J Vis (2016) 19:561–576
DOI 10.1007/s12650-015-0323-9

REGULAR PAPER

Xinyi Jiang · Jiawan Zhang

A text visualization method for cross-domain research topic mining

Received: 13 August 2015 / Revised: 28 October 2015 / Accepted: 27 November 2015 / Published online: 5 January 2016

Abstract Cross-domain research topic mining can help users find relationships among related research domains and obtain a quick overview of these domains. This study investigates the evolution of cross-domain topics of three interdisciplinary research domains and uses a visual analytic approach to determine unique topics for each domain. This study also focuses on topic evolution over 10 years and on individual topics of cross domains. A hierarchical topic model is adopted to extract topics of three different domains and to correlate the extracted topics. A simple yet effective visualization interface is then designed, and certain interaction operations are provided to help users more deeply understand the visualization development trend and the correlation among the three domains. Finally, a case study is conducted to demonstrate the effectiveness of the proposed method.

Keywords Topic mining · Text visualization · Visual analysis

1 Introduction

Different research domains possess different research areas, methods, and models. Papers published in top journals of a research domain can reflect the developments and trends of the domain. These papers focus on the hottest topics of their research domain and are considered as important references to researchers. Some research domains have similar research topics with other domains, so the research findings in one domain may affect on other research domains. For example, some disciplines integrated two or more research domains, subsequently becoming a newly emerged branch of science. These research topics always belong to a promising research area worth paying attention to. Hence, this study aims to find out the trend of cross domain. The main objectives are as follows: (1) To use a text mining method to find out the topics of each research domain and to create a hierarchical and evolutionary relationship among these topics. (2) To use visualization tools to present the topics and the relationship among these topics, and to provide interactive operations to help users find the cross-domain topics and understand the trend of cross-domain research.

Many works in science metrics have been published to show the development of a specific domain. According to the research literature, they can be classified into several categories, such as: citation analysis, co-cited analysis, coauthor analysis, co-word occurrence analysis, and word frequency analysis. For nonexperts, these works provide entry points to a domain, and as means of gaining knowledge on

X. Jiang (✉) · J. Zhang
Tianjin University, Tianjin, China
E-mail: akanea@163.com

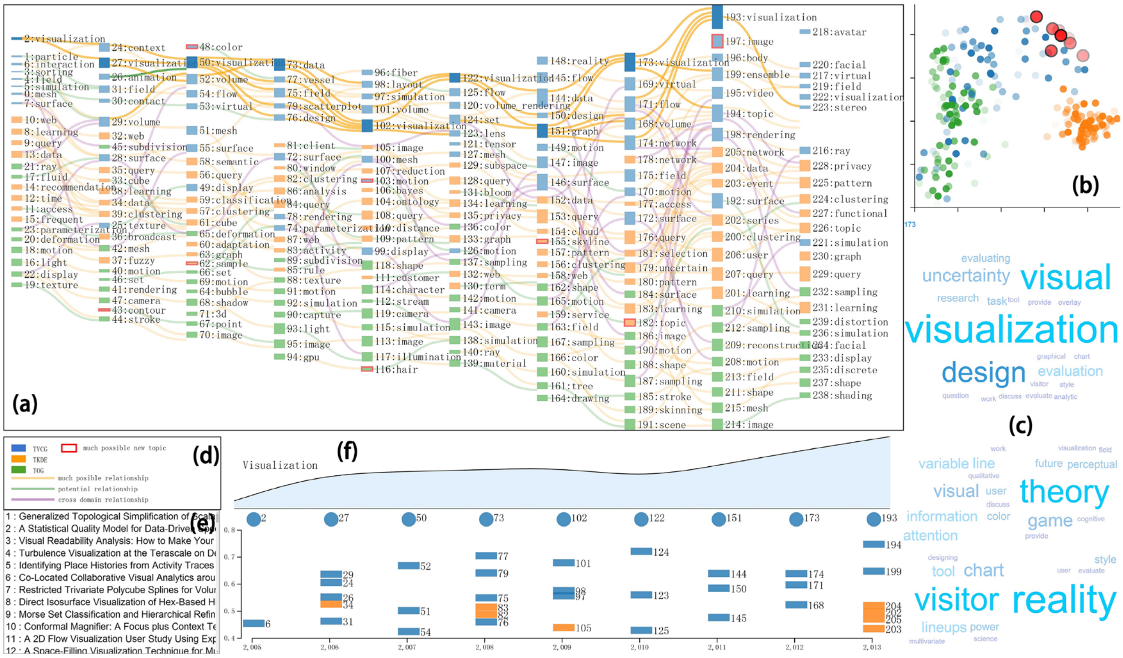
J. Zhang
E-mail: jwzhang@tju.edu.cn

both macro and micro levels. For experts, these works provide validations of perceptions and are means to quickly investigate trends and new information. However, only a few of these works consider the cross-domain research topics and design an interactive visualization interface. In the visualization domain, some works are about science research text topic mining, but few consider the hierarchy structure of topics and do not aim at finding out the evolution of cross-domain topics. In the present study, we do not focus on the text mining technique, instead, we use an existing algorithm to find out the relationship of cross science domains. We use the topic probabilistic model to find and construct a hierarchical structure of topics, analyze the evolution of topics, and design an interactive system for both experts and nonexperts.

Each research domain can be represented by the top journals of the domain since the papers reflect the latest status of the research domain and represent some new research directions. Our research domain is visualization which is a rising science with developing prospects that few people can clearly explain the exact position or the core technology of this domain. The top scholarly journal of visualization is IEEE transactions on visualization and computer graphics (TVCG), in which articles about visualization and computer graphics are published. The journal makes it more difficult to clarify the relationship between these two research domains. So we aim to find out and analyze the relationship among visualization, graphics, and data mining via a hierarchical topic model and design a visualization of evolution of the cross-domain knowledge.

The following sections explain the proposed method and visualization design (Fig. 1). The contributions of this study are as follows: first, a hierarchical topic model is used to construct a hierarchical and network structure of the cross-research domain topics; second, a visual design that enables users to interactively explore topic mining results and determine the patterns is proposed.

The rest of this paper is organized as follows. Section 2 reviews the related work on scientific topic finding, topic probabilistic model, and text visualization. Section 3 discusses a probabilistic topic modeling method, and the construction of a network and hierarchical structure of topics. An evolution analysis of topics is also performed. Section 4 details the proposed visualization design. Section 5 demonstrates the use of the proposed visualization tools in three case studies. Finally, Sect. 6 shows the conclusions.



1 : Generalized Topological Simplification of Scala
2 : A Statistical Quality Model for Data-Driven Spe
3 : Visual Readability Analysis: How to Make Your
4 : Turbulence Visualization at the Terascale on De
5 : Identifying Place Histories from Activity Traces
6 : Co-Located Collaborative Visual Analytics arou
7 : Restricted Trivariate Polycube Splines for Volur
8 : Direct Isosurface Visualization of Hex-Based Hi
9 : Morse Set Classification and Hierarchical Refin
10 : Conformal Magnifier: A Focus plus Context Te
11 : A 2D Flow Visualization User Study Using Exp
12 : A Space-Filling Visualization Technique for Mu

Fig. 1 Overview of the visualization tools. **a** sankey diagram presenting the topics for 10 years; **b** *scatter plot* showing the topics in 2D space; **c** word cloud of the selected topics and subtopics; **d** legend; **e** titles of the papers on the selected topics; and **f** stream diagram illustrating the topic trend with a scatterplot to show topic similarities

2 Related work

The following related works have been reviewed. These works can be classified into three areas: scientific topic finding, topic probabilistic model, and text visualization. In this section, some related research works that influence the present study are briefly introduced.

2.1 Scientific topic finding

In the scientometrics domain, many works exist about mapping knowledge domains. This domain is related to our work since we aim to find research topics of a domain or find the cross-point of some research domains. The main methods used in scientometrics include the following: citation analysis, co-citation analysis, co-word analysis, coupling analysis, word frequency analysis, and so on. Typical works with similar goals as those of the present study have been reviewed.

Chaomei Chen has performed excellent work on knowledge map not only through a mathematical method to construct a co-citation relationship of studies but also through a great visualization design of the co-citation network of a knowledge domain (Chen 2004, 2006). In his visualization design, certain indicators are proposed to present the turning points of a knowledge domain, such as landmarks, pivots, and hubs. Ding compares three methods in scientometrics (Ding and Chen 2014), where the result shows that hierarchical Dirichlet process (HDP), a generative probabilistic models, is more sensitive and reliable than traditional co-word and co-citation methods. Hence, the present study adopts a topic probabilistic model to detect and track topics. Boyack (2004) has generated the highest performing papers of PNAS for over a 20-year period via citation analysis. The papers describe changes and trends in the subjects that generate the highest effect within this domain. Newman (2004) has also performed coauthorship analysis, during which networks are constructed where nodes are scientists and two scientists are connected if the scientists have coauthored a paper. Through this work, scientific collaboration patterns are found. Mane (Mane and Börner 2004) has demonstrated burst detection algorithm, co-word occurrence analysis, and graph layout techniques to generate maps that support the identification of major research topics and trends. This work has determined burst words for 20 years and constructed a variation chart of ten typical words. Ginsparg et al. (2004), Landauer et al. (2004) and (Griffiths and Steyvers 2004) have used texts from papers as data sources and machine learning techniques to analyze, structure, and evolve academic literature. Romo-Fernández et al. (2013) describes an analysis of keywords which was aimed at revealing publication patterns in the field of renewable energy. The data and tasks in these works are similar with those in the present study, which serve as motivations to use a general text mining process with research papers. However, most works in this area provide insufficient interactive actions for users. The present work enables users to explore text mining results by providing certain interactive visualization tools.

2.2 Topic probabilistic model

Topic modeling provides methods to automatically organize, understand, search, and summarize large electronic archives. Via topic modeling, hidden themes infused in a text collection can be found; documents can be annotated based on themes; annotations can be used to organize, summarize, and search texts. Many topic probabilistic models have been proposed in recent years in the fields of machine learning and data mining.

A traditional vector space model uses a vector to represent terms and documents and uses relationship among vectors to correlate documents. However, a famous work called latent semantic analysis (Deerwester et al. 1990) has mapped terms and documents to a latent semantic space. The model avoids certain "noise" in the vector space of original documents and makes information retrieval more accurate. This model is a basic model on which many proposed topic probabilistic models are based.

Another widely used model is latent Dirichlet allocation (LDA), which is a model proposed by Blei et al. (2003). LDA is a generative model, in which each topic is a distribution over words, each document is a mixture of corpus-wide topics, and each word is drawn from one of these topics. This model uses posterior expectations that can visualize a hidden thematic structure in large corpora. LDA is a simple topic model used to find topics that describe a corpus; this model can be embedded in more complicated models, such as the correlated topic model (Blei and Lafferty 2007), dynamic topic model (Blei and Lafferty 2006), labeled topic model (Ramage et al. 2009) and hierarchical topic model (Mimno et al. 2007). The correlated topic model uses a logistic normal on the topic proportions to find patterns in how topics tend to co-occur. The dynamic topic model uses a logistic normal in a linear dynamic model to capture how topics change over time. Blei has indicated that time-corrected similarity is based only on topic proportions because time is factored out, given that topics associated to components differ every year.

Wang et al. (2013) proposed an algorithm to recursively construct a hierarchy of topics from a collection of content-representative documents. The framework is called constructing a topical hierarchy (CATHY). CATHY is employed in the present study to construct a hierarchical structure of topics, and this framework is combined with a dynamic topic model to capture trends and changes of the topics. Moreover, rich interactions are provided to allow users to analyze individual topics of interest.

2.3 Text visualization

Our work is related to research on text evolutionary visualization and multisource-data visualization and some visualization work in scientometrics.

The most famous evolutionary visualization is based on the river metaphor, ThemeRiver (Havre et al. 2002). In this work, each layer represents a word. The river flows from left to right representing the change of the word over time. We use a flow chart as a part of our visual analytic system. Many works are based on ThemeRiver, using topics instead of words, and each layer represents a topic. TIARA was proposed to combine word cloud and the stack graph to allow users to examine and analyze the topic content over time (Wei et al. 2010). Recently, TextFlow was proposed, introducing a seamless integration of visualization and topic mining techniques (Cui et al. 2011). In this work, the authors extracted three-level features: the topic evolution trend, the critical event, and the keyword correlation, and they made a visualization design to present the result. HierarchicalTopics integrated a hierarchical topic construct algorithm with ThemeRiver view for the users to discover interesting patterns in large textual collections (Dou et al. 2013). StoryFlow (Liu et al. 2013) used storyline to illustrate the dynamic relationships between entities in a story. ThemeDelta (Gad et al. 2015) applied sinuous, variable-width lines to show evolution of a text corpus on a timeline. Wu et al. (2014) combined Sankey graph with a tailored density map to analyze the diffusion of public opinions. Heimerl et al. (2015) featured an extended version of a streamgraph to depict clusters over time to visualize scientific literature. These works presented many good designs to represent text. However, the focus was not on cross-domain representation of topics.

Several works have focused on multisource-text data visualization. Dou et al. (2011) presented the analysis and visualization method for computing the distinction of a scientific document collection. They suggested a simple and intuitive visual design that enables users to explore the results. Oelke et al. (2014) used a topic coin view to represent the topics of different research domain. A work use multisource-data based on a widely used pairwise graph matching metrics, graph edit distance, and integrated multiple topic graphs together to support analysis (Liu et al. 2014). ParallelTopics utilized the parallel coordinate metaphor to present the probabilistic distribution of a document across topics. These works focused on multisources and not on evolution of topics (Dou et al. 2011).

Isenberg et al. (2014) described keywords, topic areas, and 10-year historic trends of visualization. The authors used keywords to find the communication of different visualization groups and provide an understanding of emerging new research trends. However, this work was a presentation of results with no provision for interactive user-operation. PNASLINK is a web-based literature-mapping system that can display most frequently co-occur terms (White et al. 2004). Morris and Yen (2004) used a crossmapping technique for visualizing multiple and overlapping relations among entity types in collections of journal articles. As mentioned in Sect. 2.1, Chen (2006) conducted visualization work on knowledge mapping. They provided a software, citespace2, for users to explore relationships and find important paper and new trends of research domains. Similar to this work, the present paper integrated text mining with some interactive visualization designs to identify the evolution of research topics in cross domain.

3 Data processing

3.1 Overview

Our work aimed to construct a visual analysis system that integrates computational methods with interactive visualizations. The system contains five key processing stages: data preprocessing, topic modeling, topic evolution constructing, topic mapping and visualizations. In Sect. 3 we focus on the first four, and in Sect. 4 we discuss visualizations.

The papers published in top journals reflect the research status of a research domain. Hence, we first collected text data from SCI database including abstract, keywords, and title of papers in three journals of different research domains. We then prepared the documents for topic models by lemmatizing and removing stop words. The cleaned data then underwent the topic modeling stage, which extracted topics from the document collection. The topic modeling we used is called CATHY, which is a hierarchical topic model. Using this model, we obtained tree-structure research topics of each domain. We then constructed the evolution relationship of topics. We constructed the relationship of topics mainly using similarities of estimate topical frequency. A topic mapping method using MDS is discussed in the last part of this section. After processing these data, we prepared a suitable data structure for visualizations. All these procedures are automated. Thus, our method is easy to apply in other data sets.

3.2 Data and topic model

3.2.1 Data

We used paper data from SCI database. Given that we would like to discuss the relationship between visualization, data mining and computer graphics, we chose the top journals of each research domain: IEEE transactions on visualization and computer graphics (TVCG), IEEE transactions on knowledge and data engineering (TKDE), and ACM Transaction on Graphics (TOG). We collected all the meta-data in SCI website from 2005 to 2014. TVCG, TKDE, and TOG have 1739, 1469, and 1683 papers respectively. The meta-data we downloaded from the website included the papers' titles, authors, affiliations, keywords, abstracts, the year of publication, and so on. Since the titles, keywords and abstracts are important parts and are good representation of the topics found in the paper itself so they can be a good outline of each paper, we used the paper's title, keywords, and abstract for topic mining. We managed title, keywords, and abstract of a paper into one document and marked the publishing year and the name of journal of each document. We obtained more than 4000 documents, which included text information from 2005 to 2014 of three journal papers. After we deleted the editor's lead papers which had no keywords and abstract, we used the rest of the papers' information to do topic mining.

Before applying the topic modeling, we performed standard text preprocessing. We lemmatized all words, added stop words, and removed the low-frequency words. Specifically, we removed all the words that only appeared one time since we considered that such low-frequency words were not important and insufficient to represent a topic. We believed that a word that strongly represents the point of view or reflect the topic of a paper should be mentioned more than once in the title, keywords, and abstract. The stop words we added included common stop words, such as functional words a, an, the, and then. Stop words also included adverb, common nouns with no meaning in the papers and several verbs, such as make, define, find, represent, guide, treat, and show. We also added method, algorithm, problem, framework, model, technique, result, and study as stop words. These words are high-frequency words in the papers regardless of the category of research domains. After preprocessing, we obtained a 14966-long vocabulary list of these three journals.

3.2.2 Hierarchical topic modeling

As mentioned in Sect. 2, many probabilistic topic models used in text mining. In this subsection, we focus on a hierarchical topic model that we used in our work. The model was proposed by Wang et al. (2013), and the name of the framework is called CATHY.

CATHY constructs a hierarchy where each topic is represented by a ranked list of topical words, such that a child topic is a subset of its parent topic. The researchers proposed an algorithm for recursively constructing a hierarchy of topics from a collection of content-representative documents. The term co-occurrence network was used for topic analysis instead of the document-term topic modeling. In the network, each node was represented by a word, and link between the nodes inferred that these two words appeared in same document.

The steps of CATHY are as follows:

1. Construct the term co-occurrence network G^o from the document collection where 0 refers to root. Set topic $t = 0$.
2. Cluster G^t into subtopic subnetworks $G^z, z \in C^t$, set C^t include all the children of topic t. And estimate subtopical frequency.

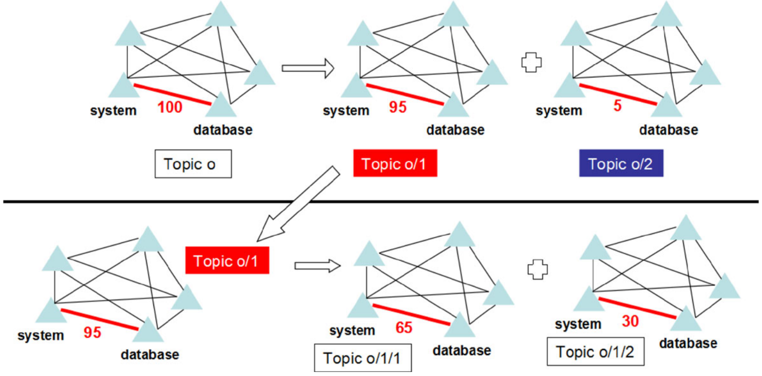


Fig. 2 Top-down recursion of CATHY

3. For each topic $z \in C^t$, extract candidate words based on estimated topical frequency.
4. For each topic $z \in C^t$, rank topical words based on topical frequency.
5. Recursively apply from (2) to (5) for each subtopic $z \in C^t$ to construct the hierarchy.

Figure 2 presents the top-down recursively inference process of CATHY. Each triangle in the co-occurrence network represents a word. The links and number represent the co-occurrence times of two words in one topic. Links would be partitioned into different subtopics and recursively apply them to construct the hierarchy.

We set some parameters for the algorithm. We set the level number to 2, which means a two-level (except root) tree-structure of topics can be obtained. Root means TVCG, TKDE, or TOG journals in a specific year. For the topic number, we set 8 to the first level and 4 to the second level. For example, the number all the papers published in TVCG in 2005 is 65. We managed title, keywords and abstract into 65 documents. Then we used these 65 documents as data source to run the algorithm. After algorithm processing, we finally get a topic mining results. A two-level tree hierarchy represented the topics. The root of the tree was the TVCG of year 2005. The first level had eight topics that reflected the major research themes of 2005. As for the second level, all the topics in the first level had four subtopics. Therefore, the second level included 32 topics in total that showed more details of the research topics of 2005 in TVCG. We repeated the algorithm for all three journals in all 10 years and finally obtained 30 tree-structure topic results for evolution construct and analysis.

We used this model on each year's documents and on all documents for 10 years as well. The topic model constructed no relationship at the same level or the evolutionary relationship; we detailed the construction of these two relationships in the next part.

3.3 Topic evolution construct

After the hierarchical topic model worked out, we obtained the tree-structure topic distribution for 10 years of these three journals. The topic number is 8 for the first level, and 4 for the second level. The number is an empirical value in our experiments. These topics have evolutionary relationship intrinsically, but we did not get it in a mathematical method. So we need to construct the relationship by an evolutionary model.

In order to get the evolutionary relationships, we need to judge the similarities between each topic. After some experiments, we choose the simple cosine similarity to measure topics' similarity.

3.3.1 Find new topic

Similar to knowledge map, our work also intends to help users gain understanding of a research domain. New topics in a research domain allow users obtain quick entry of their research in an unfamiliar domain. Identifying new topics in the past can help users understand the development of the research domain and summarize the rules and patterns for further research. We proposed a method to detect potential new topics and further combined the results with interactive visualization designs that we presented in Sect. 4 to help users find the real new topics.

For the first-level topics, we have all 240 topics from 2005 to 2014, so that is 24 topics each year. We then computed the cosine similarity of topical frequency of each word in adjacent years and obtained 9 similarity matrixes. For example, year 2005 and 2006 both had 24 topics, we counted the similarities of all the topics of 2005 and 2006 to obtain a 24×24 matrix M. Each row had the similarities of a topic of 2005 with all the topics of 2006. Each column had the similarities of a topic of 2006 with all the topics of 2005. M_{ij} represented the similarity of topic i of 2005 and topic j of 2006. The largest number of column j expresses the biggest similarity that topic j of 2006 with topics of 2005. There is an intuitive understanding that if the biggest number of column j remains small, then perhaps topic $j(j = 1\ldots$ column count $= 24)$ of 2006 is a new topic because of its difference from all the topics of last year. Based on the intuitive hypothesis, we use a simple method to judge whether a topic was a probable new topic.

We used a standard value to help determine a threshold for judging whether a topic is a new one or not. We named the standard N, which is a percentile. N was made equal to the amount of the largest number of column j that was greater than threshold and the amount of column number was divided. That is,

$N =$ sum(biggest number of column j that is greater than threshold)/sum(the column count). (j from 1 to column count).

For example, when we set N to 90 %, and the largest numbers of column j in 2005–2006 were.

0.577477399	0.527452224	0.434672896
0.690779346	0.600518303	0.572882437
0.362486444	0.484463906	0.711518011
0.673566666	0.818672777	0.646638084
0.473509978	0.622864326	0.612280624
0.631937352	0.854408554	0.677571549
0.676566291	0.338172149	0.527281327
0.506704687	0.389293723	0.64570253

The threshold can be set to 0.4 such that $22 \approx 90\ \% \times 24 = 21.4$ numbers greater. The N can be set by users; the greater the N has been set, the smaller the threshold becomes. Moreover, we used another fixed threshold 0.4 to judge if a topic is a new topic because sometimes the threshold computed by N would be too big, on this occasion, the topics are probably not new ones. Therefore, we used this fixed threshold as a filter.

3.3.2 *Evolutionary relationship*

After obtaining the hierarchical structure of topics, we needed to consider their evolutionary relationship. As mentioned, we obtained 30 tree-structure topic results after running CATHY. The 30 results belonged to three journals in 10 years. Each result of a year had 8 topics in the first level and 32 topics in the second level. We considered only the first-level topics that inferred big trends of journal topics.

The topic evolution model based on CATHY is shown as Fig. 3.

Variable x is a Boolean variable, which indicates whether the topics of two adjacent years have correlations. If x equals 1, then there is a relationship between these two topics, vice versa. The value of x

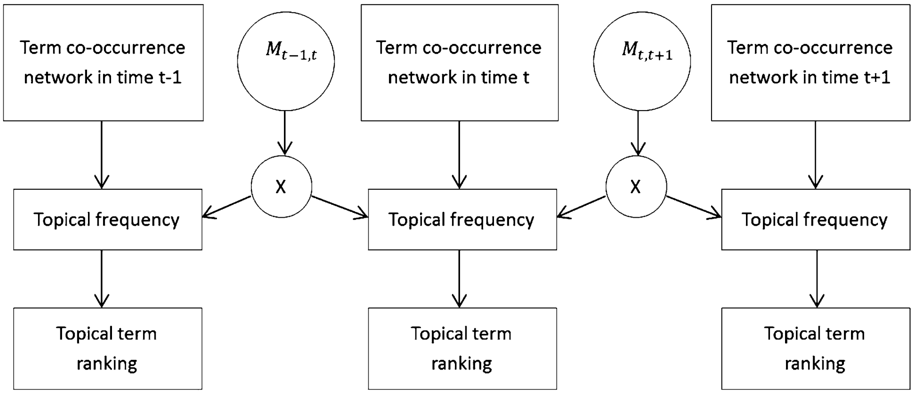


Fig. 3 The topic evolution model

depends on the matrix $M_{t,t-1}$. $M_{t,t-1} = \left(m_{ij}\right)_{t,t-1}$ is a matrix of $K_{t-1} \times K_t$, where K_{t-1} indicate the total topic number of year $t - 1$. m_{ij} is the correlation of topic i in $t - 1$ and topic j in t, if m_{ij} less than a threshold, then the corresponding variable x is 0.

We used cosine similarity to judge the correlations of topics. We counted the similarity of topics by computing all the probability of word distribution and the top ten feature probability word distribution. The computation method is the same as we introduced in finding new topics. We assigned the number of the topic to 24. Thus, we obtained two 24 × 24 matrixes M_1 and M_2 of all the words and feature words of two adjacent years. For example, for year 2005–2006, m_{1ij} inferred the similarity of all vector of topic i of words of 2005 and topic j of 2006, and m_{2ij} infers the similarity of feature vector of topic i of word of 2005 and topic j of 2006.

We now provide two definitions: highly possible evolutionary relationship and potential evolutionary relationship. Highly possible evolutionary relationship meant if these m_{1ij} and m_{2ij} were both greater than the thresholds, we marked the topic i of 2005 and topic j of 2006 had highly possible evolutionary relationship. The conception means that the two topics have an evolutionary relationship according to the text mining results. As for potential evolutionary relationship, we cannot judge whether there is an evolutionary relationship between two topics according to the text mining results, but we found a most potential relationship of a topic for users using the following method and let user judge it through visualization view.

Using these two similarities, we obtained the highly possible evolutionary relationship of 2 years. However, many topics had no high similarity with any topic of the next year. Therefore, these topics would seem like a "lonely island," and we could not see any relationship with other topics. To avoid this problem, we used another method to judge the potential evolutionary relationship between adjacent years.

Matrix M_1 we obtained from the last procedure is 24 dimension × 24 dimension. In a same example of 2005–2006, row i represents all the similarities of topic i of 2005 with all the topics of 2006 ($i = 1\ldots$column count $= 24$ in the first level). If all the numbers belonging to the row were smaller than the threshold, then using the aforementioned method, the topic i does not have a relationship with the next year's topics. On this occasion, we find the maximum number of the row: m_{1ij} ($j = 1\ldots24$), and treat topic j of 2006 and topic i of 2005 as having a potential evolutionary relationship. This method may not be an exact method to construct the evolutionary relationship, but is easy to understand and all procedures are produced automatically, which do not require users' instruction or operation. Moreover, as we will present the details in the next chapter, we provide an interactive visualization interface to let users judge and analysis the relationships between the topics.

3.3.3 Topic mapping

After CATHY ran, we obtained the topical frequency, which is a distribution over words. As mentioned earlier, the vocabulary length is 14,966. Thus, we treat each topic as a 14,966-dimension vector and each dimension is represented by the words' probability in a latent sematic space. Then, we obtain the relative position of each topic. We thus need to reduce dimension to present topics on a two-dimensional axis system.

For the first topic level, all 240 topics occur over 10 years. We use the vectors that infer the 240 topics to conduct topic mapping. We choose multidimensional scaling (MDS) to do dimensionality reduction. MDS is an algorithm that can keep the information of vectors while mapping a high-dimensional space into two dimensions. Thus, in a two-dimension axis system, the similar topics marked as nodes are placed close to each other. By this algorithm, we can easily obtain the relative space information of topics, and we can gain an intuitive understanding of the similar topics.

4 Visualization

We design views for representing the topic mining results and let users make their own analysis. Our design is to fulfill the following tasks: (1) understanding the concept a topic represents; (2) understanding the time evolution of the topics; (3) representing hierarchical structure of topics; (4) representing the similarities of topics; (5) identifying new, disappeared, and cross topics; (6) identifying the relationship inside a research domain and between some research domains. To achieve these goals, we design an interactive tool, which includes five views that include word cloud, sankey diagram, scatterplot, treemap, and stream diagram to present our visualization.

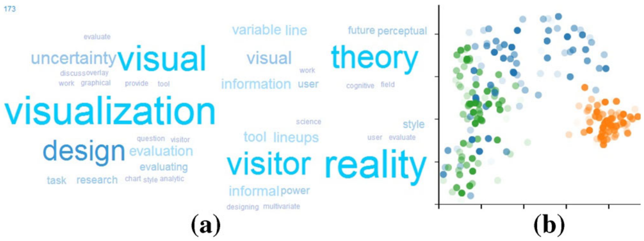


Fig. 4 **a** Word cloud of a topic from TVCG. **b** Scatter plot that shows space information of topics in 10 years. Opacity indicates the publishing year of the topic

4.1 Word cloud

After CATHY process, topics are defined as distributions over words that have different probabilities to occur within the specific topic. We wish to provide users more details of a specific topic and help them obtain a deep underlying concept of each topic. Therefore, we use a simple word cloud to display a topic. The word displayed in a word cloud is the most probable in the topic. The font size of each word represents the occurrence probability of each word in this topic. Figure 4a shows that via the simple word cloud, users can easily obtain the detailed concept of a topic represents. The bigger word cloud was the topic of the first level, and four smaller word clouds were the topics of the second level, which were subtopics of the topic of the first level.

4.2 Scatter plot

In Sect. 3, we mentioned that we obtained the relative space position relationship of each topic using MDS. In visualization, we used a scatterplot to represent the space information of each topic. Because MDS is quite stable, so the similar topics will stay close to each other. We map the topics' process results into a 2D-coordinate system as shown in Fig. 4b. The *x*-axis and *y*-axis have no meaning but to they present the relative position of each topic. We marked the topics of these three journals with different colors: blue for TVCG, orange for TKDE, and green for TOG. The opacity of each node infers the year of the topic. High opacity denotes the topics that belong to recent years, and low opacity denotes the topics that belong to past years.

It is easy to identify that TKDE is not close to other two journals. When a mouse is hovered over a node, the most probable word of the topic appears as a title. Users can obtain the primary information of this topic. When a node is clicked in this scatterplot, the respective topic word cloud and the subtopics word cloud show at the same time. Therefore, users can obtain more details of this topic. The scatterplot achieves goal (4), which represents similar topics and helps users judge the evolutionary relationship between topics.

4.3 Sankey diagram

We used Sankey diagram to represent the evolution of topics over 10 years. Sankey diagram is widely used in text visualization. Compared with stream diagram, sankey diagram represents topics using separated rectangles. We linked the similar topics of adjacent years to construct a changed flow of topics. The *x*-axis represents the year from 2005 to 2014. Thus, there are ten columns show the topics. We marked the topics from 1 to 192. Topic 1 to topic 24 denote the topics of 2005, topic 25 to topic 48 denote the topics of 2006, and so on. The topics of three journals in a specific year show on the same column of Fig. 1. The node marked with three different colors refer to three journals. The height of each node depends on the number of documents that belong to the topic in a specific year.

In Sect. 3, we used a method to find new topics. In visualization, we used rectangles with no stroke to represent the normal topics and rectangles with red-thick stroke to represent the new topics. This feature is easy for users understand and reminds users to focus to these topics. Moreover, we constructed the evolutionary relationship of topics. We use two concepts to present different probabilities of evolutionary relationships: highly possible evolutionary relationship and potential evolutionary relationship. We linked highly possible ones with orange and potential ones with green. If a topic is linked with another topic with green, then these two topics may not have a real evolutionary relationship. If this topic is not that similar as other topics, then maybe this topic is a disappeared topic that has not been discussed in next years. Figure 1

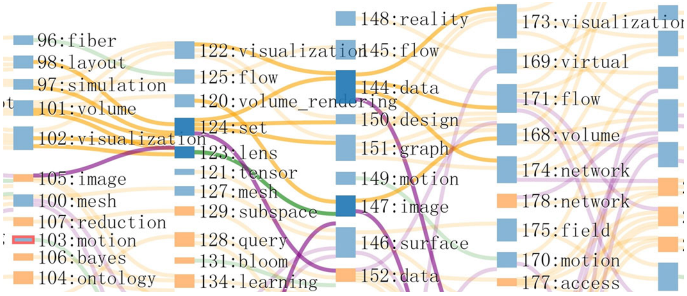


Fig. 5 Sankey diagram shows the topics relationship in adjacent years. The *curve links* shows that has a relationship between two topics. *Purple curves* denote relationships between cross topics, *orange curves* denote the highly possible evolutionary relationships, and *green curves* denote the potential evolutionary relationships between topics

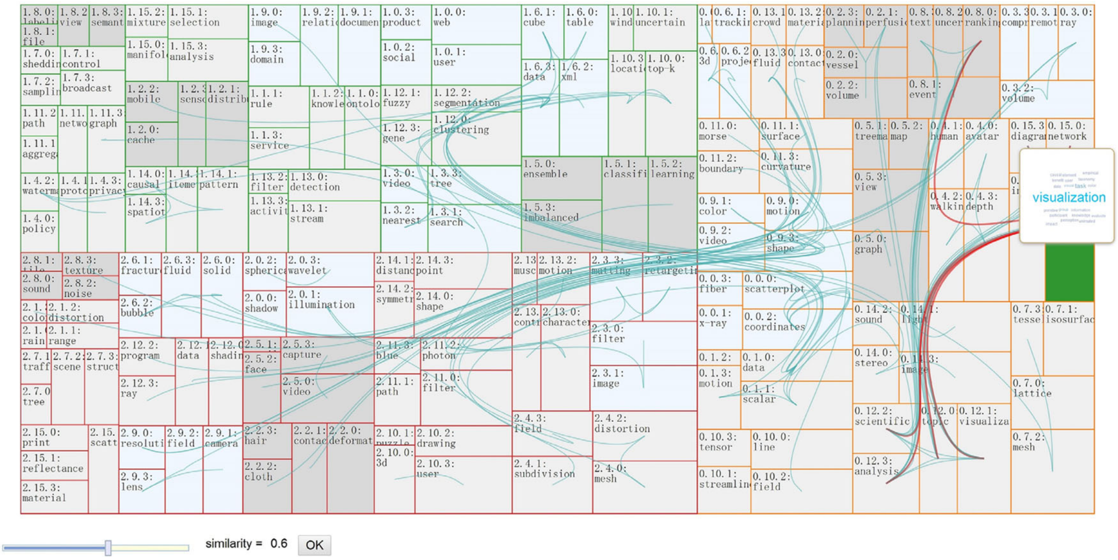


Fig. 6 Treemap of the second level the similarity threshold is 0.6 which means if the similarity of two topics are larger than 0.6, a *blue curve* links these topics to each other

indicates that most links are marked with orange, which means that potential relationships are not as strong as highly possible ones. Users can mainly analyze the potential links to judge whether a topic is disappeared.

Figure 5 shows that a topic may be similar to one or more, or no topics at all. Topics from a journal may be similar to those from other journals. In this occasion, we know that these three journals discussed common topics, which may infer the cross-point of these different research domains. In the sankey diagram, in addition to orange and green, other links are marked with purple. This color represents links of topics from one journal to another. Users can check these purple links to examine and analyze those cross topics.

4.4 Treemap diagram

In addition to analyzing the evolutionary relationship of topics, our work has another main goal: identifying the relationship inside a research domain and among some research domains. To achieve this goal, as we mentioned in Sect. 3, we use documents from the entire 10 years and divide them based on journals for data processing. Therefore, we obtained three tree-structure topic results as the hierarchical structures of topics and the similarities of topics inside one journal and among three journals. Treemap is a common visualization design to represent the hierarchical structure. Considering that we obtained a hierarchical structure in topic mining process, as well as our desire to link-related topics with curves, we choose treemap rather than node-link graph.

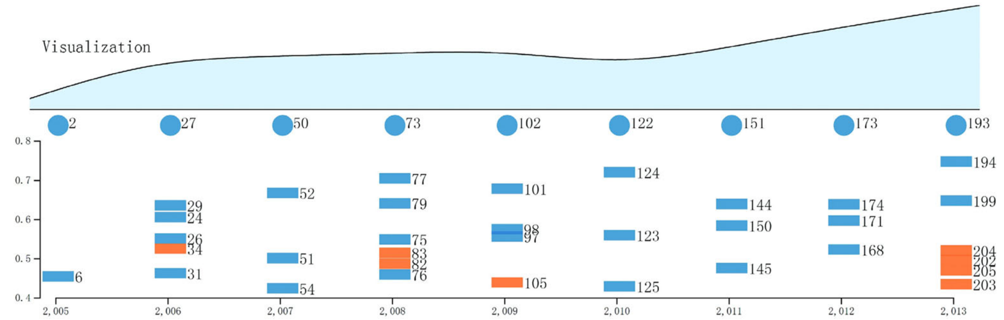


Fig. 7 Bottom of the figure as a stream diagram with a *scatterplot*. The stream diagram shows the trends of topic "visualization," and the marked scatterplot showing the similarities of topics related to topic "visualization"

On the first level, we annotate the topics with an ID from 0.0 to 0.15, 1.0 to 1.15, and 2.0 to 2.15. The first 0 denotes TVCG, 1 denotes TKDE and 2 denotes TOG. The number behind point from 0 to 15 denotes the topic ID. On the second level, we annotate the topics similarly: 0.0.0 to 0.0.3, 0.1.0 to 0.1.3, and so on. The last number is between 0 and 3, which denotes the subtopics' ID.

As Fig. 6 shows, the different stroke colors of three larger rectangles shows these three journals: red for TOG, green for TKDE, and orange for TVCG. In addition, the smaller rectangles with different colors show the different topics of the first level. Moreover, the smallest rectangles show the topics of the second level, which are the subtopics of the first-level topics. The text inside the rectangles is the most probable terms of the topic to help users get the primary information of a topic.

Moreover, we construct all the similarities between topics and want to show them in the treemap diagram. Thus, we use curves linking two topics to show that these two topics are similar to each other.

The curved lines over these rectangles represent the similarity among topics. If the similarity is larger than the threshold, we link two topics with a curve line. The threshold of similarities can be changed by users from 0.1 to 1 using the control bar under the treemap. The link shows the two topics have intrinsic connections. We can easily find the connections between two journals as well as inside journals. Figure 6 shows that when users hovering the mouse over a rectangle, the word cloud of corresponding topic shows the topic details. When users click a rectangle, the corresponding curve lines are highlighted with red for users to see more clearly and then explore the detail of topics. The Treemap diagram is independent of other view which shown in Fig. 1.

4.5 Stream diagram with similarities

Stream diagram is an effective visualization tool for representing the trends of a topic. For example, as Fig. 7 shows, we picked the topics that have the term "visualization" in each year, and estimate the word frequency, then mark in the stream diagram. The diagram shows that the visualization topic has become more and more important in the journal TVCG, because the frequency is much higher in recent years than that in 2005. At the bottom of the stream diagram, we use a scatterplot to show the similarity of other topics with topic visualization. We use circles to mark the topic visualization and rectangles to represent other topics. When a mouse is hovered over a circle or a rectangle node, the most probable word of the topic appears as a title. The y-axis is the similarity which ranges from 0.4 to 0.8 and x-axis represents the year. Similar to the sankey diagram, we used blue to represent TVCG, orange to represent TKDE, and green to represent TOG. From the figure, we can find out that there is no green rectangle which means the topics of TOG do not have high similarity with visualization. Most rectangles are blue, denoting that most topics are related to visualization in TVCG. The Fewer rectangles are orange shows TKDE includes some related topic with visualization. This view provides users with an instinctual effect of a user-appointed topic term, helping users find the relationship in cross-domain.

5 Case study

In this chapter, we used three user studies to show the effect of our topic mining results with visualization tools. The three studies are implemented in two main views: evolutionary-topic results view and all topics

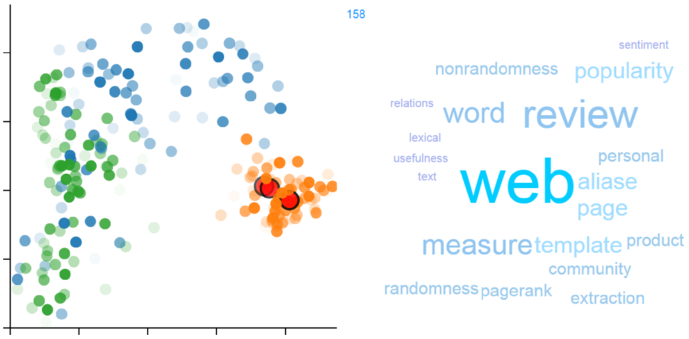


Fig. 8 In *the left* figure *red circles* with *black stroke* in the *scatterplot* show the topic 182, 158, and 202, respectively. These three *circles* are placed close to each other which mean they are similar with each other

results view. The first view helps users find the changes and trends over 10 years in these three journals and also helps users find the time point and analyze the reasons for cross-topic emergence. The second view is the overview of all topics in 10 years. Users can obtain more details of cross-domain topics. Therefore, we used three user studies, including the new topic analysis, the evolutionary cross-topic analysis, and the all cross-domain topics analysis.

5.1 New topic finding

As presented in Chapter 4, we used some highlighting design to let users check and analyze the topic mining results of TVCG, TKDE, and TOG journals. Users can easily find the highlighted information in the interface without extra work. In this occasion, we mark the new topic with a red stroke in sankey diagram so users can find it easily. As listed in Chapter 3, new topics are found in the ten-year period, which can be analyzed individually.

First we choose the topic 43: contour in 2006 of TOG. Upon clicking the topic, we find it concerns about contour, curve, shape, distance, and so on. This topic links with topic 67: point with a green curve, which indicates a potential evolutionary relationship between them. We then checked the position of these two topics in the scatterplot and found that the distance of these two nodes was not very close. However, we found another node representing topic 127 placed in the same position as node 43 in the scatter plot. We then checked the word cloud of topic 127 and noticed that the topics were about mesh, access, map, and hierarchy, which were dissimilar to topic 43. Therefore, we inferred that topic 43 appeared occasionally when the topic mining algorithm was run. Then, we chose topic 48 of TVCG in 2007. This topic shares the same term as topic 77 (color). Thus, a green link occurred between them. However, after analyzing the word cloud further, we found that topic 77 was about medical visualization, although we could not identify an appropriate domain for topic 48. The distance in the scatterplot of these two topics was not close, which also meant that these two topics had no real evolutionary relationship between them. Repeating the same procedure for topic 62, 103, 116, 155, and 197, we found that these new topics may not be the new ones, disappearing quite quickly in the next year.

We concluded that the reasons are the following: the algorithm was limited because the topic mining algorithm was running automatically. We could not change the parameter or add extra information to it. In addition, the size of data was not big enough in each year to obtain good topic mining results. We had only 100+ papers of a journal in 1 year, which do not support a good result of probability topic model.

As Fig. 8 shows, we found that topic 182 in year 2012 of TKDE was similar to topic 202 in 2013. These topics were about text, topic, and sequence. However, topic 182 was also very similar to topic 158 in 2011, which meant it may not be a real new topic. Topic 158 was about web, PageRank, among others, which had an intrinsic relationship with text. Thus, topic 182 was marked as a new topic maybe because in this year, text mining, and topic modeling were popular topics even though many related research works had already started before 2012.

Table 1 Cross topics in 10 years

TVCG and TKDE	
2005–2006	Topic 13: data, topic 29: volume
2006–2007	Topic 34: data, topic 52: volume
2008–2009	Topic 79: scatterplot, topic 105: image
2009–2010	Topic 105: image, topic 124: set
2010–2011	Topic 124: set, topic 152: data
2011–2012	Topic 144: data, topic 181: selection
2012–2013	Topic 181: selection, topic 194: topic
	Topic 179: uncertain, topic 194: topic
	Topic 174: network, topic 205: network
	Topic 178: network, topic 194: topic
	Topic 171: flow, topic 204: data
	Topic 171: flow, topic 202: series
2013–2014	Topic 194: topic, topic 228: privacy
	Topic 194: topic, topic 227: functional
TVCG and TOG	
2005–2006	Topic 7: surface, topic 45: subdivision
	Topic 21: ray, topic 29: volume
	Topic 23: parameterization, topic 28: surface
	Topic 22: display, topic 25: texture
2006–2007	Topic 28: surface, topic 65: deformation
	Topic 42: mesh, topic 55: surface
	Topic 45: subdivision, topic 55: surface
2007–2008	Topic 65: deformation, topic 72: surface
2008–2009	Topic 72: surface, topic 118: shape
	Topic 89: subdivision, topic 109: mesh
	Topic 78: rendering, topic 117: illumination
	Topic 74: parameterization, topic 118: shape
2009–2010	Topic 100: mesh, topic 136: color
	Topic 100: mesh, topic 137: sampling
	Topic 99: display, topic 143: image
	Topic 114: character, topic 126: motion
2010–2011	Topic 136: color, topic 146: surface
	Topic 137: sampling, topic 146: surface
	Topic 143: image, topic 147: image
2011–2012	Topic 147: image, topic 186: image
	Topic 146: surface, topic 184: surface
	Topic 165: motion, topic 170: motion
	Topic 162: shape, topic 172: surface
	Topic 163: field, topic 169: virtual
2012–2013	Topic 170: motion, topic 208: motion
	Topic 172: surface, topic 209: reconstruction
	Topic 190: motion, topic 195: video
	Topic 187: sampling, topic 198: rendering
2013–2014	Topic 198: rendering, topic 232: sampling
	Topic 212: sampling, topic 216: ray
	Topic 210: simulation, topic 221: simulation

5.2 Evolutionary cross-domain analysis

We marked curves with purple to denote that two topics belong to different journals. Figure 1 indicates that the amount of purple curves was not as many as color orange or green curves. It means that most evolutionary relationships between 2 years are from same journal. This is easy to understand because one journal had its own research domain and popular topics. Some links were from TVCG to TOG, or TOG to TVCG, which is normal because these two journals were both about computer graphics. We then found some lines from TKDE to TVCG, and most topics are related to "data." TKDE is a major journal of data mining, and TVCG includes papers belonging to visualization analysis, which examines data.

Those cross topics are listed in Table 1 as follows:

We first focus on cross topics of TVCG and TKDE. From 2005 to 2012, only one or no link was found between these 2 years. Such a result indicates that these two research domains are not sufficiently closely related in these years. We check these topics individually, and find that "data" in TKDE is related to

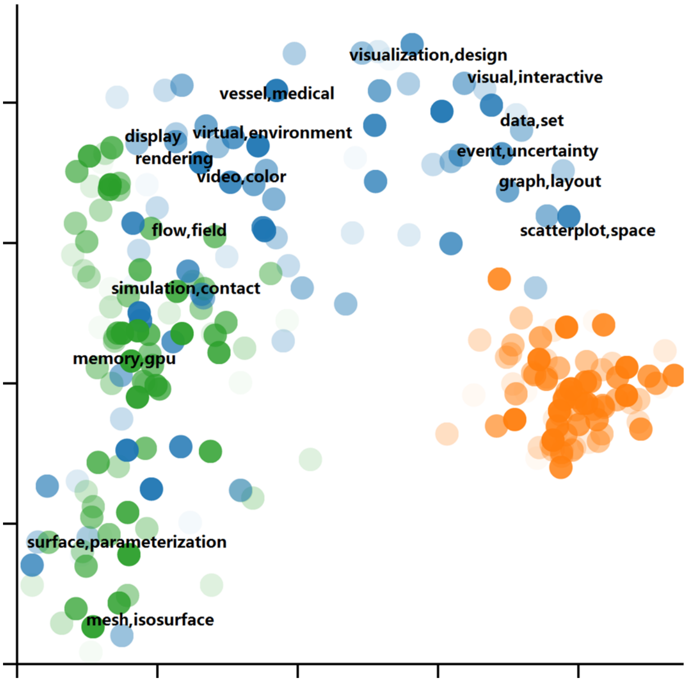


Fig. 9 The topics of TVCG in *scatterplot*

"volumetric data" in TVCG and the term "data set" in TVCG has almost the same meaning as the same term in TOG. However, a sudden increase of link amount is observed from 2012 to 2013. An obvious reason is that the papers of VAST were collected into TVCG in that year. VAST is the top conference of visualization analysis, and it is much closer to data mining than to graphics, science visualization, and information visualization. "Graph" and "network" are popular topics in these 2 years. "Flow" of TVCG in 2012, which refers to flow data, became a topic similar to the topics of TKDE. Such a result indicates that certain works using the data processing methods in data mining domain solve the problems of visualization domain.

Then we focus on the cross topics of TVCG and TOG, but we have not found the real cross-domain topics because these two journals share a same research area: computer graphics. We have checked all these topics and found that they all belong to computer graphics domain, as opposed to visualization domain.

5.3 Cross-domain topic analysis

We now use the treemap view to analyze the topic mining results of all 10 years. We first check the first level, which includes 16 topics from each journal, 48 topics in total. We change the similarity threshold and see the change of the curves. When similarity = 0.9, no curves occur in the treemap. When similarity = 0.8, curves are only present for one journal. When similarity = 0.7, cross topics begin to appear. We checked topic 0.1: data, set of TVCG and topic 1.6: data, xml of TKDE. Then, we checked the subtopic of topic 0.1 by the second level treemap of the same threshold. We found that the similarity only originated from one subtopic, which was topic 0.1.0: data, molecular, biology, and so on. We can be sure that the topic that discusses biology visualization belonged to science visualization. The result showed visualization analysis and science visualization had similar topics with data mining domain. Other cross topics originated from TVCG and TOG, and they were on computer graphics, which were not the real cross topics.

We further lowered the threshold to 0.6 and found that topic 0.0: scatterplot, space from TVCG was similar to topic 1.15: analysis, selection. Following the same steps as before, we found that the subtopics of these two topics were not similar to each other. After adjusting the threshold, topic 0.0.0: scatterplot was similar to topic 1.15.3: analysis. Topic 1.15.3 mentioned visualization, which is thus a common cross-topic. Topic 0.5: graph, layout of TVCG was quite similar to topic 1.11: graph, subgraph of TKDE. They both discussed the graph and network, which were in the same research area as data mining and visualization.

Although further lowered the threshold of the first level, no more real cross-topic emerged among these three research domains.

We manually annotated the labels in Fig. 9 according to the text mining results which show the Sankey diagram. From Fig. 9, we can find that those topics cross TVCG and TOG are in the left side of the scatterplot. Those topics are very similar with each other and all the topics belong to computer graphics. The topics: scatterplot, graph, event, data, and visual are in the right side of the figure. Some of them are close to the orange circles that represent TKDE, but we can see the topics of TVCG and TKDE is not that close to each other, although they have some cross topics. And some topics in the center part of the scatterplot such as vessel, medical which belong to the scivis topics are not very close to computer graphics.

We therefore found the relationship among these three journals. TVCG and TOG shared topics about computer graphics, but not about visualization. TVCG and TKDE shared topics primarily on data that belong to visualization analysis and data mining. Via these results, we could identify that visualization developed to be an independent subject from computer graphics in these 10 years. Visualization shares the same cross topics with data mining, but primarily in the area of visualization analysis.

6 Conclusion

In this paper, we used a hierarchical topic model to extract topics from three different research domains and to construct relationships among topics. Therefore, we designed a visualization interface and provided interaction operations for users to gain a deep understanding of the visualization development trend, as well as the correlation of three subjects. Via three user studies, we found new and cross topics for these three domains, especially for visualization. Our method also has some limitations, since user cannot adjust complex parameters of the topic model. Also for nonexpert user, parameters are not easy to appoint. We will deal with these issues in future work. Our work can be considered a brief introduction of visualization from past 10 years. We hope our work can help expand the understanding of visualization. In the future, this work can also be extensively used in other research domains or other data sources.

References

Blei DM, Lafferty JD (2007) A correlated topic model of science. Ann Appl Stat 1:17–35

Blei DM, Lafferty JD (2006) Dynamic topic models. In: Proceedings of the 23rd international conference on machine learning. ACM, pp 113–120

Blei DM, Ng AY, Jordan MI (2003) Latent dirichlet allocation. J Mach Learn Res 3:993–1022

Boyack KW (2004) Mapping knowledge domains: characterizing PNAS. Proc Natl Acad Sci 101(suppl 1):5192–5199

Chen C (2004) Searching for intellectual turning points: progressive knowledge domain visualization. Proc Natl Acad Sci 101(suppl 1):5303–5310

Chen C (2006) CiteSpace II: detecting and visualizing emerging trends and transient patterns in scientific literature. J Am Soc Inform Sci Technol 57(3):359–377

Cui W, Liu S, Tan L et al (2011) Textflow: towards better understanding of evolving topics in text. IEEE Trans Vis Comput Graph 17(12):2412–2421

Deerwester SC, Dumais ST, Landauer TK et al (1990) Indexing by latent semantic analysis. JAsIs 41(6):391–407

Ding W, Chen C (2014) Dynamic topic detection and tracking: a comparison of HDP, C-word, and cocitation methods. J Assoc Inf Sci Technol 65(10):2084–2097

Dou W, Wang X, Chang R et al (2011) Paralleltopics: a probabilistic approach to exploring document collections/visual analytics science and technology (VAST). In: IEEE conference on 2011, pp 231–240

Dou W, Li Y, Wang X et al (2013) HierarchicalTopics: visually exploring large text collections using topic hierarchies. IEEE Trans Vis Comput Graph 19(12):2002–2011

Gad S, Javed W, Ghani S et al (2015) ThemeDelta: dynamic segmentations over temporal topic models. IEEE Trans Visual Comput Graphics 21(5):672–685

Ginsparg P, Houle P, Joachims T et al (2004) Mapping subsets of scholarly information. Proc Natl Acad Sci 101(suppl 1):5236–5240

Griffiths TL, Steyvers M (2004) Finding scientific topics. Proc Natl Acad Sci 101(suppl 1):5228–5235

Havre S, Hetzler E, Whitney P et al (2002) Themeriver: visualizing thematic changes in large document collections. IEEE Trans Vis Comput Graph 8(1):9–20

Heimerl F, Han Q, Koch S, Ertl T (2015) CiteRivers: visual analytics of citation patterns. IEEE Trans Visual Comput Graphics 22(1):190–199

Isenberg P, Isenberg T, Sedlmair M et al (2014) Toward a deeper understanding of visualization through keyword analysis. arXiv preprint arXiv:1408.3297

Landauer TK, Laham D, Derr M (2004) From paragraph to graph: latent semantic analysis for information visualization. Proc Natl Acad Sci 101(suppl 1):5214–5219

Liu S, Wang X, Chen J et al (2014) TopicPanorama: a full picture of relevant topics/visual analytics science and technology (VAST). In: IEEE conference on 2014, pp 183–192

Liu S, Wu Y, Wei E et al (2013) StoryFlow: tracking the evolution of stories. IEEE Trans Vis Comput Graph 19(12):2436–2445

Mane KK, Börner K (2004) Mapping topics and topic bursts in PNAS. Proc Natl Acad Sci 101(suppl 1):5287–5290

Mimno D, Li W, McCallum A (2007) Mixtures of hierarchical topics with pachinko allocation. In: Proceedings of the 24th international conference on machine learning. ACM, pp 633–640

Morris SA, Yen GG (2004) Crossmaps: visualization of overlapping relationships in collections of journal papers. Proc Natl Acad Sci 101(suppl 1):5291–5296

Newman MEJ (2004) Coauthorship networks and patterns of scientific collaboration. Proc Natl Acad Sci 101(suppl 1):5200–5205

Oelke D, Strobelt H, Rohrdantz C et al (2014) Comparative exploration of document collections: a visual analytics approach. Comput Graph Forum 33:201–210

Ramage D, Hall D, Nallapati R et al (2009) Labeled LDA: a supervised topic model for credit attribution in multi-labeled corpora. In: Proceedings of the 2009 conference on empirical methods in natural language processing, vol. 1. Association for Computational Linguistics, pp 248–256

Romo-Fernández LM, Guerrero-Bote VP, Moya-Anegón F (2013) Co-word based thematic analysis of renewable energy (1990–2010). Scientometrics 97(3):743–765

Wang C, Danilevsky M, Desai N et al (2013) A phrase mining framework for recursive construction of a topical hierarchy. In: Proceedings of the 19th ACM SIGKDD international conference on knowledge discovery and data mining. ACM, pp 437–445

Wei F, Liu S, Song Y et al (2010) Tiara: a visual exploratory text analytic system. In: Proceedings of the 16th ACM SIGKDD international conference on knowledge discovery and data mining. ACM, pp 153–162

White HD, Lin X, Buzydlowski JW et al (2004) User-controlled mapping of significant literatures. Proc Natl Acad Sci 101(suppl 1):5297–5302

Wu Y, Liu S, Yan K et al (2014) OpinionFlow: visual analysis of opinion diffusion on social media. IEEE Trans Vis Comput Graph 20(12):1763–1772