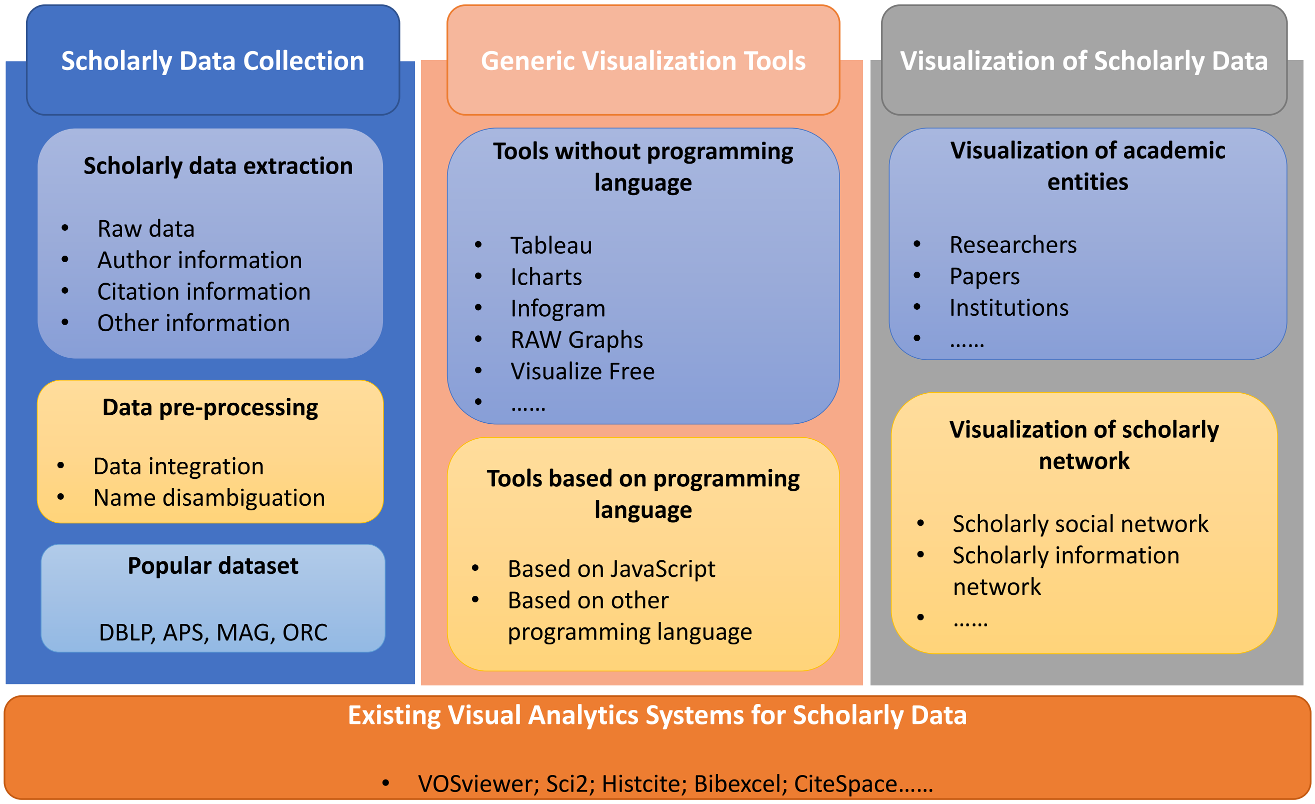

Scholarly Data Collection
Scholarly data extraction
Raw data
Author information
Citation information
Other information
Data pre-processing
Data integration
Name disambiguation
Popular dataset
DBLP, APS, MAG, ORC
Generic Visualization Tools
Tools without programming language
Tableau
Icharts
Infogram
RAW Graphs
Visualize Free
……
Tools based on programming language
Based on JavaScript
Based on other programming language
Visualization of Scholarly Data
Visualization of academic entities
Researchers
Papers
Institutions
……
Visualization of scholarly network
Scholarly social network
Scholarly information network
……
Existing Visual Analytics Systems for Scholarly Data
VOSviewer; Sci2; Histcite; Bibexcel; CiteSpace……

Fig. 1. Framework of scholarly data visualization

also important to analyze and visualize scholarly big data. Figures are usually to vividly present the architecture of a paper and most results are presented with figures or tables. Algorithms are used to describe the core process and idea of the proposed method. The Acknowledge section usually indicates scholars who also contribute to the paper and grants information. The extraction of these scholarly data types can support the scholarly data visualization with additional information. However, existing work in collecting these data are limited to raw data extraction [29].

B. Data Pre-processing

Data cleaning has always been an important process in big data area. Before data visualization, we need to clean the acquired dataset. Data pre-processing techniques have been extensively discussed in big data related article [30]. In this section, we mainly discuss the two fundamental issues in scholarly data pre-processing including data integration and name disambiguation.

1) Data Integration: With the development of academic social networks and academic search engines, scholarly information can now be inferred from divers data sources. For example, we extract a scholar's personal information including institution, teaching, research interest, and work experience from his/her personal information; The publication list and citation information of the given scholar can be extracted from the Google Scholar search engine; Linkin provides the scholar's location and friends. We can infer comprehensively data of a given scholar by integrating information from various data sources. However, the data is usually dynamic and uncertain. A scholar may change his/her institution as well as research interest. It takes time to integrate all the information together and the accuracy should also be considered.

big data

2) Name Disambiguation: One important challenge for processing scholarly big date is the author name ambiguity [31], [32], [33]. On the one hand, scholars may use different name (full name or short name) in different papers within the same digital library. On the other hand, different digital libraries may adopt different name styles. These factors may result in the author name ambiguity problem. What's worse, two scholars may share a common name. For example, there are more than seventy scholars with the name of "Wei Zhang" in the DBLP digital libraries. Thus, it is critical to conduct author name disambiguation before data visualization.

Existing author name disambiguation mainly adopt additional information from raw data beyond author name to tackle the issue of author disambiguation. For example, Schulz et al. [32] propose to take advantage of citation network covered by the Web of Science for disambiguating author names. They propose to merge common name papers with a pairwise publication similarity metric based on common authors, self-cation, shared references, and citations. Liu et al. [34] design an author name disambiguation system for PubMed with a machine-learning method by scoring the features for disambiguating a pair of papers with ambiguous author names.

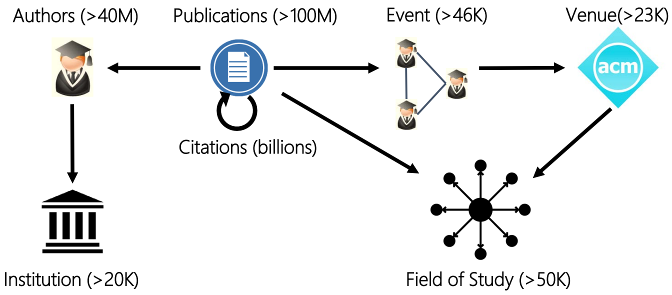


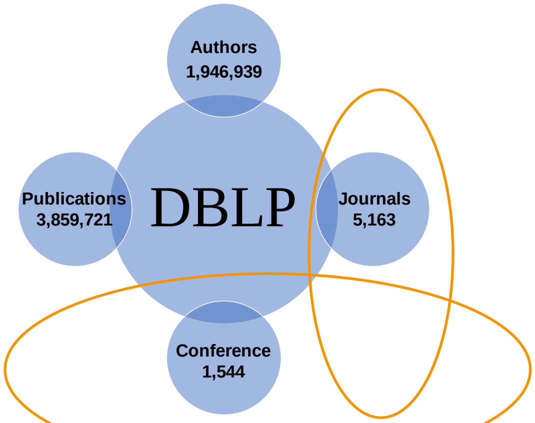
Fig. 2. Entities in scholarly dataset.

TABLE I
STATISTICS OF POPULAR SCHOLARLY DATA SET.

Name	Size	Field	Citation
DBLP	382 MB	Computer Science	No
APS	1.2 GB	Physics	Yes
MAG	29.8 GB	Multidisciplinary	Yes
ORC	7 GB	Computer Science/Neuroscience	Yes

C. Popular Dataset

In modern academia, more and more scholars are willing to share their datasets. Many academic search engines, digital libraries, and research institution have made their scholarly dataset available to help explore science itself. Among them, DBLP, APS (American Physical Society), MAG (Microsoft Academic Graph), ORC (Open Research Corpus) are widely used. These datasets contain the basic information of a given scholar extracted from the publication information of a certain discipline or various disciplines. A brief overlook of DBLP dataset can be seen from Fig. 3. The basic information of these datasets can be seen in Table I.



图中数据错误 journals和conference数据反了

Fig. 3. Entities in DBLP digital library.

1) DBLP: DBLP digital library is an on-line reference bibliographic information focusing on the field of computer science area. It contains over 3.8 million publications, published by more than 1.9 million authors. It indexes more than 32,000 journal volumes and more than 31,000 conferences or workshop proceedings. More importantly, it has done author name disambiguation for every scholar. They have gave an additional number for scholars with the same name, e.g, “Wei Wang 0077”. However, One of the limitation of the data set is that it does not contains the citation information. The data set can be downloaded directly from the link http://dblp.dagstuhl.de/xml/ in XML data format.

dataset？

2) APS: APS data set is provided by the American Physical Society, which is a corpus of Physical Review Letter, Physical Review, and reviews of Modern Physics. It is composed of more than 450,000 articles dating back to 1893. Scholars may now request access to the APS dataset from the link https://journals.aporg/datasets with accepting the terms and conditions governing the use of the data sets. The dataset contains two sub datasets including article metadata and citing article pairs. The article metadata contains the raw data of all APS journal articles including DOI, venue, page number or the first and the last page, title, authors, affiliations, etc. The citing article pairs dataset contains pairs of APS articles that cite each other. The dataset are in the format of comma-separated values (CSV).

3) MAG: The MAG dataset is a heterogeneous academic graph containing scientific publication records and the citation relationships. It mainly consists of six entity types including authors, papers, institution, journals, conferences, and field of study. The data set now can be accessed via the Microsoft Cognitive Services Academic Knowledge API with the link http://research.microsoft.com/en-us/projects/mag/. The biggest advantage of this dataset is that it contains the papers from all the fields. However, name disambiguation is necessary before utilizing this dataset for further analysis.

4) ORC: The ORC dataset is provided by the Semantic Scholar project. It contains more than 7 million paper from the fields of Computer Science and Neuroscience. It contains the raw data including paper title, abstract, keywords, paper url, author name, in and out citation, publication date, and venue.

most of them have?

III. GENERIC VISUALIZATION TOOLS

With the development of visualization tools, most of them has integrated many useful functions (liked data preprocessing, visual analysis) into a library or common plugins. Therefore, it enables users to simplify the procedure of data visualization by using programming language to invoke the function or use the pre-integrated function directly in tools. Visualization tools also give users ability to transform every element of the data into interactive charts and pictures [**?**]. Based on these intuitive charts and pictures, the analysis of the generated charts and graphs is more effective than the raw data. Data visualization is frequently used in BI (Business Intelligence), scientific visualization, information visualization and visual analysis. Therefore, when the researchers handled the large, complex data sets, it demands the traditional data mining techniques with high levels of data processing. According to user’s preference, we divide the tools into two categories, visualization tools without programming language and tools with programming language.

并列

A. Tools without programming language

1) Tableau: Tableau is a desktop software for business intelligence and analytics. Tableau can link to the data files

并列

on both local and server, and it supports a variety of data file formats such as txt, excel, csv. It also has a number of database interfaces for importing data from online servers, liked Oracle and MySQL. The updated data in the server can synchronize to the local automatically. Tableau uses a novelty operation method. It extracts the header of each item in the linked or imported data set automatically. The graph is immediately generated when users drag and drop these header into row and column and choose a chart type. The map function in Tableau is easy to use, such as the geographic information in users' data set can be marked automatically and displayed on a map when use the map function. It is suitable for users to analyse the data geographically. In Fig. 4, we use the map function in Tableau to visualize the data into a world map. Kale and Balan draw some charts to analyse the job vacancy in New York by using Tableau [35]. Tableau contains several features: 1) It is easy to visualize the data, users only need to choose a chart type and drag the header after importing the data. 2) These flexible, interactive graphs allow users to analyze the characteristics of their data from various perspectives. 3) The public edition is for free and it is commercial for using desktop edition, server edition, web-based edition, and even mobile application.

2) ICharts: ICharts[1] is a commercial web-based application that integrates the optimized API connector for NetSuite, Salesforce, Google Cloud Platform, and many others database platforms in the official. It is mainly used for BI. ICharts can combine data from CRM (Customer Relationship Management), ERP (Enterprise Resource Planning), even on-premise data stores that can help users take comprehensive analysis of the data. ICharts declares itself as a real-time business intelligence tool because the databases it linked are automatically updated, and it provides a number of chart types for users to visualize their data, each type of the chart can be fully customized. Some features of ICharts includes: 1) It connects to the real-time database, avoiding the secondary update of the data. 2) ICharts can take visual analysis automatically and build report of user's datasets periodically. The report is easy for sharing, thus the other users can take visualization analysis individually. 3) ICharts can combine multiple different types of dataset into the dashboard by creating fully customized interactive charts.

3) Infogram: Infogram[2] is a web-based application for making graphs and charts about informations, and this tool has a quick response and it can complete the data visualization quickly. Registered users could upload their own data file (.xls, .csv, .xslx) to the website, as well as importing data on
并列 GoogleDrive, Dropbox, OneDrive or JSON feed. The problem of Infogram is that the project is created by a public URL, thus the privacy of the data is in public. If users want to protect the privacy of their data, becoming a paid member is the only way. Infogram has opened its global sources which include all public themes and charts created by the other users to let users share their inspiration to each other. Infogram also enhances the function of sharing, so that users could embed their charts on webpage by using the codes which are automatically generated or shared by URL and Email. This application is easy for users to visualize their data. A game leaderboard is embedded into the game using Infogram to enhance the game players experience that they can see their progress and their performance in the game when compete with other players [36]. In Fig. 5, we create a graph based on Infogram. It shows all country names in the data, and the size of names are represented the scale of the graduate students and their supervisors. Infogram contains several features: 1) It shows a friendly users experience, users can communicate with the technical staff online, which make their works easier. 2) Public chart type library shared by other users is a good place to share charts and get original inspirations from others production. 3) It also provides a real-time data processing, and supports multi-terminal display. 4) The uploaded data in this tool's online database is public unless the user upgrades to a paid member then he can make sure its privacy.

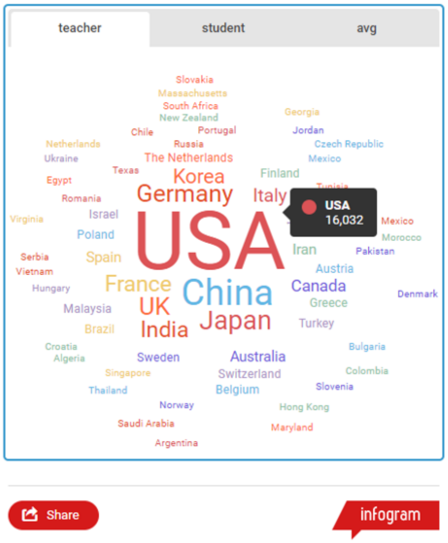
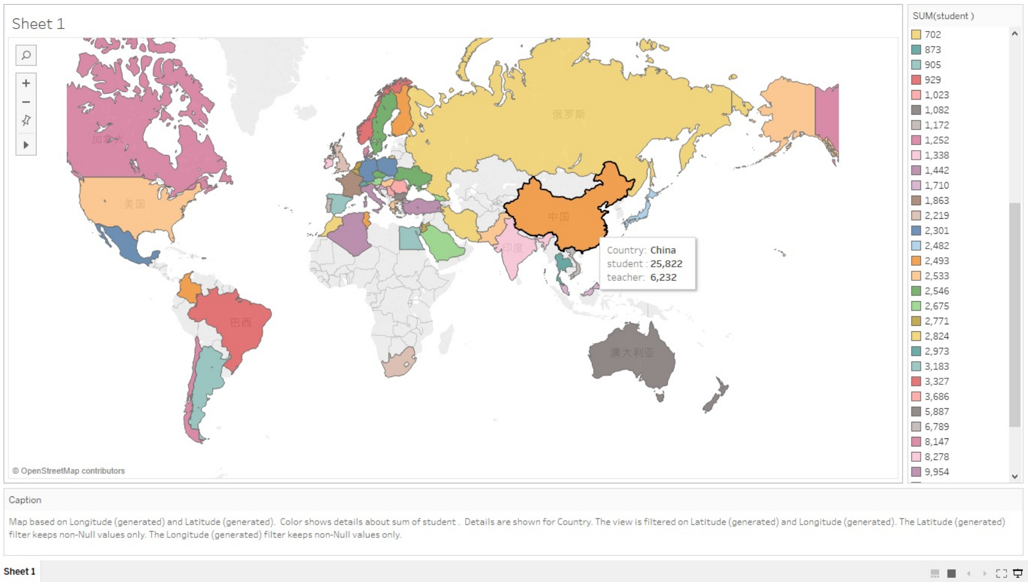


Fig. 5. The size of the countries' name in this graph shows the number of teachers and students, it is scaled based on the quantity, visualized by Infogram.

4) RAW Graphs: RAW Graphs[3] is an open web-based tool that can be used directly without registration. It supports the data format such as .tsv, .dsv, .csv, .JSON or .xls file, even the online data with a public API or from a public cloud platform. RAW Graphs processes the data only used the web browser in the local but not upload it to server, that can ensured the data safety. This application offers users 21 kinds of chart models for their data visualization and also supports to create custom vector-based visualizations on top of the d3.js library by Bostock at local[4]. In Fig. 6, we visualize the data into a circular dendrogram. Users can choose a chart type and map the dimensions by dragging the visual variables of the selected layout into its attributes for convenient chart generating. They also can export the generated chart as a vector (SVG) or raster (PNG) image or embed their graphs into webpages by using

[1] http://icharts.net
[2] http://infogr.am
[3] http://rawgrphs.io
[4] https://github.com/densitydesign/raw

The can also



Country: China
student: 25,822
teacher: 6,232

SUM(student)
702
873
905
929
1,023
1,082
1,172
1,252
1,338
1,442
1,710
1,863
2,219
2,301
2,482
2,493
2,533
2,546
2,675
2,771
2,824
2,973
3,183
3,327
3,686
5,887
6,789
8,147
8,278
9,954

Caption

Map based on Longitude (generated) and Latitude (generated). Color shows details about sum of student . Details are shown for Country. The view is filtered on Latitude (generated) and Longitude (generated). The Latitude (generated) filter keeps non-Null values only. The Longitude (generated) filter keeps non-Null values only.

Sheet 1

Fig. 4. This graph shows the number of graduate students and their supervisors between 63 countries in a world map which is generated by Tableau. It mapped the students from different countries in different colors, furthermore the color of supervisors are the same with students.

codes generated in RAW Graphs automatically. RAW Graphs contains some features: 1) It is easy for users to visualize their data in charts. 2) The imported data is safe, because it is only processed by the web browser but not on the online server. 3) RAW Graphs is open for users to create new charts by d3.js, but not insert in web application for the customized charts.

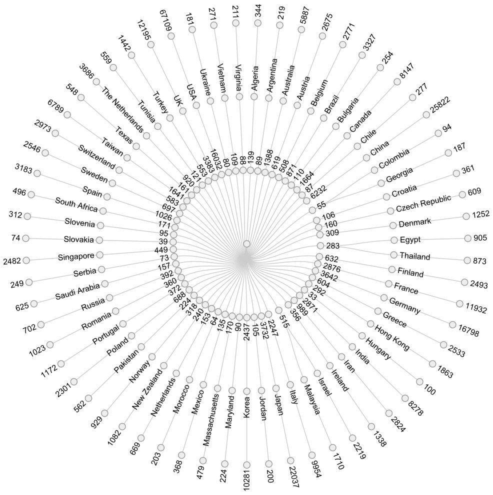


Fig. 6. This circular dendrogram shows the number of graduate students and their supervisors between countries, drawing by RAW Graphs. From inside to out side represents: number of supervisors, Country, number of students.

5) Visualize Free: Visualize Free[5] is a free and light web-based application which needs registration before using. Users can upload their data files with a 5MB file limit, and it supports Excel file (both .xls and .xlsx) and text file (.csv and tab-delimited .txt). Users can easily visualize their data into multiples of beautiful charts by dragging and dropping the data into correct layout for shaping the chart's dimension. The free visual analytics is provided that users can take a detailed analysis about the uploaded data. In Visualize Free, the uploaded data is private for users, and the generated charts can be shared by moving them into the shared folder or downloading them in .pdf, .xls, or .ppt format. Visualize Free contains some features: 1) It is easy for using and suitable for users to visualize the small amount of data into exquisite charts and visual analysis is free for analysing their data. 2) Common charts and maps are available for visualizing the data and it is easy for users to take visual analysis based on them.

B. Tools based on programming language

Visualization tools without programming language are easy for users to visualize their data into common charts or graphs by the guidance of these tools. Some visualization tools has opened their API that can enhance the function for chart plotting. It enables a flexible way to design their own style charts and graphs by handling the raw data with codes, but it will cost time for the basic users to command a new programming language. A part of the visualization tools are combined JavaScript, other part of tools uses the programming language like Python, JAVA, PHP, R. So we divide these visualization tools which are based on programming language into two parts: tools based on JavaScript and tools based on other programming languages. Users can choose a tool and access to its official documents and relevant tutorials to learn how it works and practice to visualize the data into charts by themselves. In TABLE II, we list a table for comparing the basic information of five visualization tools based on JavaScript. In TABLE III, we compare five visualization tools with other programming language.

并列

[5] https://www.visualizefree.com/

1) Tools based on JavaScript:

a) D3.js: D3.js [37] is a program of the open source JavaScript graphics library that combines HTML and CSS techniques, and the graphs it gathered are all in .svg format after visualizing the imported data [38]. D3.js completes the data visualization that it runs as a coded html file in browser platform under a server environment. D3.js is requested to run with its official document library for function invocation. In the website[6], D3.js provides plenty of examples (i.e. graphs, charts and their source codes) for users which can inspire them to design their own charts or use the examples directly. In Fig. 7, we combine line chart with histogram on a coordinate axis to show the number of graduated students and their supervisors between countries by using D3.js.

b) Chart.js: Chart.js[7] is also a program of the open source JavaScript graphics library. It uses canvas on HTML5, so the rendering performance is good in all modern browsers (above IE9). Chart.js can visualize the data into several common chart types by invoking the script language and its official chart library such as the color parsing library, the chart.js file, etc. [39], [40], [41]. Users should insert these libraries into the source code file by coding, then they can use the API from the library to set its parameter and process the chart [42].

c) FusionCharts: FusionCharts[8] is a commercial JavaScript library suite that combines the technologies including JavaScript and ActionScript3.0. It can run on multi-devices, browsers, and platforms. FusionCharts has more than 90 types of charts and over 1000 maps which are included all of continents. It supports processing .xml and .JSON files and exporting the generated charts as .jpg, 并列 .png and .pdf files [43]. The function of FusionCharts can be extended for embed the generated interactive charts to user's applications with several wrappers in official offered plugins, 并列 such as JSP charts, PHP charts, jQuery charts, Django charts, etc.

d) Flot Charts: Flot Charts[9] is focused on simple usage, attractive exterior and interactive charts. It is an extension for the jQuery library that supports HTML5 charts which is combined canvas and VML. This library separates the functional logic from HTML structure and uses DOM (Document Object Model) element to complete plotting. It contains ready-made component for the four basic chart types: charts-bar, line, point, and segment. Users can extend these charts easily and indefinitely with changing a wide variety of configuration parameters of them [44].

e) ZingChart: ZingChart[10] integrates Angular, React, jQuery, PHP, Ember, and Backbone in its declarative, efficient, and simple JavaScript library. ZingChart supports over 35 types of chart and model and allows users exports their 并列 visualization graphs in .png, .jpg and .pdf formats. It also offers integrated chart arrangement capabilities and has the basic drill-down function that users can select a data item within a chart [45].

data不可数

2) Tools based on other programming language:

a) Gephi: Gephi is a free open-source network visualization software which can implement network analysis. It is written by JAVA on the NetBeans platform. The typical feature of Gephi is that the process of spatialization can be presented vividly. The default layout algorithm of Gephi is ForceAtlas2, it is defined as a continuous force-directed layout algorithm [46]. Users can import their CSV data files or type their data on the spreadsheet of Gephi directly. The data file is divided into two parts: edges table and nodes table, thus users need pre-processing their data into two parts. The network supports the number of edges and nodes both up to 1 million, and the visualization is automatical when the data is imported. Users can choose a algorithm (ForceAtlas, ForceAtlas2, Fruchterman Reingold, Noverlap, etc.) to analyze the network, and export the generated network graph in .svg, .pdf or .pdf formats directly. 并列

b) NodeBox: NodeBox[11] is a free open-source and node-based Mac OS X application for creating 2D visuals (static, animated or interactive) that is based on Python programming codes. Users can combine the kinds of functionalities optionally by writing Python scripts [47]. NodeBox has integrated various document formats, such as users can manipulate the vector images in details by invoking the additional SVG library. It also supports the NodeBox Core Image library to create layered images, and exports the generated visuals to a PDF-document and the animations can be exported as QuickTime movies. 首字母大写？

c) ggplot2: Ggplot2 is an open source software package for graphs and visualization of statistical data creating. This package is based on the graphic grammar of R. It allows users to edit the plotting component at a high level of abstraction that is compared with the basic R graphs. This tool attends many details of plotting that makes it fiddly to plot charts or graphs. It is easier to produce complex multi-layer graphics by providing the powerful graphics models and a set of independent building blocks that users can plot a graph piece by piece by implementing the layered grammar of graphics (an extension of Hadley Wickham's grammar of graphics). It means that users will create a more complicated plotting by using faceting that users can concentrate more on graphs but less attention to the normalized programming language. Ggplot2 has its wiki in GitGub[12] for providing users a annual case study competition to show their graphs to others in a venue. It also highlights the large range of graphs which are created by using the richness of grammar. Users has the chance to be the developer when they are veteran to ggplot2, as a return, they can contribute codes back to ggplot2 [48], [49].

d) Processing: Processing [50] is an open source programming language based on JAVA that uses the simplified JAVA grammar. It provides users with a graphical interfaces and runs in the Java environment, and serves as a flexible software sketchbook. Processing is created for teaching fundamentals of computer programming within a visual context. The

[6] https://github.com/d3/d3/wiki/Gallery

[7] http://www.chartjs.org/

[8] http://www.fusioncharts.com/

[9] http://www.flotcharts.org/

[10] https://www.zingchart.com/

[11] https://www.nodebox.net/code/index.php/Home

[12] https://github.com/tidyverse/ggplot2/wiki

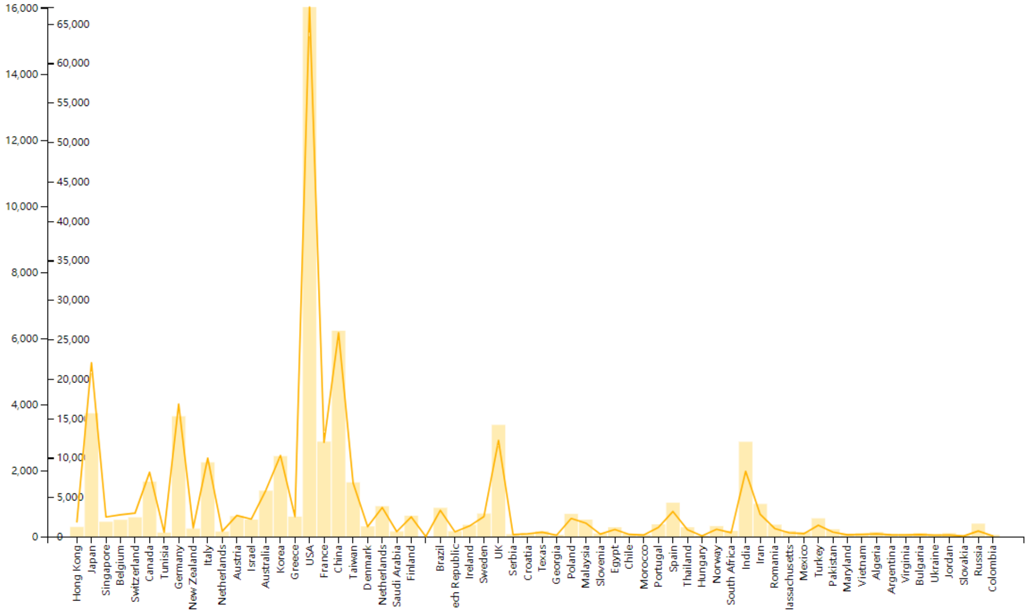


Fig. 7. This chart shows the number of graduated students and their supervisors between countries, of which the line chart shows the number of their supervisors which is mapped to the right side and the histogram shows the number of students which is mapped to the left side of Y axis.

TABLE II
BASIC INFORMATION OF VISUALIZATION TOOLS BASED ON JAVASCRIPT

Framework Name	Input Data Format	Rendered Charts By	Charts and Maps Type	Open Source	License From
D3.js	JSON, CSV, XML	HTML5 canvas, SVG and CSS	A powerful D3 gallery with multiple charts, graphs, and maps including the world map and the US maps.	YES	BSD-3
Chart.js	JavaScript API	Only HTML5 Canvas	8 chart types, including over 23 charts and graphs.	YES	MIT LICENSE
FusionCharts	JSON and XML	SVG, VML	90+ charts and graphs, 1000 + maps including all continents, major contries, and all US states.	YES	Free basic edition and advanced Commercial edition.
Flot Chart	JavaScript API	Only HTML5 Canvas	The charts of lines, points, filled areas, bars and any combinations of these charts. Not support maps.	YES	Free
ZingChart	JavaScript API	HTML5 Canvas, SVG and VML	Plenty of chart and graph types in its ZingChart gallery.Support almost every countries and areas.	YES	Free basic edition and advanced commercial edition.

language has high expansibility that users can write additional codes or integrate existing Java libraries to extend its function. The official website of Processing[13] is served as the online communication hub to host the relevant references, examples. In this website, it shows a public exhibition about kinds of project which are designed by Processing [51], [52]. Up to now, Processing is supported to alternative programming interfaces including JavaScript[14], Python[15], Ruby[16], it also can run on Android for users to create Android application[17].

首字母大写？

e) jpGraph: jpGraph[18] is an object-oriented library for creating graphs. The library is based on PHP5 (version above 5.1) and PHP7, and completely written by PHP and compatible with any PHP scripts. The commercial professional version of jpGraph supports the additional graph types: odometer, windrose, and barcodes.

并列

[13] Processing.org

[14] P5.js,https://p5js.org/

[15] Processing.py, http://py.processing.org/

[16] Ruby-Processing, https://github.com/jashkenas/ruby-processing

[17] http://android.processing.org/

[18] http://jpgraph.net/

TABLE III
BASIC INFORMATION OF VISUALIZATION TOOLS WITH OTHER PROGRAMMING LANGUAGE

Tools	Input Data Format	Language Based	Features	License From
Gephi	CSV, Excel file,	Java, OpenGL	Powered by OpenGL engine. Force-based layout algorithms shape the graph	GUN GPL
Nodebox 3	CSV	Python and Clojure	Intergrate all the functional parts in the nodes	GPL
ggplot2	R API	R	Plotting based on layers. Graphs composed of layers	GUN GPL V2
Processing	Multiples of formats are available in its library	Java, plugin for Python and JavaScript	Integrate the OpenGL engine. Over 100+ libraries offered to expand its usage	GPL,LGPL
JpGraph	CSV, From database liked mySQL	PHP	Tiny size of Generated images. Anti-spam images is supported. 3D effects supported	Free QPL, paid for commercial

IV. VISUALIZATION OF SCHOLARLY DATA

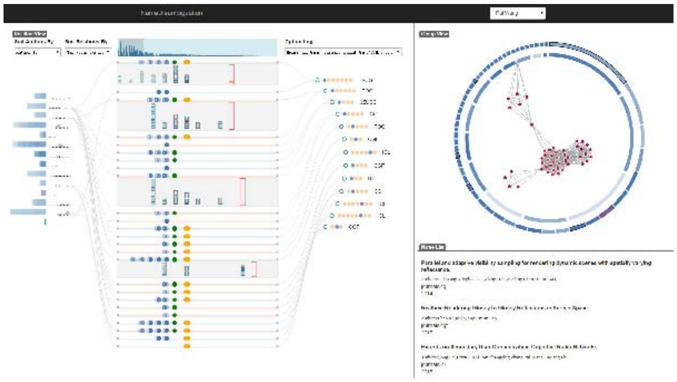
Scholarly data contains multiple entities, such as papers, authors, or journals. Based on it, the generated scholarly network, wherein nodes represent these academic entities and links represent the relationships such as citation, coauthorship, etc. This section describes the visualization techniques specifically designed for the simple attributes and heterogeneous networks of scholarly data.

A. Visualization of academic entities

Scholarly metadata is essential to carry out the efficient management of scholarly documents. Extracting the metadata of a paper such as title, authors, keywords, algorithms, figures, and tables is vital for developing scholarly services. In order to have a better understanding of topics or trends in science, there are some efforts to visualize the metadata for scholarly documents [53], [54], [55]. Such efforts become important pieces of scholarly data visualization for enabling expressing how academia develops.

1) Visualization of researchers: Current bibliographic databases usually provide the service of article searching and author retrieval. Using article metadata constituents such 并列 as authors' names, affiliations and research grants to build author profiles can help scientists obtain extensive author-related information. The well-organized information can help scientists to analyze scientific team formation and to have a comprehensive learning about the research interactions in the science of team science.

Name ambiguity, which means many-to-many corresponding relationships between persons and their names [56], is a common problem particularly common in Chinese names in the scientific venues. Shen et al. designs a novel visual analytics system for author name disambiguation called NameClarifier [55]. Not same with the traditional black box solution, NameClarifier changes the solution into a white box process with the visual method. Beyond that, it provides a way to guide the users to classify rather than give the classification results simply. The system consists of there parts: Relation View, Temporal View and Group View as well as a list for users to refer back to the original metadata (see Fig. 8). In addition, the system also provides a rich and practical user interaction, such as view correlation, iterative disambiguation, backtracking, query and so on. It ensures the effectiveness of person search and shed light on scientific community detection.



contain?

Fig. 8. Four interfaces for the NameClarifier, which contains: (A) The relation view: contrasts papers which contains ambiguous author names with confirmed authors to classify the ambiguous names easier; (B) The group view: supports the relation view; (C) The temporal view: verifies whether the specific paper can match into a confirmed authors publications; and (D) A list: contains all papers with ambiguous author names, for users to refer back to the original metadata [55].

2) Visualization of paper: In regards to aggregation levels, papers are the basic research unit, which can be aggregated into several higher research units, such as the author unit, the journal unit, the institution unit, or the country unit. It is an efficient way for researchers to have a comprehensive cognition about their research fields by reading academic papers. By analyzing the scientific literature, it becomes much easier to understand the trends of research or discover links and patterns among scientific documents. It also can help scientists keep track of the lasted developments and trends in the hot topic [57], [58], [59]. Recently, researchers have shown an increased interest in visualization of paper to provide multiple views of published articles. Their aim is to discover explicit or implicit relationships between them [?], [60], [61].

Matejka et al. [62] designs Citeology System (a portmanteau of citation and genealogy) to explore the relationship between publications. It is implemented as a Java applet and could be a useful tool for finding related work in a specific research field. The dataset of Citeology contains the papers as well as their citations published at CHI (ACM Conference on Human

Factors in Computing Systems) and UIST (ACM Symposium on User Interface Software and Technology) between 1982 and 2010. Citeology System presents users visualization results on the basis of a "family tree" of sorts, which can represent the generations of the referenced papers built upon on the target paper. Once a paper is selected, it shows the shortest path from hovered paper to the selected paper. Fig. 9 is the display of main components of the Citelogy interface.

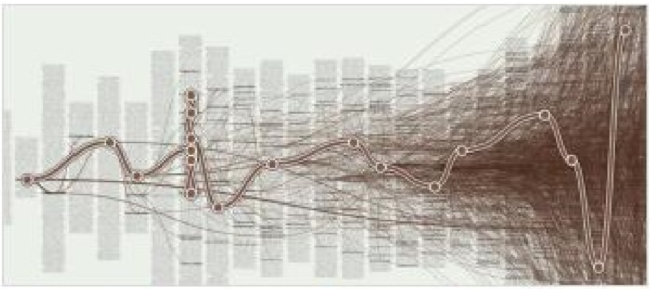


Fig. 9. The longest direct path between two CHI papers in the Citelogy which is an 18 generation gap [62].

Some interesting phenomena could be found through this straightforward visualization of citation generation. For example, Fig. 10 is the visualization of the main longest direct path between CHI papers. It turns out to be an 18 generation gap. Based on these discoveries, it broadens the collection of papers (especially topics) beyond the particular disciplines. The system can make the connection between researchers and the new conferences or topic areas they were not aware of previously.

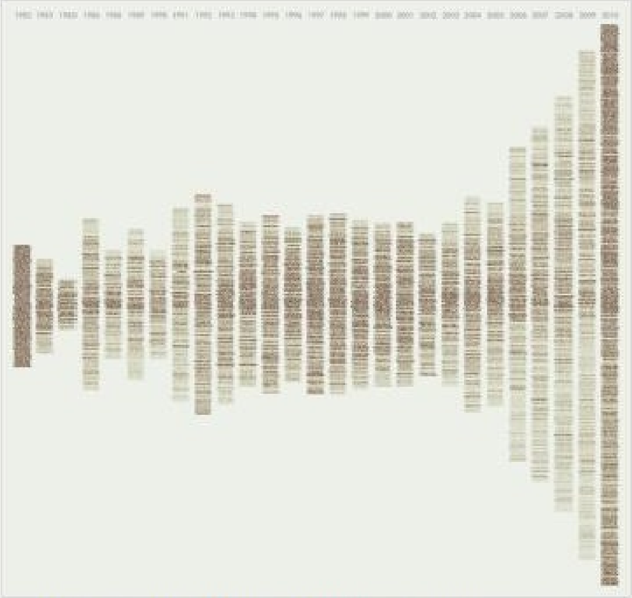


Fig. 10. The heatmap of click counts per paper during 3 week deployment in Citeology System [62].

There are numerous tools and techniques already for visualizing citation networks, but few focus on the impact among each nodes in the network. Maguire et al. [63] propose a solution to compare the impact between publications. The design can be divided into three interconnected parts: impact graphs, impact glyphs, and impact overviews (see Fig. 11). Impact graphs show the specific information of a focus paper, as well as its references and citations. Different patterns are used to identify the varying impact of publications. Impact glyphs are compact versions of the impact graphs to show the comprehensive importance of a paper. Finally, impact overviews position impact graphs for a subject area, author, institution and so on. It represents the core concepts of impact graph and impact glyphs. This part provides a way to layout the glyphs in 2D space. The design can outline mass summarizations of publication impact across a database. Jiang et al. [64] conduct the relationships among topics in three different research domains based on a hierarchical topic model. They also provide users a visualization interface and interactive operations to enhance their comprehension of connection among the cross domains, as well as the development trend of visualization. The model aims to represent the hierarchical structure and the similarities between topics. The interactive tool includes five views: word cloud (displays a topic), sankey diagram (represents the evolution of topics), scatter plot (presents the relative position of each topic), treemap (analyzes the relationship of topics), and stream diagram (represents the trends of a topic). Fig. 12 is the combination of the whole design. It enables users to explore topic mining results interactively and determine the proposed patterns, as well as draw a brief picture of visualization over the past 10 years.

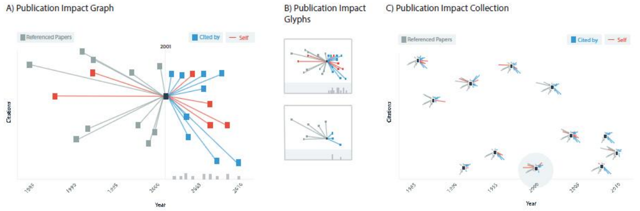


Fig. 11. An overview of publication impact, which contains: (A) Impact graphs: present the specific information of the paper including references and citations; (B) Impact glyphs: compact the impact graphs to show the comprehensive importance of the paper; (C) Impact overviews: position the impact graphs and impact glyphs for the related information [63].

3) Visualization of institution: Institution is part of the entities, which makes the scholarly data into a complex system. It encompasses various information including the name, the ranking, members, and location, etc. The relevant information of institutions such as members can be visualized through different techniques (refer to the above). For example, Acemap[19] is a website that can visualize affiliations onto a map (as shown in Fig. 13). Each node on the map represents an individual affiliation.

For each affiliation, it can show the collaboration network of the authors in the affiliation. For instance, Fig. 14 is the collaboration network of the California Institute Of Technology. When clicking on the specific institution on the network, users can also see the total number of publications and authors in the affiliation easily in the webpage.

[19] http://acemap.sjtu.edu.cn/app/affiliationMap/index.html

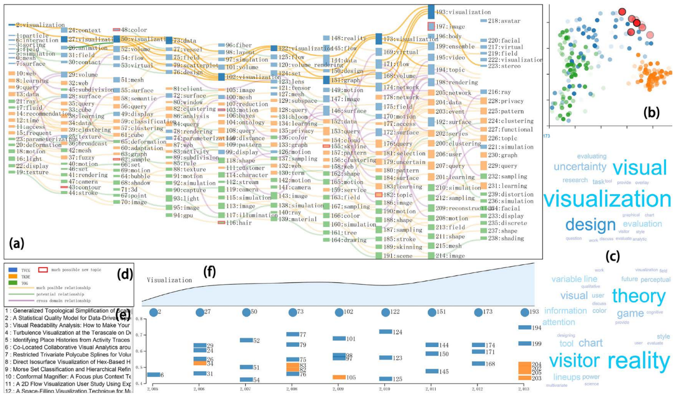


Fig. 12. An overview of cross-domain-research model, which contains: (a) Word cloud: displays a topic; (b) Sankey diagram: represents the evolution of topics, (c) Scatter plot: presents the relative position of each topic; (d) treemap: analyzes the relationship of topics), (e) Stream diagram: represents the trends of a topic [64].

为什么有的是逗号有的是分号

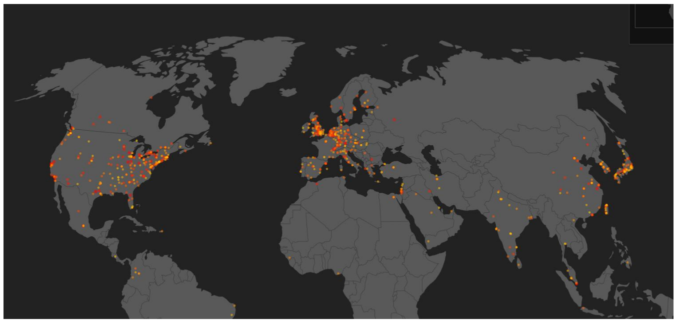


Fig. 13. An overview of visualizing the institutions contained in the datasets of Acemap. Each node on the map represents an individual affiliation.

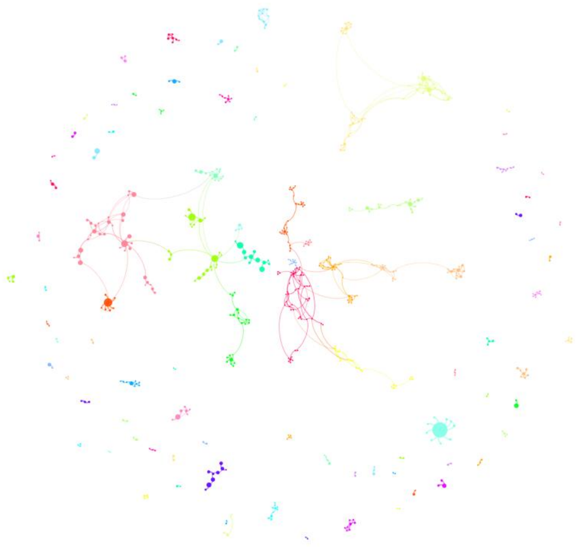


Fig. 14. Visualization of the collaboration network in the California Institute Of Technology in Acemap.

B. Visualization of scholarly network

One of the important review articles authored by Newman et al., distinguish four categories of real-world networks: social networks, information networks, technical networks, and biological networks [65]. Based on this, scholarly networks can be distinguished as social networks (e.g., collaboration networks) versus information networks (e.g., citation networks).

加逗号或者and

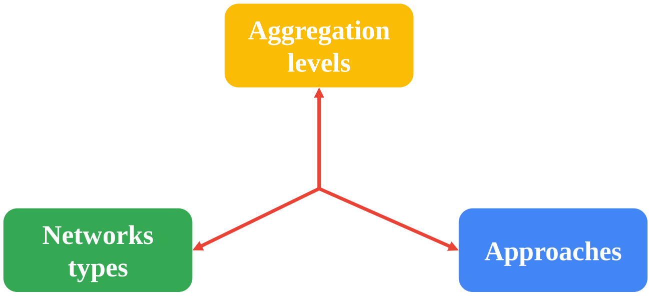


Fig. 15. Three-dimensional presentation of scholarly network-based bibliometric studies.

1) Visualization of scholarly social network: Recently, there has been renewed interest in the study of various types of scholarly networks. It provides scientists the opportunities to advanced the comprehension of the interactive research aggregates [66]. Fig. 15 presents the three-dimensional presentation of scholarly network-based bibliometric studies including approaches, networks types, and aggregation levels. Social network analysis are dividing into two typical ways: personal center network analysis (ego-centric analysis), and group network analysis (sociocentric analysis) [67]. On this basis, we show the current conditions of study in scholarly networks visualization from following two aspects: visualization of scholarly ego-centric network and visualization of scholarly sociocentric network.

Ego is the central node of ego-centric network, and alters are associated nodes. Depending on the distance from the alters to the ego, alters can be divided into 1-degree alter (nodes connected with ego directly), 2-degree alter (nodes connected with ego's alters), and so on. Ego-centric network focuses on the impact of the network on the ego. The networks have multidimensional attributes and change with time. How to display these characteristics that can make scientists understand, analyze and solve practical problems becomes a central issue of visual research.

As shown in Fig. 16, the scholarly tree is designed by Fung et al. [18] based on a botanic tree metaphor. It is a web-based, interactive visual interface. The different parts of the tree (e.g., leaves, branches and trunks) on behalf of different characteristics of the scholars' published papers. This project aims to show the details of collaboration information based on the unique tree features. The patterns of visualization encourage scientists to examine the personal career and also help to promote their self-development.

Botanical tree is focused on the association between the ego and 1-degree alters and emphasizes on the information representation. In order to figure out the evolution of personal

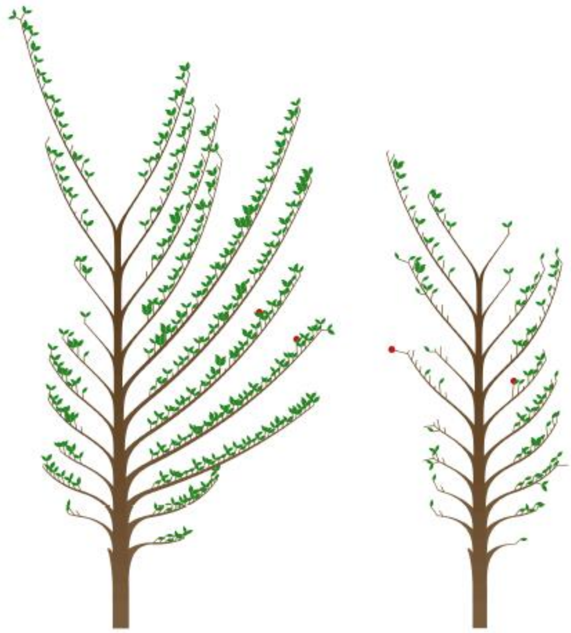


Fig. 16. The scholarly tree of two active researchers. Each branch on the tree encodes the publications of two years. The trees display their details of the publications between 1993-2013 for the left one and 1995-2012 for the right one [18].

networks from different perspectives, Wu et al. [68] design the egoSlider System. Fig. 17 uses different colors and shapes on the ring to show the relationship between the ego and the alters. It also presents the number of 1-degree alters (stripes) sustaining in a continuous-time period based on the glyph encoding. The glyph focuses on showing the overall statistical information, which is different from the line chart-based exhibition of the specific information about 1-degree alters and 2-degree alters. The design provides a wealth of interaction. It can change the glyph coding and the group method of lines flexibly according to user requirements.

EgoSlider can be applied to explore the DBLP collaboration network. It can help to discover a scientist's academic career, the change of the research direction, and the closeness of the collaboration with other scientists. Moreover, it also can make a major contribution for explaining some interesting phenomena by combing with the practical information found in this system. Compared with the baseline system based on the node link diagram, it will be more efficient to use egoSlider to complete the same task.

2) Visualization of scholarly information network: Co-citation count represents times of two papers cited jointly. Visualization of co-citation contributes much to the understanding of similarity in the document influence network. Early efforts in visualizing document co-ciation similarity employed cluster visualization. Noel et al. [69] provide examples of influence network visualization for both co-citation count and co-citation correlation and observe that the correlation-based visualization exhibits chaining effects.

Shen et al. [70] present a system called BiblioViz that gives exhaustive views of the bibliography data combining the features with capabilities of techniques in a unified fashion.

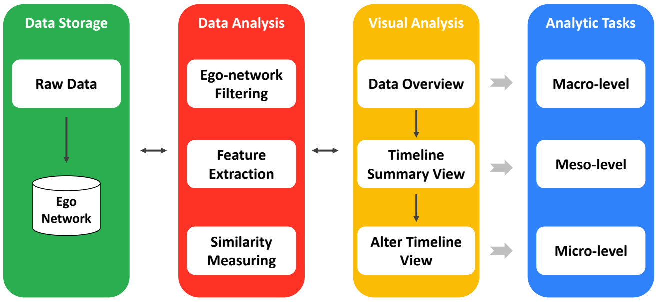
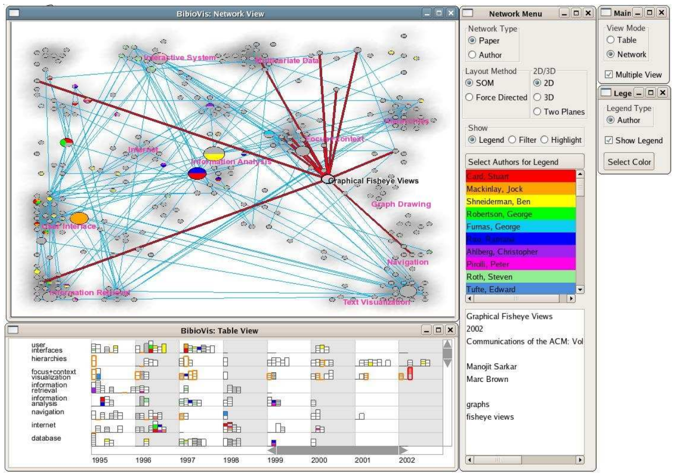


Fig. 17. The overview of the egoSlider visualization pipeline, which contains four parts: (A) Data storage: extracts the ego-network structures from the raw data and stores into MongoDB; (B) Data analysis: integrates several analytical methods to process the dynamic ego-network sequences; (C) Visual analysis: performs visual analysis of the data to let users interactively navigate; (D) Analytic tasks: addresses a different level of ego-centric analytical tasks.



Network Type ● Paper ○ Author
Layout Method ● SOM ○ Force Directed
2D/3D ● 2D ○ 3D ○ Two Planes
Show ● Legend ○ Filter ○ Highlight
Select Authors for Legend: Card, Stuart; Mackinlay, Jock; Shneiderman, Ben; Robertson, George; Furnas, George; Ahlberg, Christopher; Pirolli, Peter; Roth, Steven; Tufte, Edward

View Mode ○ Table ● Network
☑ Multiple View
Legend Type ● Author
☑ Show Legend

user interfaces, hierarchies, focus+context visualization, information retrieval, information analysis, navigation, internet, database
1995 1996 1997 1998 1999 2000 2001 2002

Graphical Fisheye Views
2002
Communications of the ACM: Vol
Manojit Sarkar
Marc Brown
graphs
fisheye views

Fig. 18. The double view mode layout of BiblioViz, which contains: (A) Network View: shows the citation network. (B) Table View: displays the paper details [70].

并列

The system includes five parts: Table View, Network View, Paper Details Panel and User Control Panel. The views are linked, thereafter, interaction with ones can affect the other manageably. With the help of BiblioViz, users can explore bibliography information easily.

Recently, Wu et al. [71] have designed PathWay to discover and understand the trends in the bibliographic data on the basis of individual professionals' co-authorship as well as the citation network of their publications. Implemented with JavaScript, HTLM5, and the Scalable Vector Graphics (SVG), the system displays researchers' career path in terms of their collaboration networks (see Fig. 19). The design can help users understand the social process better under the situation of challenges emerging.

V. Existing visual analytics systems for scholarly data

For conducting the research of scholarly data, researchers usually need to extract the based information of the large- scale scholarly data sets (such as DBLP, WOS, MAG, and so on) for analyzing the science of science. There are some of groups built a series of visual analytics systems for scholarly data,

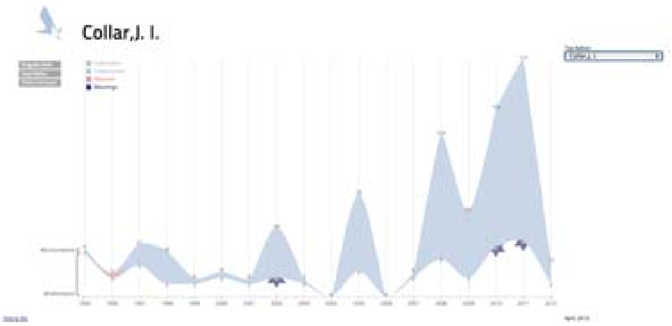


Fig. 19. The career path of scientist Collar in the Pathway which gives a quick and memorable overlook of Collars professional activities. In this figure, the altitude represents number of publications and the thickness of the path represents the number of acquaintances of Collar. Different color represents different type of acquaintances [71].

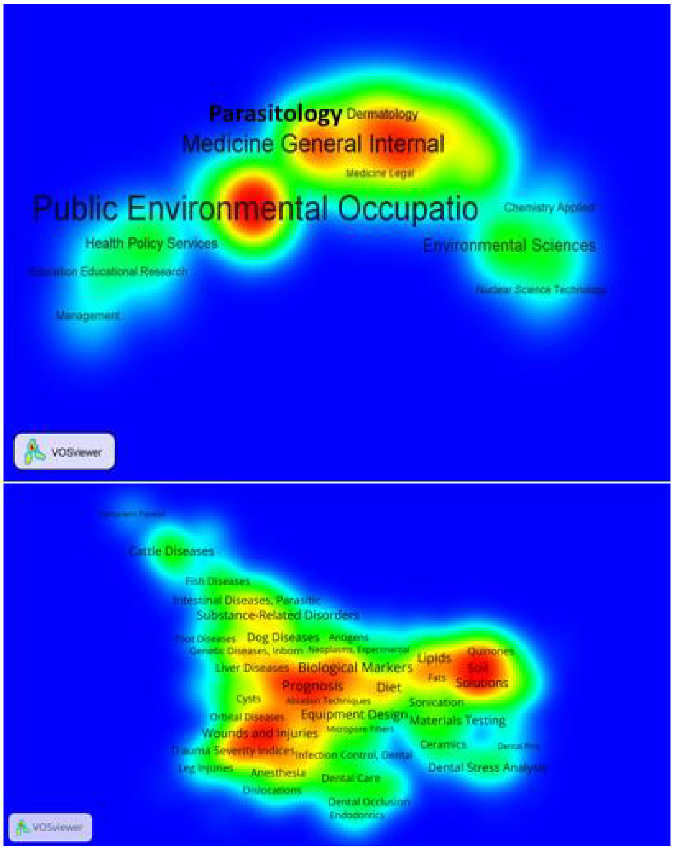


Fig. 20. The pubmed area studies (medical subject heading) with VOSviewer in a heat map. The above one shows the most important categories in WOS Categories which contains public environ-mental occupation, medicine general internal, and parasitology. The below one shows the most important area studies in Pubmed area studies which contains prognosis, wounds and injuries, biological markers, soil solutions [73].

并列

which has brought a great convenience for researchers that they can save times to concentrate on processing and analyzing the scholarly data by using these systems. In this section, we introduce 5 typical visual analytics systems for scholarly data below. In TABLE IV, we list a table for comparing the basic information of these five visual analytics systems.

A. VOSviewer

VOSviewer is a free visualization and analysis tool for constructing and visualizing bibliometric networks[20]. It facilitates the analysis of clustering solutions by visualizing the scholarly data into bibliometric networks. Actually, another tool named CitNetExplorer[21] is also used to cluster publications and analyze the resulting clustering solutions. CitNetExplorer is focused on the analysis at an individual publications level, while VOSviewer is focused on the analysis at an aggregate level. It supports users to create the network by importing the data files from WOS (web of science), Scopus, PubMed[22], RIS(Reference Information Systems), Pajek, and GML(Graph Modelling Language). VOSviewer can create biblometric networks and handle them by using the advanced layouts and clustering techniques. The visualization of these networks can be saved as bitmap or vector format [72]. Poreau et al. [73] mapped WOS Categories with VOSviewer in a heat map to show the most important categories. They also mapped the pubmed area studies (medical subject heading) with VOSviewer in a heat map to show the most important area studies (see Fig. 20).

B. Sci2

The Science of Science (Sci2 [74]) tool is an open source modular toolset which supports the temporal, geospatial, topical, and network studies[23]. Sci2 can visualize the scholarly data sets into kinds of networks. The visualizations of small data sets can be explored interactively, and the large-scale data sets are rendered into Postscript files that users can convert it into .pdf files and retrieve its information as a filter, such as searching the specific text in the visualization [16]. Osili et al. [75] geolocate donors and recipients based on their combined cities, states, and country information in existing data using the Bing! geocoder[24] available in the Sci2 tool. They also extract a bimodal network of the major donors and the six merged subsectors and visualize them by using the Sci2 tool (see in Fig. 21), and the detailed instructions are available in http://wiki.cns.iu.edu/display/SCI2TUTORIAL/Bipartite+Network+Graph.

C. Histcite

Histcite [76] is a software package that runs on Windows computers with the Internet Explorer. This system is used for scholarly data visualization and bibliometric analysis including the productive authors, the scales of journals, the frequency of words, the type of documents, and the ranking of countries and institutions. Histcite converts the bibliographies data sets into time-based network called historiograph, and makes it easier for users to observe and understand the main publishing

[20] VOSviewer,http://www.vosviewer.com/
[21] CitNetExplorer,http://www.citnetexplorer.nl/
[22] https://www.ncbi.nlm.nih.gov/pubmed
[23] https://sci2.cns.iu.edu/user/index.php
[24] http://wiki.cns.iu.edu/display/CISHELL/Bing+Geocoder

TABLE IV
BASIC INFORMATION OF VISUAL ANALYTICS SYSTEMS FOR SCHOLARLY DATA

System Name	Supported data file format	Features	Operating Environment
VOSViewer	Data from WOS, Scopus,PubMed, RIS, Pajek, andGML. and后边空格	Density and overlay visualizations. Create bibliometric networks based on co-authorship, bibliographic coupling, and co-citation networks, etc. Natural language processing techniques are available for creating term co-occurrence networks.	Windows, Mac OS X, Other systems with the support of Java 6 or later updates, the web client based on Java installed.
Sci2	TXT, CSV, Network data (in-memory graph/network object or network files saved as Graph/ML, XGMML, NWB, Pajek .net or Edge list format), Matrix data (Pajek.mat), In-memory database, Tree data (TreeML)	Visualize the scholarly data sets into kinds of networks. Perform different types of analysis with the most effective algorithms available. Access science datasets online or load their own.	Mac OS, Windows, Linux
Histcite	Data from WOS	Scholarly data visualization and various types of bibliometric analysis	Windows, based on the browser
Bibexcel	Plain text data from WOS, SCIE, DII (Derwent Innovations Index), Medline	Able to do various types of bibliometric analysis. Export its processed data into other visualization tools (Gephi, Pajek, VOSviewer, etc.) that can take a comprehensive visual analysis.	Windows
CiteSpace	Data from WOS	Visualize and analyze the patterns and trends in scientific literature	Windows, Require Java 8

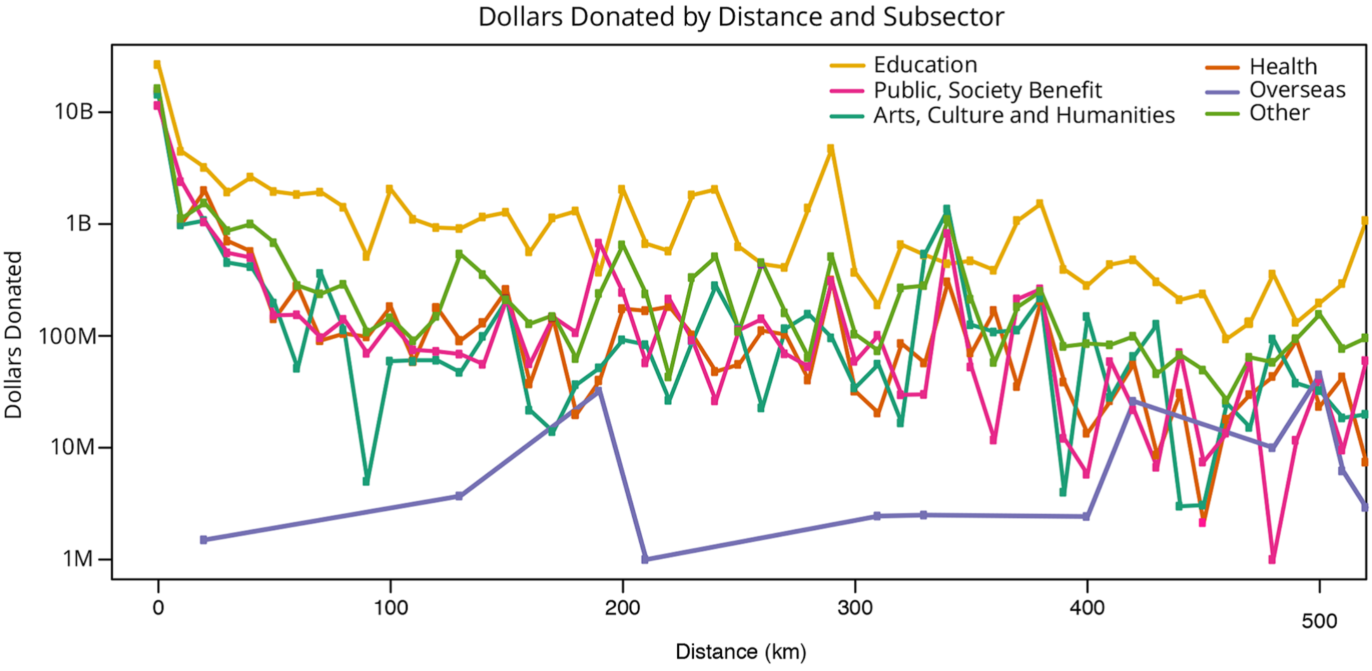


Fig. 21. Dollars of million-dollar-plus gifts (2000-2014) over distance for major subsectors rounded to the nearest 10 km as a form of binning presented by Sci2 [75].

events of the subject and the impact of the chronology in the networks [77], [78].

D. Bibexcel

BibExcel is a multifunctional bibliometric toolbox developed by Persson [79]. BibExcel is used to do various types of bibliometric analysis, such as citation analysis, cluster analysis, co-citation analysis, and so on. It allows users to analyze their scholarly data by selecting a catalogue which exist in their data (such as the authors) and adding it as a variable in a data matrix of the output files which is created by BibExcel. This software also allows users to export the files which include the data matrix and can import to another visualization tools (Gephi, Pajek, VOSviewer, etc.) for visualization [80].

E. CiteSpace

Citespace[25] is a free Java application that designed by Chen [20], [81]. It runs on the the java virtual machine so that it requires the Java runtime enviorment. The function

[25] Citespace,https://sites.google.com/site/citespace101/

of this software is to detect, visualize and analyze emerging trends and critical changes in scientific literature. It combines information visualization methods, bibliometrics with the algorithms of data mining in an interactive visualization tool to extract the patterns in citation data . Muthukrishnan et al. [77] use CiteSpace to map and cluster the top 20 countries' publication output of the British Journal of Cancer during 2005-2015 (see in Fig.22).

data后面有空格

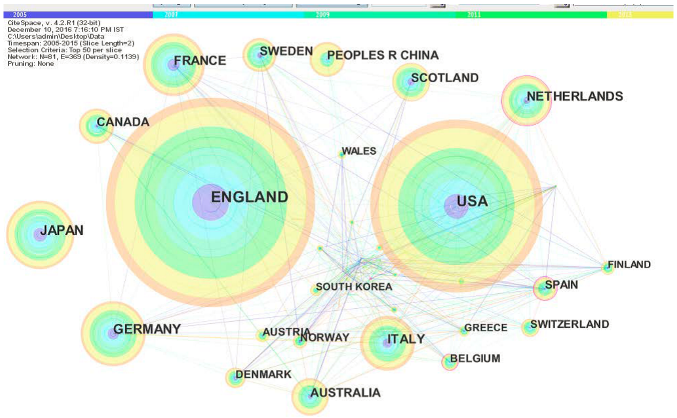

2005
2007
2009
2011
CiteSpace, v. 4.2.R1 (32-bit)
December 10, 2016 7:16:10 PM IST
C:\Users\admin\Desktop\Data
Timespan: 2005-2015 (Slice Length=2)
Selection Criteria: Top 50 per slice
Network: N=81, E=369 (Density=0.1139)
Pruning: None
FRANCE
SWEDEN
PEOPLES R CHINA
SCOTLAND
NETHERLANDS
CANADA
WALES
ENGLAND
USA
JAPAN
FINLAND
SOUTH KOREA
SPAIN
GERMANY
AUSTRIA
NORWAY
ITALY
GREECE
SWITZERLAND
BELGIUM
DENMARK
AUSTRALIA

Fig. 22. Mapping and cluster on publication output of top 20 institutions in Citespace [77].

VI. Open Issues and Outlook

The majority of the techniques and systems discussed in this survey specifically address one or two facets of scholarly data visualization. Benefiting from the development and popularization of these techniques, scientists have opportunities to study "science of science" from a new perspective. However, large data also brings numerous challenges to the field of scholarly data visualization. We discuss the issues that seem promising for further research as follows.

One of the main challenges in scholarly data visualization is information integration. Previous visualization tasks mainly focus on presenting a single relationship, for example, citation visualization or collaboration visualization. However, Scholarly data contains various entities including papers, authors, institutions, etc. There are diverse relationships among these entities. How to visualize different relationships in a single task is meaningful and challenging. At the same time, scholarly relationships are hidden in different data sources. For example, the collaboration information can be gained from the digital libraries i.e., DBLP, while the friend relationships are hidden in online social networks i.e., Facebook. How to integrate and information from different data sources is a promising open issue in scholarly data visualization.

Although quantities of visualization methods have proposed, some specific visualization techniques of scholarly attributes are encouraged to be improved. For example, very little attention has been paid to the visualization of the academic institutions. Such scholarly data often contains abundant information. How to mine the useful information through visualization is still a critical problem to be solved. The effectiveness of needs to be enhanced due to the increasing complex network structure as well. Another challenge is how to combine the visualization techniques with the analysis. The visualization theory and techniques on scholarly data visual analysis are not applied in practice.

VII. Conclusion

Scholarly big data brings a variety of opportunities and challenges to the field of scholarly data analysis. Nowadays, researches have realized the significance of applying the visualization technologies on different datasets to comprehend the science itself. Thus scholarly data visualization plays a key role in addressing the problems arising from large-volume, multi-variety, and important-value data. Since it make sense to concentrate more on this topic.

To provide new insights into scholarly data visualization, we review the emerging area of it in this survey. We present state-of-the-art scholarly data visualization techniques, with a focus on the visualization tools and analytic systems. According to the characteristics of scholarly data, visualization techniques for scholarly data are presented in two aspects: simple attributes and heterogeneous networks. These techniques are applied to various scholarly data analytic system, to map visualization and multivariate data visualization. A challenge is information integration of the complex scholarly data. Another challenge is how to combine virous visualization techniques with the analysis suitably. A future study investigating these questions would be very significant.

Appendix A
Proof of the First Zonklar Equation

Appendix one text goes here.

Appendix B

Appendix two text goes here.

Acknowledgment

The authors would like to thank...

References

[1] F. Xia, W. Wang, T. M. Bekele, and H. Liu, "Big scholarly data: A survey," *IEEE Transactions on Big Data*, 2017.

[2] Y.-R. Lin, H. Tong, J. Tang, and K. S. Candan, "Guest editorial: Big scholar data discovery and collaboration," *IEEE Transactions on Big Data*, vol. 2, no. 1, pp. 1–2, 2016.

[3] C. Caragea, J. Wu, K. Williams, S. Das, M. Khabsa, P. Teregowda, and C. L. Giles, "Automatic identification of research articles from crawled documents," in *Proceedings of the Workshop: Web-Scale Classification: Classifying Big Data from the Web, New York, NY*, 2014.

[4] S. Lehmann, A. D. Jackson, and B. E. Lautrup, "Measures for measures," *Nature*, vol. 444, no. 7122, pp. 1003–1004, 2006.

[5] W. Wang, J. Liu, S. Yu, C. Zhang, Z. Xu, and F. Xia, "Mining advisor-advisee relationships in scholarly big data: A deep learning approach," in *Digital Libraries (JCDL), 2016 IEEE/ACM Joint Conference on*. IEEE, 2016, pp. 209–210.

[6] M. E. M. Barak and J. S. Brekke, "Social work science and identity formation for doctoral scholars within intellectual communities," *Research on Social Work Practice*, vol. 24, no. 5, pp. 616–624, 2014.

[7] G. Cormode, S. Muthukrishnan, and J. Yan, "People like us: mining scholarly data for comparable researchers," in *Proceedings of the 23rd International Conference on World Wide Web*. ACM, 2014, pp. 1227–1232.

[8] C. D. Hansen and C. R. Johnson, *Visualization handbook*. Academic Press, 2011.

[9] T. L. Naps, G. Rößling, V. Almstrum, W. Dann, R. Fleischer, C. Hundhausen, A. Korhonen, L. Malmi, M. McNally, S. Rodger *et al.*, "Exploring the role of visualization and engagement in computer science education," in *ACM Sigcse Bulletin*, vol. 35, no. 2. ACM, 2002, pp. 131–152.

[10] B. H. McCormick, T. A. DeFanti, and M. D. Brown, "Visualization in scientific computing," *IEEE Computer Graphics and Applications*, vol. 7, no. 10, pp. 69–69, 1987.

[11] D. A. Keim, "Information visualization and visual data mining," *IEEE transactions on Visualization and Computer Graphics*, vol. 8, no. 1, pp. 1–8, 2002.

[12] J. J. Thomas, *Illuminating the path:[the research and development agenda for visual analytics]*. IEEE Computer Society, 2005.

[13] U. Demšar, K. Buchin, F. Cagnacci, K. Safi, B. Speckmann, N. Van de Weghe, D. Weiskopf, and R. Weibel, "Analysis and visualisation of movement: an interdisciplinary review," *Movement ecology*, vol. 3, no. 1, p. 5, 2015.

[14] G. Andrienko, N. Andrienko, P. Bak, D. Keim, and S. Wrobel, *Visual analytics of movement*. Springer Science & Business Media, 2013.

[15] S. Li, S. Dragicevic, F. A. Castro, M. Sester, S. Winter, A. Coltekin, C. Pettit, B. Jiang, J. Haworth, A. Stein *et al.*, "Geospatial big data handling theory and methods: A review and research challenges," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 115, pp. 119–133, 2016.

[16] R. P. Light, D. E. Polley, and K. Börner, "Open data and open code for big science of science studies," *Scientometrics*, vol. 101, no. 2, pp. 1535–1551, 2014.

[17] D. Keim, J. Kohlhammer, G. Ellis, and F. Mansmann, *Mastering the information age solving problems with visual analytics*. Eurographics Association, 2010.

[18] T. L. Fung and K.-L. Ma, "Visual characterization of personal bibliographic data using a botanical tree design," in *Proceedings of IEEE VIS 2015 Workshop on Personal Visualization: Exploring Data in Everyday Life*, 2015.

[19] C. Chen, "Visualising semantic spaces and author co-citation networks in digital libraries," *Information processing & management*, vol. 35, no. 3, pp. 401–420, 1999.

[20] ——, "Citespace ii: Detecting and visualizing emerging trends and transient patterns in scientific literature," *Journal of the American Society for information Science and Technology*, vol. 57, no. 3, pp. 359–377, 2006.

[21] A. Sinha, Z. Shen, Y. Song, H. Ma, D. Eide, B.-j. P. Hsu, and K. Wang, "An overview of microsoft academic service (mas) and applications," in *Proceedings of the 24th international conference on world wide web*. ACM, 2015, pp. 243–246.

[22] Z. Guo and H. Jin, "A rule-based framework of metadata extraction from scientific papers," in *Distributed Computing and Applications to Business, Engineering and Science (DCABES), 2011 Tenth International Symposium on*. IEEE, 2011, pp. 400–404.

[23] H. Li, I. Councill, W.-C. Lee, and C. L. Giles, "Citeseerx: an architecture and web service design for an academic document search engine," in *Proceedings of the 15th international conference on World Wide Web*. ACM, 2006, pp. 883–884.

[24] L. Yao, J. Tang, and J. Li, "A unified approach to researcher profiling," in *Web Intelligence, IEEE/WIC/ACM International Conference on*. IEEE, 2007, pp. 359–366.

[25] J. Tang, J. Zhang, L. Yao, J. Li, L. Zhang, and Z. Su, "Arnetminer: extraction and mining of academic social networks," in *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2008, pp. 990–998.

[26] I. G. Councill, C. L. Giles, and M.-Y. Kan, "Parscit: an open-source crf reference string parsing package." in *LREC*, vol. 2008, 2008.

[27] E. Cortez, A. S. da Silva, M. A. Gonçalves, F. Mesquita, and E. S. de Moura, "Flux-cim: flexible unsupervised extraction of citation metadata," in *Proceedings of the 7th ACM/IEEE-CS joint conference on Digital libraries*. ACM, 2007, pp. 215–224.

[28] F. Peng and A. McCallum, "Information extraction from research papers using conditional random fields," *Information processing & management*, vol. 42, no. 4, pp. 963–979, 2006.

[29] S. R. Choudhury, S. Tuarob, P. Mitra, L. Rokach, A. Kirk, S. Szep, D. Pellegrino, S. Jones, and C. L. Giles, "A figure search engine architecture for a chemistry digital library," in *Proceedings of the 13th ACM/IEEE-CS joint conference on Digital libraries*. ACM, 2013, pp. 369–370.

[30] X. Wu, X. Zhu, G.-Q. Wu, and W. Ding, "Data mining with big data," *IEEE transactions on knowledge and data engineering*, vol. 26, no. 1, pp. 97–107, 2014.

[31] C. L. Giles, H. Zha, and H. Han, "Name disambiguation in author citations using a k-way spectral clustering method," in *Digital Libraries, 2005. JCDL'05. Proceedings of the 5th ACM/IEEE-CS Joint Conference on*. IEEE, 2005, pp. 334–343.

[32] C. Schulz, A. Mazloumian, A. M. Petersen, O. Penner, and D. Helbing, "Exploiting citation networks for large-scale author name disambiguation," *EPJ Data Science*, vol. 3, no. 1, p. 11, 2014.

[33] M. Khabsa, P. Treeratpituk, and C. L. Giles, "Large scale author name disambiguation in digital libraries," in *Big Data (Big Data), 2014 IEEE International Conference on*. IEEE, 2014, pp. 41–42.

[34] W. Liu, R. Islamaj Doğan, S. Kim, D. C. Comeau, W. Kim, L. Yeganova, Z. Lu, and W. J. Wilbur, "Author name disambiguation for pubmed," *Journal of the Association for Information Science and Technology*, vol. 65, no. 4, pp. 765–781, 2014.

[35] P. Kale and S. Balan, "Big data application in job trend analysis," in *Big Data (Big Data), 2016 IEEE International Conference on*. IEEE, 2016, pp. 4001–4003.

[36] D. Honeyman and D. Walker, "Evolving customer engagement: Using mobile technology and gamification to improve awareness of and access to library services," in *Theta*, 2015.

[37] M. Bostock, "D3. js-data-driven documents," *URL: https://d3js. org*, 2016.

[38] F. Bao and J. Chen, "Visual framework for big data in d3. js," in *Electronics, Computer and Applications, 2014 IEEE Workshop on*. IEEE, 2014, pp. 47–50.

[39] N. Downie, "Chart. js— open source html5 charts for your website," *Chart. js*, 2015.

[40] C. Bergstrom, "Eigenfactor: Measuring the value and prestige of scholarly journals," *College & Research Libraries News*, vol. 68, no. 5, pp. 314–316, 2007.

[41] R. Murphy, "An employee performance simulation to aide in managerial decision making in a target driven work environment," 2016.

[42] R. Raghav, S. Pothula, T. Vengattaraman, and D. Ponnurangam, "A survey of data visualization tools for analyzing large volume of data in big data platform," in *Communication and Electronics Systems (ICCES), International Conference on*. IEEE, 2016, pp. 1–6.

[43] S. Nadhani and P. Nadhani, *FusionCharts Beginner's Guide: The Official Guide for FusionCharts Suite*. Packt Publishing Ltd, 2012.

[44] P. Pokornỳ and K. Stokláska, "Chart visualization of large data amount," in *Computer Science On-line Conference*. Springer, 2017, pp. 460–468.

[45] R. L. Rothfeld, "Advancing web-based dashboards: Providing contextualised comparisons in an air traffic discovery dashboard," 2015.

[46] M. Jacomy, T. Venturini, S. Heymann, and M. Bastian, "Forceatlas2, a continuous graph layout algorithm for handy network visualization designed for the gephi software," *PloS one*, vol. 9, no. 6, p. e98679, 2014.

[47] T. De Smedt, L. Lechat, and W. Daelemans, "Generative art inspired by nature, using nodebox," *Applications of Evolutionary Computation*, pp. 264–272, 2011.

[48] H. Wickham, "ggplot2," *Wiley Interdisciplinary Reviews: Computational Statistics*, vol. 3, no. 2, pp. 180–185, 2011.

[49] ——, "A layered grammar of graphics," *Journal of Computational and Graphical Statistics*, vol. 19, no. 1, pp. 3–28, 2010.

[50] C. Reas and B. Fry, "Processing. org," *Processing. org*, vol. 3, no. 06, 2012.

[51] ——, "Processing. org: programming for artists and designers," in *ACM SIGGRAPH 2004 Web graphics*. ACM, 2004, p. 3.

[52] ——, "Processing. org: a networked context for learning computer programming," in *ACM SIGGRAPH 2005 web program*. ACM, 2005, p. 14.

[53] 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*, vol. 23, no. 1, pp. 771–780, 2017.

[54] J. Whittaker, "Creativity and conformity in science: Titles, keywords and co-word analysis," *Social Studies of Science*, vol. 19, no. 3, pp. 473–496, 1989.

[55] Q. Shen, T. Wu, H. Yang, Y. Wu, H. Qu, and W. Cui, "Nameclarifier: A visual analytics system for author name disambiguation," *IEEE Transactions on Visualization and Computer Graphics*, vol. 23, no. 1, pp. 141–150, 2017.

[56] J. Kim, H. Kim, and J. Diesner, "The impact of name ambiguity on properties of coauthorship networks," *Journal of Information Science Theory and Practice*, vol. 2, no. 2, pp. 6–15, 2014.

[57] M. Callon, J. P. Courtial, and F. Laville, "Co-word analysis as a tool for describing the network of interactions between basic and technological research: The case of polymer chemsitry," *Scientometrics*, vol. 22, no. 1, pp. 155–205, 1991.

[58] Q. He, "Knowledge discovery through co-word analysis," *Library trends*, vol. 48, no. 1, p. 133, 1999.

[59] J. Law, S. Bauin, J. Courtial, and J. Whittaker, "Policy and the mapping of scientific change: A co-word analysis of research into environmental acidification," *Scientometrics*, vol. 14, no. 3-4, pp. 251–264, 1988.

[60] A. Khan, J. Matejka, G. Fitzmaurice, and G. Kurtenbach, "Spotlight: directing users' attention on large displays," in *Proceedings of the SIGCHI conference on Human factors in computing systems*. ACM, 2005, pp. 791–798.

[61] F. Heimerl, Q. Han, S. Koch, and T. Ertl, "Citerivers: Visual analytics of citation patterns," *IEEE transactions on visualization and computer graphics*, vol. 22, no. 1, pp. 190–199, 2016.

[62] J. Matejka, T. Grossman, and G. Fitzmaurice, "Citeology: visualizing paper genealogy," in *CHI'12 Extended Abstracts on Human Factors in Computing Systems*. ACM, 2012, pp. 181–190.

[63] E. Maguire, J. M. Montull, and G. Louppe, "Visualization of publication impact," *arXiv preprint arXiv:1605.06242*, 2016.

[64] X. Jiang and J. Zhang, "A text visualization method for cross-domain research topic mining," *Journal of Visualization*, vol. 19, no. 3, pp. 561–576, 2016.

[65] M. E. Newman, "The structure and function of complex networks," *SIAM review*, vol. 45, no. 2, pp. 167–256, 2003.

[66] E. Yan and Y. Ding, "Scholarly network similarities: How bibliographic coupling networks, citation networks, cocitation networks, topical networks, coauthorship networks, and coword networks relate to each other," *Journal of the American Society for Information Science and Technology*, vol. 63, no. 7, pp. 1313–1326, 2012.

[67] J. Scott, *Social network analysis*. Sage, 2017.

[68] Y. Wu, N. Pitipornvivat, J. Zhao, S. Yang, G. Huang, and H. Qu, "egoslider: Visual analysis of egocentric network evolution," *IEEE transactions on visualization and computer graphics*, vol. 22, no. 1, pp. 260–269, 2016.

[69] S. Noel, C.-H. H. Chu, and V. Raghavan, "Co-citation count vs correlation for influence network visualization," *Information Visualization*, vol. 2, no. 3, pp. 160–170, 2003.

[70] Z. Shen, M. Ogawa, S. T. Teoh, and K.-L. Ma, "Biblioviz: a system for visualizing bibliography information," in *Proceedings of the 2006 Asia-Pacific Symposium on Information Visualisation-Volume 60*. Australian Computer Society, Inc., 2006, pp. 93–102.

[71] M. Q. Y. Wu, R. Faris, and K.-L. Ma, "Visual exploration of academic career paths," in *Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*. ACM, 2013, pp. 779–786.

[72] N. J. van Eck and L. Waltman, "Citation-based clustering of publications using citnetexplorer and vosviewer," *Scientometrics*, vol. 111, no. 2, pp. 1053–1070, 2017.

[73] B. Poreau, "Scientometrics on public health research in iran: Increase of area studies despite embargoes? a review article," *Iranian journal of public health*, vol. 46, no. 3, p. 281, 2017.

[74] S. Team, "Science of science (sci2) tool," *Indiana University and SciTech Strategies*, 2009.

[75] U. O. Osili, J. Ackerman, C. H. Kong, R. P. Light, and K. Börner, "Philanthro-metrics: Mining multi-million-dollar gifts," *PloS one*, vol. 12, no. 5, p. e0176738, 2017.

[76] T. Reuters, "Histcite," http://interest.science.thomsonreuters.com/forms/HistCite/, 2014, [accessed 29-september-2017].

[77] M. Muthukrishnan and R. Senthilkumar, "Mapping of publications productivity on british journal of cancer during 2005-2015: A study based on web of science database," *Asian Journal of Information Science and Technology*, vol. 7, no. 1, pp. 42–46, 2017.

[78] E. Garfield, "From the science of science to scientometrics visualizing the history of science with histcite software," *Journal of Informetrics*, vol. 3, no. 3, pp. 173–179, 2009.

[79] O. Persson, "Bibexcel: a toolbox for bibliometricians," Tech. Rep.

[80] O. Persson, R. Danell, and J. W. Schneider, "How to use bibexcel for various types of bibliometric analysis," *Celebrating scholarly communication studies: A Festschrift for Olle Persson at his 60th Birthday*, vol. 5, pp. 9–24, 2009.

[81] C. Chen, "The citespace manual," 2014.

PLACE
PHOTO
HERE

Michael Shell Biography text here.

John Doe Biography text here.

Jane Doe Biography text here.