Visual Informatics ▮ (▮▮▮▮) ▮▮▮–▮▮▮

Contents lists available at ScienceDirect

Visual Informatics

journal homepage: www.elsevier.com/locate/visinf

VISTopic: A visual analytics system for making sense of large document collections using hierarchical topic modeling

Yi Yang [a,b,*], Quanming Yao [a], Huamin Qu [a]

[a] Hong Kong University of Science and Technology, Hong Kong
[b] Lenovo Group Limited, Hong Kong

ARTICLE INFO

Article history:
Available online xxxx

Keywords:
Topic-modeling
Text visualization
Visual analytics

ABSTRACT

Effective analysis of large text collections remains a challenging problem given the growing volume of available text data. Recently, text mining techniques have been rapidly developed for automatically extracting key information from massive text data. Topic modeling, as one of the novel techniques that extracts a thematic structure from documents, is widely used to generate text summarization and foster an overall understanding of the corpus content. Although powerful, this technique may not be directly applicable for general analytics scenarios since the topics and topic–document relationship are often presented probabilistically in models. Moreover, information that plays an important role in knowledge discovery, for example, times and authors, is hardly reflected in topic modeling for comprehensive analysis. In this paper, we address this issue by presenting a visual analytics system, VISTopic, to help users make sense of large document collections based on topic modeling. VISTopic first extracts a set of hierarchical topics using a novel hierarchical latent tree model (HLTM) (Liu et al., 2014). In specific, a topic view accounting for the model features is designed for overall understanding and interactive exploration of the topic organization. To leverage multi-perspective information for visual analytics, VISTopic further provides an evolution view to reveal the trend of topics and a document view to show details of topical documents. Three case studies based on the dataset of IEEE VIS conference demonstrate the effectiveness of our system in gaining insights from large document collections.

1. Introduction

Nowadays, the development of digital libraries, such as JSTOR, IEEE Xplorer and ACM, has made repositories of knowledge more and more available to the public. Searching documents based on keywords, as the typical approach that users interact with these libraries, is effective to obtain targeted knowledge. However, due to the vast amount of information, it is challenging for users to learn and gain new knowledge from the corpus in an exploratory way.

In recent years, much attention has been paid on text mining techniques that automatically analyze large volume of textual data with computational power. Topic modeling is one of such techniques. Based on unsupervised machine learning models, topic modeling can discover a thematic structure describing what topics are covered and how topics and documents are related within the corpus, from a statistic point of view (Landauer et al., 1998; Hofmann, 1999; Blei et al., 2003). Many different corpora, including humanities (Blei, 2012), social media (Xu et al., 2013) and online reviews (Titov and McDonald, 2008), have been analyzed when the thematic summarization of those specific domains are investigated in research. Although powerful, topic modeling has some limitations when directly applied to real-world analytical tasks. First, as presented in the probabilistic way, topic models are less human-readable. It makes knowledge discovery less effective as users have to interpret every topic and topic–document relationship based on numerical values throughout entire mode output. Second, it often requires a linkage of topics and other information such as times, people and locations to discover insight from document collections. However, these perspectives are available in few text mining tools, resulting in a gap between techniques of text mining and requirements in piratical usage.

In this paper, we present a visual analytics system, VISTopic, to leverage interactive visualization techniques in making sense of large document collections with topic modeling. In specific, we use hierarchical latent tree model (HLTM) (Liu et al., 2014), which is a novel topic model that extracts a set of hierarchical topics, to

* Corresponding author.
E-mail address: yyangao@connect.ust.hk (Y. Yang).
Peer review under responsibility of Zhejiang University and Zhejiang University Press.

http://dx.doi.org/10.1016/j.visinf.2017.01.005

Please cite this article in press as: Yang, Y., et al., VISTopic: A visual analytics system for making sense of large document collections using hierarchical topic modeling. Visual Informatics (2017), http://dx.doi.org/10.1016/j.visinf.2017.01.005.

summarize the corpus at different levels of abstraction. To address the above limitations, VISTopic first provides a topic visualization design that focuses on the structural and semantic features of HLTM to foster better understanding of modeling results. Then, times, authors and other detailed meta data are extracted from the documents and integrated into VISTopic for the purpose of visual analytics. By this means, VISTopic can facilitate users in analyzing document collections from two levels of details: overall exploration and investigation of the topic system, and close examination on topics of interest from comprehensive perspectives, including semantic patterns within the topic hierarchy, trend at temporal domain and aggregated patterns in regards of authorship or venue distribution. The contributions of this work can be summarized as follows:

- The design and implementation of a comprehensive visual analytics system with three linked views to investigate the document collections based on multi-perspective information;
- Novel visualization designs, which are tailored to the characteristics of a new hierarchical topic model, facilitate overall understanding and interactive exploration of the topic organization;
- Comprehensive case studies based on the dataset of IEEE VIS conference to show how our system can lead to interesting insights and handle various analytical tasks.

The rest of this paper is organized as follows. Section 2 reviews the related areas, including topic modeling and visualization techniques. Section 3 presents design rationals for VISTopic. The system design, including data processor and visual interface, is introduced in Section 4. Case studies in which VISTopic was applied to explore and analyze IEEE VIS corpus are shown in Section 5. Finally, Section 6 discusses and concludes the paper.

2. Related work

Two lines of works, namely topic modeling and topic visualization techniques, were the main inspiration for the design of VISTopic.

2.1. Topic modeling

In topic modeling research, a "topic" is defined as set of words highly correlated with each other. Topic modeling refers to the approach that applies probabilistic models to extract hidden topics from large document collections. These models often define a generative process for document generation, describing how words in documents get into their place. Finally, each document will be represented as a mixture of topics. Latent Dirichlet Allocation (LDA) (Blei et al., 2003) is the most commonly used topic model on, e.g., analyzing social media content (Hong and Davison, 2010) and historical documents (Yang et al., 2011). Unfortunately, LDA and other traditional models generate a large number of topics, making it difficult for users to have understandings of underneath documents at different level. Thus, they fail to address the issue of scalable analysis.

To fix this issue, hierarchical topic models are proposed to organize topics into tree structures (Blei et al., 2010; Ghahramani et al., 2010; Liu et al., 2014). Among them, HLTM (Liu et al., 2014) is emerging as a novel approach to detect hierarchical topics. It builds the hierarchy in a bottom-to-top manner. The most specific topics are taken as the bottom nodes, they are generated from separating vocabulary into independent sets of related words. Then, nodes are merged together as larger parent nodes based on their semantics similarities. Thus, HLTM shows an advantage of presenting a clear semantic structure on the bases of all vocabulary words. Note that we focus on describing the model features of HLTM instead of the technical details of how topic hierarchies are built. The interested reader is referred to Mourad et al. (2013) and Liu et al. (2014) for a more thorough technical description.

2.2. Topic visualization

Visual presentations of topic modeling results have been significantly beneficial to the use of topic models. Based on the purpose that topic-based information is visualized, we can generally classify these visualizations into two categories: model assessment tools and text analytical tools.

Visualizations in model assessment tools facilitate model designers and users to evaluate the model quality. Usually model-driven visualizations are designed to capture the features of target topic models. For instance, parallel coordinate metaphor has been used in Paralleltopics (Dou et al., 2011), LDAexplore (Ganesan et al., 2015) and other LDA-based topic models to reveal the patterns in probabilistic discriminations. With the development of hierarchical topic models, tree visualizations have became the prevalent approach to present topic hierarchies (Dou et al., 2013; Smith et al., 2014). Unlike these applications, VISTopic further takes the model features of HLTM into consideration and brings about a model-specific semantic design based on the established approaches. Lastly, there have been some user interfaces, such as UTOPIAN (Choo et al., 2013), that focus on interactive model verification and refinement. Considering the model complexity of HLTM, VISTopic extracts topic hierarchy off-line and excludes this model refinement function from the interface.

In text analytical tools, high-level information such as text summarization is usually extracted from textual data for visualization. As one of the pioneer work, the Maps of Computer Science (MoCS) project (Fried and Kobourov, 2014) designed a geographic map metaphor to visually summarize underlying topics in the DBLP database. Besides, temporal patterns of themes are crucial to many analytical tasks. In this regard, ThemeRiver (Havre et al., 2002) presented a general solution to depict the evolving patterns of multiple themes. Thereafter, the visually appealing river metaphor has been widely accepted and extended for trend analysis (Cui et al., 2011; Wei et al., 2010; Sun et al., 2014). In real-world applications, faceted information need to be explored and analyzed together. However, comparatively little effort has been made to bring together semantic, temporal and bibliographic information into a single analytical tool. Thus we develop VISTopic, based on three separated but coordinated views, to facilitate comprehensive analytical tasks.

3. Design rationals

A user-centered design process is followed to develop and improve our visual analytics system. We worked closely with two domain experts, one is a professor majored in information visualization and the other is a researcher in the field of machine learning, to discuss design requirements and present the prototype of VISTopic iteratively. The following design guidelines are derived based on their feedback.

R1. An overview of topic organizations. An effective visualization is required by experts to allow them to easily form a full picture of the topic organization, including its total number of branches, topics with large semantic coverage and so on. Moreover, a visual representation to depict the semantic characteristics of topics is required, not only to facilitate the discovery of salient topics, but also to enable a deeper understanding of the model mechanisms.

Please cite this article in press as: Yang, Y., et al., VISTopic: A visual analytics system for making sense of large document collections using hierarchical topic modeling. Visual Informatics (2017), http://dx.doi.org/10.1016/j.visinf.2017.01.005.

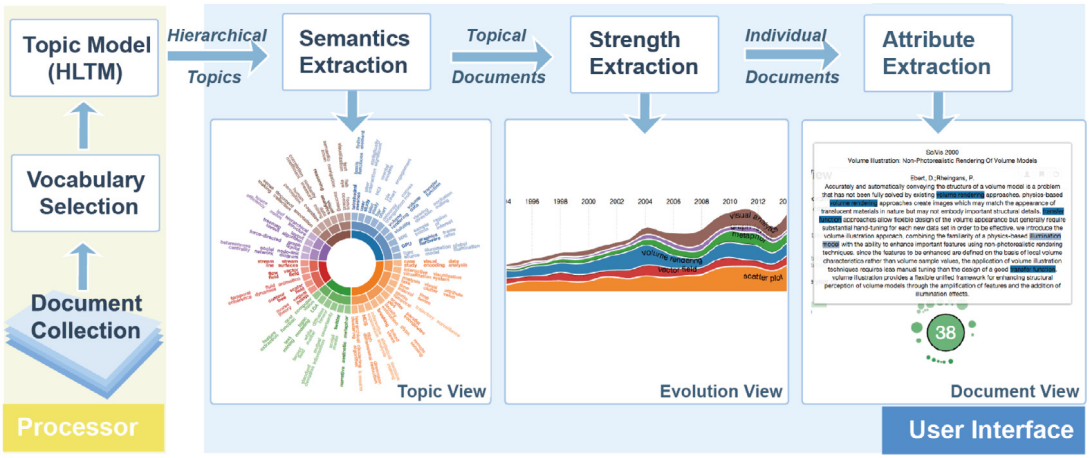


Fig. 1. System architecture of VISTopic.

R2. Multi-perspective analysis of topics. Our collaborators required access to multiple aspects of a topic for in-depth analysis. Hence, the topical information is desired to be highlighted within the global view of topic hierarchy, followed the spirit of "focus+context". A timeline is needed to provide for the strength changes of topical documents over time. Experts further required visual representations showing document distribution over venues, times, article impacts, authors and other perspectives to obtain a comprehensive understanding of the topic.

R3. Detail exploration via user interactions. Appropriate exploration of details needs to be done for practical applications. Particularly, the exact figures behind the interactive visualizations, such as the population statistics of a topic and the exact weight of a topical document, should be provided when users request them. Meanwhile, the full content of articles should be available for users who are interested to read. We also provide representations that can be effortlessly flipped through, such as the two perspectives to show the temporal strength of topics, thus allowing users to achieve a variety of analytical tasks.

4. System design

VISTopic is a visual analytics system that consists of two major components: a data processor and an interactive visual interface. An overview of the system architecture is shown in Fig. 1.

4.1. The data processor

In the data processor, the documents are collected and converted to plain texts. The cleaned data then underwent two processes, namely vocabulary selection and topic modeling. First, we determine a vocabulary to form the word space in topic modeling. The words can be either automatically generated based on natural language processing approach or decided with domain knowledge. Then, documents are transformed to word-based representation, which is a word-by-document matrix, as the training data for topic modeling. All topic modeling results and meta data of individual documents are stored in our database and available for use. An illustrative example of data processing can be found in case studies in Section 5.

4.2. The user interface

As illustrated in Fig. 2, VISTopic comprises three primary views: a Topic view, an Evolution view and a Document view. Each view provides patterns at a specific perspective and, more importantly, these views are interlinked to allow users to learn and gain knowledge through a joint analysis of topics. In addition, VISTopic includes an article viewer to allow users to read the original document.

4.2.1. Topic view

The topic view (Fig. 2(a)) presents a data-driven view of hierarchical topic organization based on a combination of two intuitive visualizations, namely Sunburst diagram and tag cloud. The visualization aims at providing an overview of topic organizations (R1) and supporting in-depth analysis of topic semantics (R2). In this section, we present the specific designs guided by these rationals.

Sunburst diagram for structure visualization. As surveyed by Tree.net (Schulz, 2011), a considerable number of tree visualizations have been developed to visualize hierarchical data. We choose the Sunburst diagram (Stasko and Zhang, 2000) for topic hierarchy visualization by taking two aspects, i.e., scalability and node shape, into consideration.

First, the Sunburst diagram features a radial layout which scales well for large trees. As is shown in Fig. 3(a), nodes on each level of the tree data are arranged as a concentric ring around the root. With level increasing, larger area is allocated to the levels that have larger number of nodes, thus, the space is used efficiently.

Then, node has the shape of slice arc that sweeps out an area. With this space-filling fashion, the connection between parent and children is implicitly shown by arc adjacency. Moreover, it is intuitive to identify the number of descendants beneath a node with the angular width of its arc. By this means, users can quickly discern topics with large semantic coverage to investigate.

Based on the Sunburst diagram, we further adopt the contextual anchor proposed in Smith et al. (2014). As is shown in Fig. 3(b), the contextual anchor shows where a node is located by highlighting its ancestors in the hierarchy. Such visual encoding is intuitive to remind users of the context around a topic.

Modified tag clouds for semantic visualization. The basis of topics is sets of words that weighted by probability. Tag cloud, as the text visualization that encodes word weights with font color or size, is widely used in topic semantic visualization. For instance, a segment of topic hierarchy generated by HLTM is presented by hierarchically arranged tag clouds in Fig. 4.

However, with the conventional use, it is not effective to discern how a parent topic is semantically close to its children in tag clouds. Therefore, we design a modified tag cloud to emphasize on the overlapped words in hierarchical topics. More specifically, we first

Please cite this article in press as: Yang, Y., et al., VISTopic: A visual analytics system for making sense of large document collections using hierarchical topic modeling. Visual Informatics (2017), http://dx.doi.org/10.1016/j.visinf.2017.01.005.

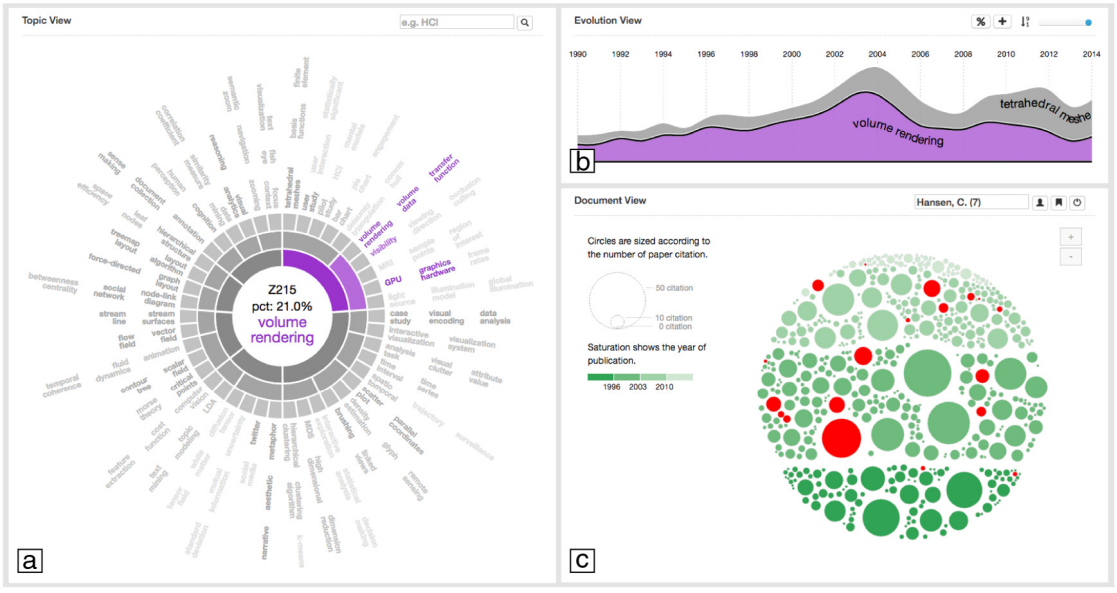
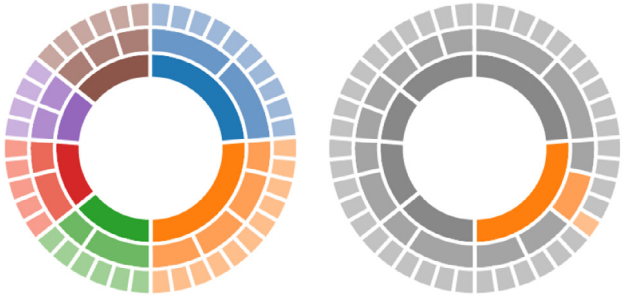


Fig. 2. Overview of VISTopic interface. The interface has three main views: Topic view (a), Evolution view (b) and Document view (c). A user is exploring Topic Z215 (light purple) in detail. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

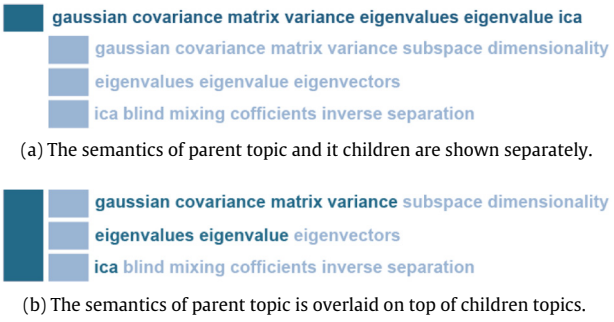


(a) Overview of the tree data. (b) Highlight of the location.

Fig. 3. The Sunburst diagram with radial layout and space-filling fashion for hierarchy visualization.

- gaussian covariance matrix variance eigenvalues eigenvalue ica
  - gaussian covariance matrix variance subspace dimensionality
  - eigenvalues eigenvalue eigenvectors
  - ica blind mixing coefficients inverse separation

Fig. 4. Conventional tag cloud for a segment of topic hierarchy generated by HLTM.



(a) The semantics of parent topic and it children are shown separately.

(b) The semantics of parent topic is overlaid on top of children topics.

Fig. 5. A modified tag cloud that shows semantic closeness.

turn the tag clouds shown in Fig. 4 to those in Fig. 5(a), by mapping the level that a word occurs to its color saturation. Then we break down the tag cloud of parent topic and overlay all the words to where it occurs at the lowest level (Fig. 5(b)). For larger topic hierarchies, the shown color is consistent to the highest topic level that uses a word, as its occurrence at specific topics can be easily inferred.

With the modified tag clouds, there is no more necessary to explicitly present tag clouds for high level topics, and the space to display a topic organization can be reduced. Moreover, the way to integrate above visualizations becomes clear: radiating the tag clouds from the border of the Sunburst diagram and aligning them with the leaf nodes (Fig. 6(a)). Alternatively, users can understand the combination as rolling up multiple sets of tag clouds in Fig. 5 around a center.

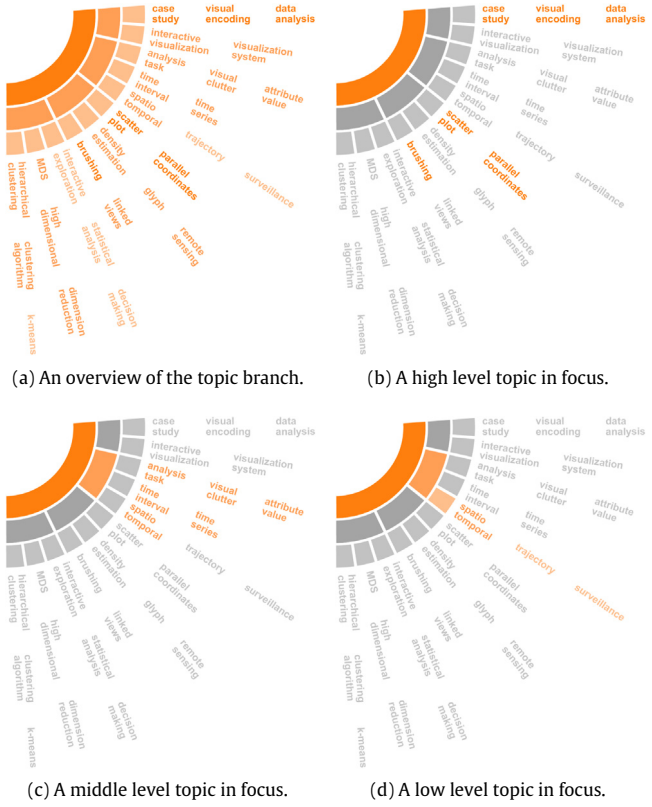
Overview and detailed analysis of topic hierarchy. With the Sunburst visualization, users can quickly identify the scale of entire topic organization. We further choose color encoding schemes to enrich the information displayed in the overview. First, as mentioned in the design of modified tag cloud, we use color saturation to distinguish topics at different levels of abstraction. Besides, we map the subtrees that directly separate the topic space with color hues, to show the major topic branches in the specific domain (R1).

When a topic of interest is identified, its semantics can be emphasized within the "context" of overall semantics with user interaction. At that time, all topics except the focused one and its ancestors are deemphasized in gray color as a reminder. Fig. 6 illustrates the effect with focus shifting along a path in a topic branch (R2). In addition, as is shown in Fig. 2(a), information including topic name, abbreviated label and its strength is displayed in the diagram center to facilitate understanding of the topic (R3).

4.2.2. Evolution view

In the evolution view (Fig. 2(b)), a ThemeRiver-based timeline is presented to show the temporal strength of topics. In this section, we introduce several designs to help users understand topic trend and gain more insights by comparing multiple topics (R2).

Topic trend visualization. Since we are analyzing documents accumulated over time, the temporal strength can be derived by summing up the topical documents within distinct time frames. Then,



(a) An overview of the topic branch.

(b) A high level topic in focus.

(c) A middle level topic in focus.

(d) A low level topic in focus.

Fig. 6. Interactive visualization for overview and detailed analysis of topic hierarchy.

multiple topics are presented by the ThemeRiver visualization, as introduced in Section 2.2, to reveal the overall temporal patterns. A set of stacked rivers forms the basis of this visualization, where each river flowing from left to right represents a specific topic and its width varying shows how the temporal strength evolves.

Two perspectives are available in the evolution view to measure the temporal strength: the absolute count and the proportion. The considerations behind this design are two-folds. The first is to flexibly meet the requirements in different scenarios. The second is to provide a joint view to understand the strength of changes. In fact, any one perspective alone potentially leads to a biased understanding. For example, a temporal strength can both increase in absolute count and decrease in proportion simultaneously. This may happen when the corpus grows at an even faster rate than topical documents over time. Therefore, this design allows users to examine the trend with either representation and switch to the other to verify the understanding.

Topic trend comparison. We allow users to compare topics that branch out from the same parent topic. Although this task is common in real-word applications, ThemeRiver fails to provide a visual cue to make the comparison intuitive. In our evolution view, we permit sorting the multiple topics at every few time stamps to detect the events when topics have their weight re-ordered. More specifically, users are allowed to set the frequency to detect the re-ordering patterns. During every detection, topics are ranked based on current strength, and rivers for the stronger topics will flow downwards while those for the weaker will flow upwards in the stacked graph (Fig. 7(a)). By this means, users can identify not only the dominant topics in different period, but also the overall evolution patterns for the specific topic branch.

Furthermore, we provide a tooltip to show the exact figure for temporal strength on demand (R3). The tooltip keeps hidden until users interacting with a river in the graph. Then the tooltip shows up at the corresponding time stamp, and the statistics can be found in either red or green, indicating whether the temporal strength is stronger or weaker than its average strength across time (Fig. 7(b)).

4.2.3. Document view

The document view (Fig. 2(c)) provides an interactive visualization to accommodate the high-dimensional meta data in topical documents. Users can either analyze them with overall patterns at various perspectives (R2), or get access to a document in detail (R3). In this section, we introduce the visual designs and user interactions that facilitate the above use scenarios.

Bubble-based document representation. We represent individual documents as bubbles for two reasons. First, it is intuitive to identify the multi-dimensional information encoded by the size, position and color of a bubble. More importantly, the bubbles can be packed together to provide a space-efficient solution, as shown in Fig. 8. The scalability issue is urgent as a large number of documents can be retrieved when the topic is board.

In a pack layout, the position of bubbles usually address nothing but a stable status determined by graph drawing algorithms. In spite of that, we are motivated to make the layout more meaningful by enforcing shorter distance for documents of same attribute. To this aim, we refer to the force-directed algorithm (Kamada and Kawai, 1989) for bubble layout. In this algorithm, the overall bubbles are modeled as a physical system with two distinct forces: an attractive force to reduce empty space between pairs of bubbles and a repulsive force to ensure that they do not overlap. We further add a third force into the system to preserve groups of bubbles residing in vertical layers. As a result, one dimension of document attribute, for example, publishing time, can be used to organize the bubbles in visualization to facilitate document navigation and knowledge discovery.

The bubble size encodes the importance of a document. In our case studies, VISTopic is applied to exploring and analyzing academic publications. Therefore, we determine the importance as the total number of following publications that have cited it within this academic archive. By default, we utilize the color channel partially, simply providing a double encoding of the time information with color saturation (Fig. 8(a)).

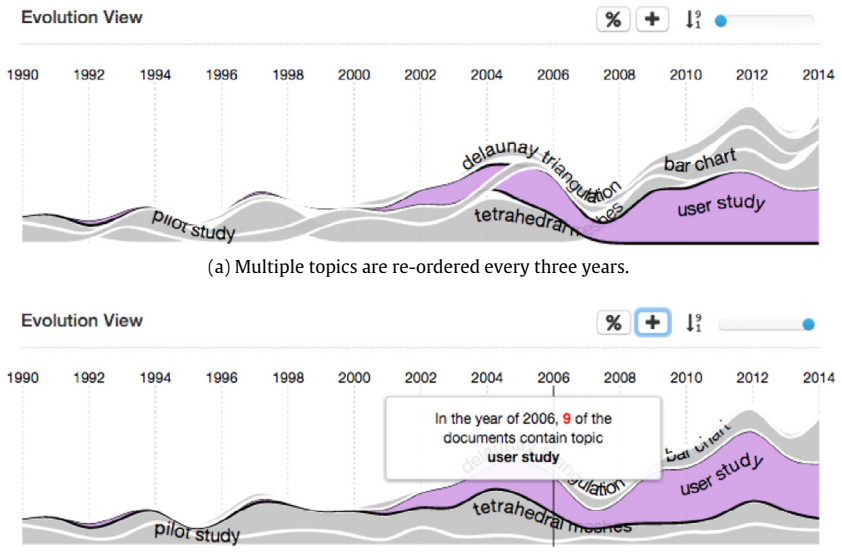
Interactive exploration and analysis. With the visual encodings described above, users can have a basic understanding of the topical documents, such as the temporal distribution, the amount of influential documents and so on. Moreover, they can interact with the visualizations to search salient patterns, switch visual encodings to document clustering at a specific dimension and find meta data in detail.

First, we allow users to retrieve topical documents that are authored or co-authored by identical people. For example, the nineteen highlighted documents marked in red are published by same author Hansen, C. in Fig. 2(c). For users who may not have sufficient knowledge about the topic and domain, we also allow them to navigate a list of authors for exploratory search. In particular, the list is ranked by an h-index value, which suggests how much impact or contribution author made to the community corresponding to the collection.

Then, the default color encoding can be replaced with categorical colors to show document clustering results at a specified perspective (Fig. 8(b)). For example, the perspective is determined as the three child conferences of IEEE VIS in our case studies. As a result, users can readily combine multiple visual cues to discern patterns and answer comprehensive questions, such as which venue is more dedicated to this topic.

Furthermore, the attribute used to encode bubble size, for example, the number of citations in our case study, is directly shown

Please cite this article in press as: Yang, Y., et al., VISTopic: A visual analytics system for making sense of large document collections using hierarchical topic modeling. Visual Informatics (2017), http://dx.doi.org/10.1016/j.visinf.2017.01.005.



(a) Multiple topics are re-ordered every three years.

(b) Exact figure is shown in a tooltip.

Fig. 7. Specific designs in the evolution view.

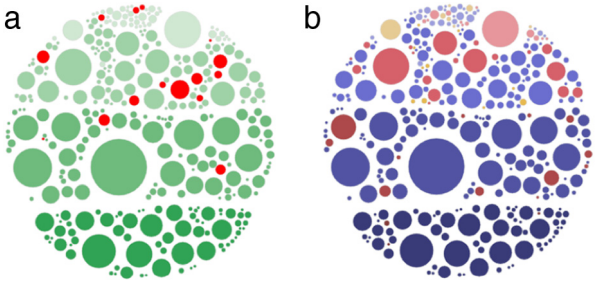


Fig. 8. Color schemes representing information at different aspects of data. (a) double encoding of time information with salient documents highlighted. (b) categorical colors for documents clustering.

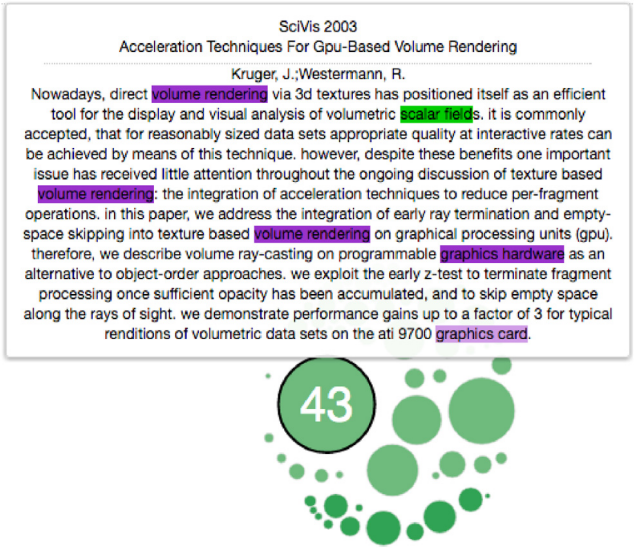


Fig. 9. Tooltip for detailed analysis of meta data and topic assignment in the document view.

when users hovering over a bubble (Fig. 9). A tooltip describing the document with detailed meta data will also show up. The tooltip can be further extended for a full text of the paper abstract upon user request. We further use small markers with the color hue and saturation corresponding to the branch and level of a topic to annotate keywords in the text. By this means, users can easily discover whether a document concentrates on a single topic or mixes several of them and how the abstract is composed based on the topical words.

5. Case studies

The IEEE VIS conference dataset is used as an exemplary corpus for our VISTopic system. In this section, we first introduce the data processing approach with focus on the domain-specific vocabulary selection. Then, we present two use cases that apply VISTopic to analyzing IEEE VIS corpus and finding interesting patterns.

5.1. Data processing and topic modeling

To start, we collected all the VIS publications ranging from 1990 to 2014. The entire collection contains 2592 documents which are both indexed by Vispubdata (Isenberg et al., 2015) and also available to download. We also clean and store the meta data associated with each document.

To have an appropriate vocabulary, we iteratively refine the selection approach based on topic modeling results. In the pilot experiment, all the words were taken as candidates, and we employed the score of TF–IDF (Jones, 1972), which is a general natural language processing metric to discover important words, to rank them. However, the top words are dominated by general words such as *visualization*, *node* and *link*, while some common terminologies, such as *parallel* (*parallel coordinates*) and *transfer* (*transfer function*) are excluded. Although the topic modeling works fine, our collaborators required an more domain-driven vocabulary to show the domain-specific patterns in topic modeling results.

Therefore, we make an effort to elaborate candidates manually rather than including all words in collection. Two authoritative resources are referred to guide our candidate selection. The first one is keywords that authors assign to their conference paper. We collected a list of 2629 author keywords including words and phrases from keyvis.org (Isenberg et al., 2014). The other one is a textbook introducing the problems and methodologies

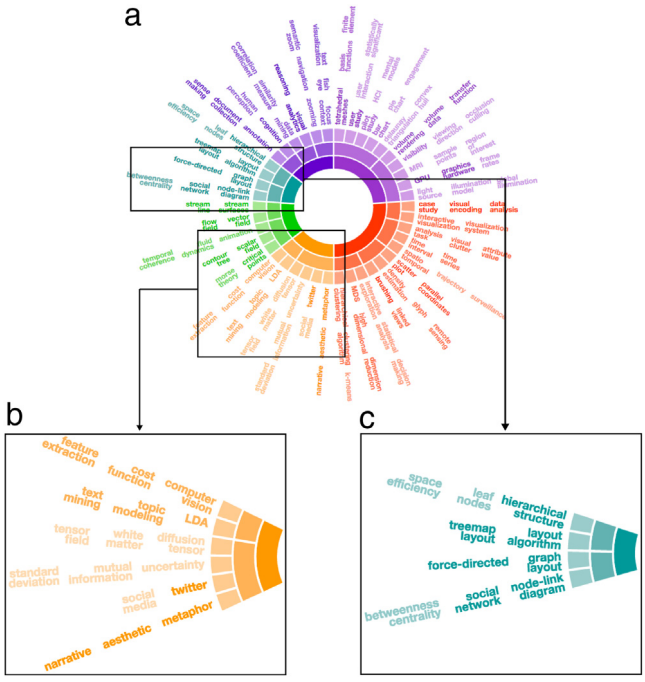


Fig. 10. Model examination with the topic visualization, discovering two topic branches with diverse semantic patterns. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

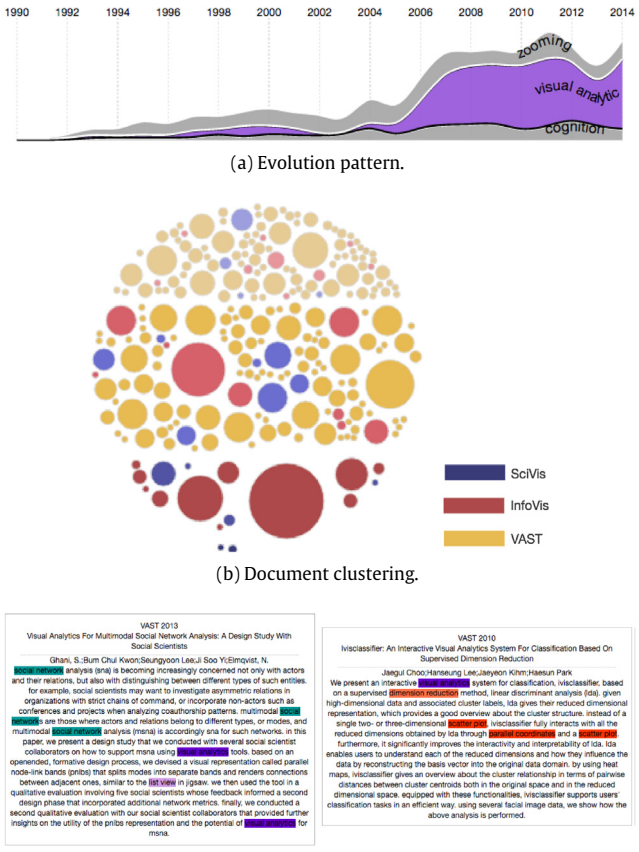
in the research field. We use the book *visualization analysis & design* (Munzner, 2014) and collect 450 terms in the concept index at the end of the book. We further screen the candidates based on both words and phrases. A vocabulary of 220 terms with high TF–IDF scores were finally selected.

With the vocabulary, a hierarchical topic organization of 63 topics is generated with HLTM. The topic hierarchy scales 3 levels, with 42, 15 and 6 on the first (top), second and third (bottom) level, respectively.

5.2. Case study I: assessing topic modeling results

The assessment of topic modeling results is an area of interest for a broad audience, including model designers and users. Although it is not difficult to find the meaning of a word-based topic, noticing the semantic patterns among multiple topics, for example, the hierarchical topics, is a challenging task. In this case, we show the effectiveness of our topic visualization in providing visual cues for users to identify structural and semantic patterns in the topic organization.

The hierarchical topics extracted from the corpus of IEEE VIS are presented in the topic view as shown in Fig. 10(a). First, we quickly perceived six topic branches occupying the domain, each of which was identified with a distinct color hue. Besides, the branches are not evenly weighted, where the light purple one and the orange one have covered up to half of the total specific topics. Then, we investigated the topics from the perspective of semantics, with a focus on the semantic closeness among hierarchical topics. We noticed some leaf topics showing a pattern of dominant impact to their parents, compared with brother topics. For example, all the three words, *metaphor*, *aesthetic* and *narrative* in Fig. 10(b) are colored with the saturation representing a highest level of abstraction, meaning a repeated usage in the description of hierarchical topics. Such patterns can invoke users' thinking from the perspective of how the HLTM constructs hierarchical topics. As an example, we assume this leaf topic residing near the "center" of the semantic



(a) Evolution pattern.

(b) Document clustering.

(c) Abstract marked with topic assignment.

Fig. 11. Investigation of the research field of "visual analytics" from multiple perspectives.

coverage for this branch, while the others have a relatively larger distance, viewed from a high-dimensional space. In contrast, the blue branch (Fig. 10(c)) shows a much more balanced patterns in terms of high-level semantics distribution. During this process, we gain a deeper understanding of topic model collect the observations to suggest potential extension of this model.

5.3. Case study II: investigating a research field

It is important for researchers to survey a particular field of study to summarize related works and gain insights. However, it is not an easy task to have a quick and comprehensive understanding of a research field, as a lack of domain knowledge and previous experience. In this case, we show the way to investigate a research field with VISTopic and present some domain-relevant findings.

We select the field represented by topic "Visual Analytics" as the targeted field. This topic, attracting us as a rapid growth during recent years, is presented in the evolution view (Fig. 11(a)). The visual patterns match our expectation as Visual Analytics Science and Technology (VAST), one of the children conference of IEEE VIS focusing on visual analytics, initiates for research at this specific field since the year of 2006. Furthermore, we verify our hypothesis about research venue with the document view. The bubble visualization in Fig. 11(b) confirmed this hypothesis by showing yellow as the dominate color. The visual cues also reflect the youth and vitality of this field as a large number of small-sized (meaning less cited in field) bubbles are presented. When zooming into the details of documents, we find many of them have colorfully annotated abstracts (Fig. 11(c)). This means that these documents are

Please cite this article in press as: Yang, Y., et al., VISTopic: A visual analytics system for making sense of large document collections using hierarchical topic modeling. Visual Informatics (2017), http://dx.doi.org/10.1016/j.visinf.2017.01.005.

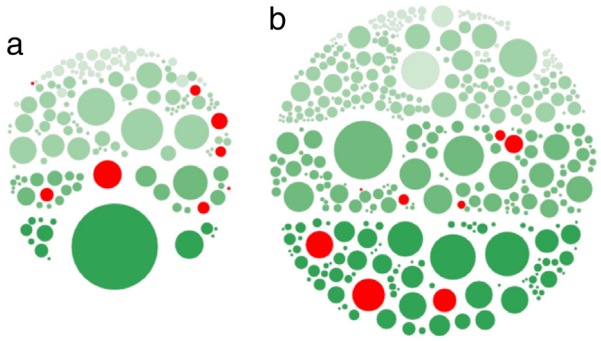


Fig. 12. Two data-driven views of topical documents with highlighting an identical author. (a) Documents labeled by the topic "graph layout", (b) Documents labeled by the topic "vector field".

composited in an interdisciplinary manner. It is not surprising to find this pattern as visual analytics techniques and methodologies solve problems across a variety of datasets and use scenarios.

5.4. Case study III: reasoning an author's topic

During the exploration of topical documents, we also obtain some interesting findings related to an author' topic evolution.

The name of an author, van Wijk, J.J., has appeared in the documents of multiple topics. Moreover, publications authored or co-authored by him have an overall high influence to the following work, as identified from the size of bubbles. We further identify an interesting pattern that the authors have published papers in both "graph layout"(Fig. 12(a)) and "vector field" (Fig. 12(b)), which are two diverse research field in VIS. The documents for "vector field" were mostly published in the 1990s, while those for "graph layout" started from 2000. Based on such visual cues, we make a hypothesis about his research evolution through these years. The domain expert confirmed the topic evolution of this author despite that patterns are more complex and noisy in real world.

6. Conclusion

In this paper, we present VISTopic, a visual analytics system to make sense of large document collections with the hierarchical latent tree model (HLTM). Interactive visualizations are designed and presented in linked views to help users simultaneously discover knowledge with the topical information, temporal information and bibliographic information.

In the case studies, we obtained interesting findings and verify our hypothesis about the research field of VIS. As a future work, it is desired to have a user study to explore the effectiveness of VISTopic for non-experts who may not have the knowledge of information visualization. Feedback of such user study should be valuable for us to identify the design that are less intuitive and suggest potential directions for improvement. Besides, the underlying techniques and approaches in VISTopic are general and not limited to VIS datasets, we will later apply it on others as well, such as historical newspapers and social media.

Acknowledgments

We would like to thank Professor Nevin L. Zhang and his Ph.D. student Peixian Chen from the department of computer science and engineering in HKUST for their kind technical support on topic modeling. In addition, we would like to thank Siwei Fu for preparing the pdf files of IEEE VIS corpus and discussion on data processing. This project is funded by a grant proposal (Ref: YBCB2009041-44) of Huawei Technologies Noah's Ark Lab. We are grateful for Huawei's generous support, as well as data acquisition. Finally, we thank the anonymous reviewers for their valuable feedback.

References

Blei, D., 2012. Topic modeling and digital humanities. J. Digital Humanities 2 (1), 8–11.

Blei, D., Griffiths, T., Jordan, M., 2010. The nested chinese restaurant process and bayesian nonparametric inference of topic hierarchies. J. ACM 57 (2), 1–30.

Blei, D., Ng, A., Jordan, M., 2003. Latent dirichlet allocation. J. Mach. Learn. Res. 3, 993–1022.

Choo, J., Lee, C., Reddy, C., Park, H., 2013. Utopian: User-driven topic modeling based on interactive nonnegative matrix factorization. IEEE Trans. Vis. Comput. Graphics 19 (12), 1992–2001.

Cui, W., Liu, S., Tan, L., Shi, C., Song, Y., Gao, Z.J., Qu, H., Tong, X., 2011. Textflow: towards better understanding of evolving topics in text. IEEE Trans. Vis. Comput. Graphics 17 (12), 2412–2421.

Dou, W., Wang, X., Chang, R., Ribarsky, W. 2011. Paralleltopics: a probabilistic approach to exploring document collections, in: IEEE Conference on Visual Analytics Science and Technology, pp. 231–240.

Dou, W., Yu, L., Wang, X., Ma, Z., Ribarsky, W., 2013. Hierarchicaltopics: Visually exploring large text collections using topic hierarchies. IEEE Trans. Vis. Comput. Graphics 19 (12), 2002–2011.

Fried, D., Kobourov, S., 2014. Maps of computer science. In: IEEE Pacific Visualization Symposium. pp. 113–120.

Ganesan, A., Brantley, K., Pan, S., Chen, J., 2015. LDAEXplore: Visualizing topic models generated using latent Dirichlet allocation. Technical Report, Dept. of Computer Science & Electrical Engineering, University of Maryland.

Ghahramani, Z., Jordan, M., Adams, R., 2010. Tree-structured stick breaking for hierarchical data. In: Advances in Neural Information Processing Systems. pp. 19–27.

Havre, S., Hetzler, E., Whitney, P., Nowell, L., 2002. Themeriver: Visualizing thematic changes in large document collections. IEEE Trans. Vis. Comput. Graphics 8 (1), 9–20.

Hofmann, T., 1999. Probabilistic latent semantic indexing. In: Proceedings of the 5th Conference on Uncertainty in Artificial Intelligence. pp. 50–57.

Hong, L., Davison, B. 2010. Empirical study of topic modeling in twitter, in: Proceedings of the 1st workshop on Social Media Analytics, pp. 80–88.

Isenberg, P., Heimerl, F., Koch, S., Isenberg, T., Xu, P., Stolper, C., Sedlmair, M., Chen, J., Möller, T., Stasko, J. 2015. Visualization publication dataset. Dataset: URL http://vispubdata.org/. Published Jun. 2015.

Isenberg, P., Isenberg, T., Sedlmair, M., Chen, J., Möller, T. 2014. Toward a deeper understanding of visualization through keyword analysis, Technical Report INRIA.

Jones, K., 1972. A statistical interpretation of term specificity and its application in retrieval. J. Doc. 28 (1), 11–21.

Kamada, T., Kawai, S., 1989. An algorithm for drawing general undirected graphs. Inform. Process. Lett. 31 (1), 7–15.

Landauer, T., Foltz, P., Laham, D., 1998. An introduction to latent semantic analysis. Discourse Process. 25 (2–3), 259–284.

Liu, T., Zhang, N. and Chen, P. 2014. Hierarchical latent tree analysis for topic detection, in: European Conference on Machine Learning and Knowledge Discovery in Databases, pp. 256–272.

Mourad, R., Sinoquet, C., Zhang, N., Liu, T., Leray, P., 2013. A survey on latent tree models and applications. J. Artif. Intell. Res. 47, 157–203.

Munzner, T., 2014. Visualization Analysis and Design. CRC Press.

Schulz, H.-J., 2011. Treevis. net: A tree visualization reference. IEEE Comput. Graph. Appl. 31 (6), 11–15.

Smith, A., Hawes, T. and Myers, M. 2014. Hiérarchie: Interactive visualization for hierarchical topic models, in: ACL Workshop on Interactive Language Learning, Visualization, and Interfaces, pp. 71–78.

Stasko, J., Zhang, E., 2000. Focus+context display and navigation techniques for enhancing radial, space-filling hierarchy visualizations. In: IEEE Symposium on Information Visualization. pp. 57–65.

Sun, G., Wu, Y., Liu, S., Peng, T., Zhu, J., Liang, R., 2014. Evoriver: Visual analysis of topic coopetition on social media. IEEE Trans. Vis. Comput. Graphics 20 (12), 1753–1762.

Titov, I., McDonald, R. 2008. Modeling online reviews with multi-grain topic models, in: Proceedings of the 17th international conference on World Wide Web, pp. 111–120.

Wei, F., Liu, S., Song, Y., Pan, S., Zhou, M., Qian, W., Shi, L., Tan, L., Zhang, Q., 2010. Tiara: a visual exploratory text analytic system. In: Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. pp. 153–162.

Xu, P., Wu, Y., Wei, E., Peng, T.-Q., Liu, S., Zhu, J., Qu, H., 2013. Visual analysis of topic competition on social media. IEEE Trans. Vis. Comput. Graphics 19 (12), 2012–2021.

Yang, T.-I., Torget, A., Mihalcea, R., 2011. Topic modeling on historical newspapers. In: ACL Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities. pp. 96–104.