SCIENCE CHINA

January 2013 Vol. 53 No. 1: 1–18
doi:

Topic Hypergraph: Hierarchical Visualization of Thematic Structures in Long Documents

WANG Guizhen[1], WEN Chaokai[1], YAN Binghui[1], XIE Cong[1], LIANG Ronghua[2] & CHEN Wei[1]*

[1]*State Key Laboratory of CAD&CG, Zhejiang University, Hangzhou* 310058*, China,*
[2]*School of Information Engineering, Zhejiang University of Technology, Hangzhou* 310014*, China,*

Received November 3, 2012; accepted January 4, 2013

Abstract Thematic information of a long document (i.e., a novel) can be multi-faceted: an interleaving of multiple topics, a sequential evolution of a set of themes, or a crossing superimposition of topics and themes. Conventional topic-based visualization approaches are inefficient to capture this complicated thematic structure. This paper introduces a novel topic-based model, called the topic hypergraph, that characterizes the thematic structure of a long document with a hypergraph representation. Each hypergraph node represents a unique document piece, and encodes its theme as a composition of multiple topics. Two types of relationships among nodes are modeled: an edge that connects two consecutive themes to present their sequential transition, and a hyperedge that encodes a topic. The new representation is essentially a 2D reformulation of the linear streamgraph representation, and can be adaptive by constructing a multi-level hierarchy. We design a suite of visualization and interaction tools to allow users to interactively analyze the theme evolution, theme diversities, and topic interleaving. Our approach is also suitable for comparing multiple long documents.

Keywords Topic Model, HyperGraph, Document Visualization, Text Data Analysis

Citation Wang G Z, WEN C K, Yan B H, et al. Topic Hypergraph: Hierarchical Visualization of Thematic Structures in Long Documents.

1 Introduction

Many efforts have been paid on topic-based text mining, linguistic analysis, and document visualization [1]. For instance, the pioneering work of document visualization [2] employs the wavelet technique to extract a set of topics with respect to the word frequency, and visualizes them as a set of hierarchical islands. Its goal is to present the *relationships* among varied topics within a compact layout. More sophisticated topic-based representations can be employed, e.g., the widely used topic model [3] and its variations [4]. By explicitly encoding the *sequentiality* of a document as a time axis, the thematic evolution in terms of topics can be visualized. The ThemeRiver technique [5] visualizes the theme distribution along time with a streamgraph. Many recent work leverage the topic model and the ThemeRiver technique to express a text stream, like the TIARA system [6]. Yet, it is hard for users to infer the semantic relations at different time points in a 2D visualization because the timeline representation is linear and inefficient for cross reference. A narrative structure [7] provides a higher level description, but still requires a time stamp for visualization.

This situation is much more challenging for a very long document with a more complicated thematic structure than a short document. Despite numerous efforts on document visualization, little attention has been paid to the specific care for a long document. Concerning the complexity, topics can be used as the primary primitives for

*Corresponding author (email: chenwei@cad.zju.edu.cn)

analyzing a long document. The main challenge is to characterize the topic evolution without the loss of the description to the thematic organization in a document. The thematic organization can be as simple as a chapter-based title list, or be as complicated as a storyline that maintains a plot with several stages, such as exposition, conflict, rising action, climax, a falling action and final resolution [8]. Therefore, of great importance is the employment of an appropriate representation for long documents.

Few existing approaches can completely fulfill the task. For instance, the topic island representation [2] is not designed for presenting the order of the topic islands. The TIARA system [6] concentrates on the comparison of a series of text flows. It is, however, difficult to analyze repetitive patterns or similar themes that appear at distant time points. Recently proposed FacetAtlas representation [9] enables multifaceted visualization for rich text corpora. It focuses on exploring the cross-relationship hidden in multivariate datasets.

This paper presents our efforts on representing and analyzing a very long document by leveraging the thematic information built with the topic model [3]. The topic model can be used to generate a set of topics for a collection of documents, and their probability distributions in each document. By regarding a long document as a collection of document pieces, a set of topics can be produced, together with their probabilities in each piece. The probabilities of all topics in a document piece characterize its thematic information, and thus form a theme with respect to the document piece. Thus, representing a long document can be regarded as building a data structure based on the topic set and the themes of all document pieces.

We propose *the topic hypergraph*, a new representation that employs themes as nodes and contains two types of edges: a hyperedge that encodes a topic and its relationships with related themes in different locations, and a normal edge that connects two consecutive themes to express the sequential transition. Essentially a topic hypergraph is a 2D reformulation of the streamgraph representation, and yields a topological description to the thematic information. By building different levels of thematic structures, a topic hypergraph can be hierarchically constructed to express diverse granularity of thematic information. Figure 5 shows an example of visualizing a long novel.

We additionally design and implement a set of visualization and interactive exploration techniques, like dynamic context switching and focus+context. Our integrated system allows users intuitively study the dramatically switched parts in a long document and compare multiple long documents. Moreover, it can simultaneously characterize the document sequentiality with the edges, and the topic interference with the hyperedges. Specifically, it enables the following features:

- **Navigating the theme proximity** An embedded 2D layout of the sequence of themes can help the user investigate the proximities among different document pieces;
- **Exploring the theme transition** Connecting different themes with edges characterizes the document sequentiality and the variations of themes;
- **Studying the topic interference** The topics are distributed among various document pieces, and may interleave with others in various document pieces;
- **Comparing multiple long documents** By constructing a topic hypergraph for multiple documents, their differences with respect to the topics and theme distribution can be clearly shown.

The rest of this paper is organized as follows. We review related work in Section 2, present the new representation in Section 3, and describe its visualization and interaction design in Section 4 and Section 5 respectively. Experimental results and analysis are given in Section 6. Finally, we summarize this paper and highlight the future work in Section 7.

2 Related Work

A large amount of work has been devoted to document analysis and visualization. We briefly review the ones on the analysis and visualization of themes and topics.

2.1 Analysis

Many text mining approaches focus on mining the latent thematic information of documents. Methods of document theme analysis, such as the *bag of words* representation [10] and *LSI* [11], employ the word term frequencies to encode the overall thematic information. The wavelet analysis [2] is used to describe thematic information in lengthy documents. However, these methods can only represent one text part with only one mixed thematic subject. To extract the latent thematic information of the sequence relationships of words, a list of graphic probability models like LDA [3], PLSI [12] and Theme Topic Mixture Model (TTMM) [13] are proposed. Instead of using one topic, this kind of methods represent documents as a mixture of multiple topics.

The thematic organization of a document typically follows certain routine, which is diverse in different cases. Novels and scientific literatures are well organized to precisely conveying the thematic information: "Stories and novels consist of three parts: narration, which moves the story from point A to point B and finally to point Z." [14]. The journal articles, evolve with another routine [15]: "the anecdotal lead — an initial story, often involving dialogue and characters, that presents a microcosm of the larger news story — and the nut graf — a paragraph explicitly describing the news value of an article."

Regardless of which type of thematic organization, the thematic parts are inherently related in a document: themes in remote parts are of great difference, and those in consecutive pieces are similar. Meanwhile, a long document evolves with multiple topics that may be shared by multiple documents [1]. Therefore, more complicated and hierarchical representations are required to represent a long document.

2.2 Visualization

The tag cloud [16] is perhaps the most common document visualization technique. It provides a visual summary of a collection of texts by showing the tag statistics with varied font sizes. However, this technique is not good at conveying trends between multiple tag clouds, even with the recently proposed sparkline enhancement [17]. The timeline [18] and ThemeRiver [5] techniques are two means to depict the thematic variations over time by using a time axis. For the purpose of exploring large corpus, the TIARA system [6] extracts topics and depicts the topic evolution over time by leveraging the ThemeRiver technique. Its variation [19] uses river metaphors to represent multiple facets of time-series documents along a timeline.

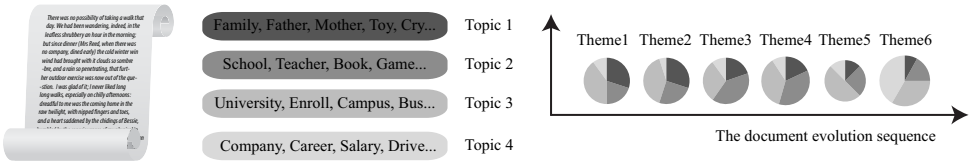
Concerning a long document, it is of great importance to visualize the topic transition. By taking periodical outputs of the clustering operations for a time-series of documents, the T-scroll technique [20] visualizes the relationships between clusters as scrolls. The narrative visualization approach [7] combines the ThemeRiver technique with the clustering methods, and allows for placing news stories in their historical and social context. By counting the word frequency, Mao et al. [21] propose a framework to hierarchically and sequentially represent a document as smooth curves based on the locally weighted bag of words (Lowbow) representation [10].

Several methods have been proposed to analyze the local relation patterns. For instance, the Topic Islands technique [2] uses the wavelet analysis method to summarize the characteristics of the narrative flow in the frequency domain. The Word Tree approach [22] supports the exploration of the prefix relation between words, and the Phrase Net technique [23] allows for effective relationship definition. None of these techniques focus on the sequential local pattern in a long document. By capturing the repetition of exact or similar expressions in a corpus, the FeatureLens approach [24] generates several levels of granularity to visualize interesting text patterns. More recently, the FactAtlas [9] uses colored links to represent connections between different document facets.

Projection-based techniques [25] are widely used to visualize relationships in a document collection. The key idea is to map them into a 2D space with respect to the document proximity. These methods can reveal the patterns in the document scale, while they are intractable for illustrating the relationships in a long document.

3 The Topic Hypergraph Model

The research on document analysis [14] indicates that a document can be regarded as a composition of a sequence of document parts, or a crossing evolution of multiple storylines. Following the notations of the topic model [3], two kinds of thematic entities used in our approach are described below.



(a) A document (b) Topics (c) Representing a document evolution with themes

Figure 1: Illustration of topics and themes, where a document contains 4 topics and 6 themes.

Topic A topic is a thread that runs through several document pieces. Take the novel "Jean Eyre"[26] for example: the romantic love of the main characters is a thematic storyline through the entire book, and the family love is anther storyline in some pieces. Multiple topics combine together to construct the thematic content of the underlying document. Figure 1 (b) shows four topics and their keywords, since each topic is represented with a probability distribution over the document vocabulary.

Theme A theme is the thematic summarization of a document piece. It is a proportional mixture of the topics (see Figure 1 (c)). In this paper, a theme is regarded to be identical to a document piece. As such, the collection of themes is an ordered set, and connecting them with their orders in the document can reflect the thematic evolution of a document.

A popular way to encode the thematic evolution (Figure 2 (a)) of a long document or a document corpus is the streamgraph representation (e.g., the ThemeRiver [5]) (see Figure 2 (b)). This representation flattens the thematic organization within a linear time axis, and thus is not suitable for characterizing complex thematic structure such as the topic correlation. To express the thematic organization and topic interferences of a long document, we leverage a hypergraph to organize the extracted themes and topics, to yield a 2D reformulation of the streamgraph representation (see Figure 2 (c)). This representation can be hierarchical by recursively merging the nodes in terms of their thematic proximities, as illustrated in Figure 2 (d). Below we elaborate the details.

3.1 The Representation

This paper defines a topic hypergraph as a combination of cluster graph [27] and hypergraph, that is, $THG = (CG, T)$, CG is a compound graph consisting of a hypergraph, HG, and a simple directed graph, G. T is a rooted tree, depicting the hierarchical structure, whose leaf nodes are the finest themes, nodes of both HG and G. In the tree T, upper level nodes are coarser themes made of a set of child themes, and its root is the whole theme of the long document. The simple directed graph G denotes sequential evolution of themes. In a certain level i, G_i has an edge $e_i = (\alpha, \beta)$ if a child theme of α connects to a child theme of β. Mathematically, a hypergraph is expressed as $HG =< V, E >$, where V is the theme node set, and E, hyperedges, denotes topics that distribute in a set of non-empty subsets of V [28]. Compared to a graph edge that contains a pair of nodes, a hyperedge is an arbitrary set of nodes, and can therefore contain an arbitrary number of nodes. In a level i, a topic hypergraph, HG_i consists of nodes, themes in the corresponding level, and topics. The detail interpretation of topic hypergraph is as followed:

- Setting each theme as a node;
- Setting each topic as a hyperedge which includes the themes, with respect to each of which the distribution proportion of the topic is non-zero;
- Setting the weight of a hyperedge with respect to a node as the corresponding distribution proportion of the topic in the node (theme);
- If the weight of a hyperedge with respect to a node is lower than a given threshold (e.g., 0.05), the node is excluded from the hyperedge;
- Making a graph edge between a pair nodes which is consecutive with respect to their evolution orders in the document.

Table 1 shows the conversion from a topic model [3] to a topic hypergraph. In general, there are a large amount of themes (document pieces) for a very long document. From the viewpoint of linguistics, a theme can be divided into smaller ones [14]. To provide adaptive exploration, a topic hypergraph can be hierarchical, meaning that a set of nodes can be grouped into a larger node. The construction algorithm will be given in Subsection 3.2.

(a) Topic Model

#Theme	Topics		
	$t1$	$t2$	$t3$
*theme*1	0.1	0.5	0.4
*theme*2	0.5	0.2	0.3
*theme*3	0.4	0.1	0.5

$\rightarrow$

(b) Topic Hypergraph

#Node	Hyperedges		
	he_1	he_2	he_3
no_1	0.1	0.5	0.4
no_2	0.5	0.2	0.3
no_3	0.4	0.1	0.5

Table 1: The conversion from a topic model to a topic hypergraph.

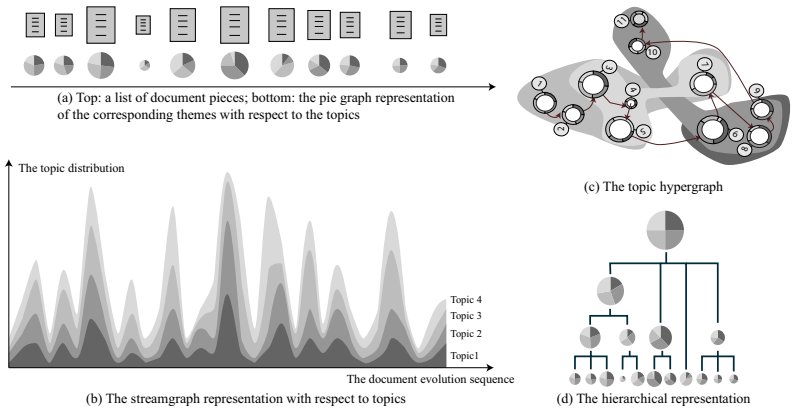


Figure 2: Comparison of the streamgraph representation and the topic hypergraph model.

A hyperedge (topic) indicates important thematic information:

- **The distribution** The layout of the involved themes indicates the distribution of its corresponding topic in the document. Based on the distribution, we can see whether a topic is contained in the entire document, or only exists in several document pieces. In Figure 1 (b), the third topic about the work activities uniformly distributes in the document. In contrast, the fourth topic about jobs mainly appears in the last piece.

- **The importance** The significance of a topic with respect to a theme varies with its proportion or other optional measures. Topics that have high distribution proportions in themes are regarded as important topics, and others are secondary.

- **The correlation** The co-occurrence of multiple topics (hyperedges) can be used to reveal their relationships and the inferences with the themes. A strong correlation indicates that the topics jointly advance the thematic revolution, while a week correlation may reflect a large thematic variation.

Representing themes as nodes helps reveal the following properties:

- **The proximity** By using the construction approach described below, the proximity between two themes implies the closeness of their thematic information. Normally, consecutive themes tend to have a stronger correlation than the ones that belong to different document pieces. However, in some cases, themes that are not adjacent may share a common topic and hence show a strong connection.

- **The sequentiality** The adjacency of two themes usually means that they may share some common topics. Likewise, the thematic transition in the entire document can be revealed by studying the theme sequence of the topic hypergraph.

- **The hierarchy** The thematic evolution of a long document can be summarized adaptively. In the coarsest granularity, the entire document can be summarized into a theme, and in a finer granularity, multiple themes characterize how the document evolves. In Figure 2 (d), the themes are organized into a hierarchical tree, of which each leaf node is a theme, and a child node denotes a subset of its father node. By nature, a hyperedge (a topic) can be represented as a collection of nodes of the tree. Adaptively exploring the topic hypergraph (Figure 2 (c)) equals the jumping among different tree nodes.

3.2 The Construction

Constructing a hierarchical topic hypergraph involves two stages: the generation of topics and themes, and the construction of a hierarchical hypergraph.

3.2.1 Topics and Themes

The complicated thematic organization of a long document elicits a partition of the document to favor a level-of-details description to the thematic evolution. Thus, the first step of the construction is to segment a long document into a set of non-overlapping pieces to generate the finest levels.

The scheme of the document partition determines the granularity of constructed topics and themes, and consequently the expressiveness of the resulting topic hypergraph. Simply speaking, the more detailed the document is segmented, the higher the granularity of a topic is. In practice, the segmentation scheme can be varied: manual subdivision; uniform partition; leveraging existing document structure (e.g., the chapter structure); employing advanced document analysis approaches [29, 30, 31]. The partition number is configured to be adequately large to accommodate a hierarchical topic hypergraph.

After the partition, a list of topics and themes are extracted by applying the LDA algorithm to the document pieces [3]. It yields a set of topics, and their distribution probabilities in each document piece, which compose the themes in distinctive pieces.

3.2.2 Hierarchical Tree

We build a hierarchical topic hypergraph in a bottom-to-up fashion under a unique theme proximity measure. By nature the themes of the finest levels are the leaf nodes. A higher level can be built by iteratively merging nodes from bottom to up.

Supposed that there are n topics and m themes in the finest level, meaning that n nodes and m leaf nodes exist in the topic hypergraph. Each topic t_i is represented with a probability distribution $\{t_{i_k}, k = 1,2,3,...,N_z\}$ over the document vocabulary whose size is N_z, and each theme T_j can be represented as the probability distribution $\{w_{ji}, i = 1,2,3,...,n\}$ with respect to the topics in the corresponding document piece (see Figure 2 (c)). We define the proximity d_{jl} of two themes T_j and T_l as: $d_{jl} = \sqrt{\sum_{i=1}^{n}(w_{ji} - w_{li})}$.

Based on the theme proximity, a hierarchical tree is iteratively constructed by applying the K-means clustering algorithms to the nodes of the current level and choosing the clustering center as the upper node. Specifically, the merged node in an upper level is represented as a weighted sum of the probability distributions of its M child nodes:

$$\{wm_i = \frac{\sum_{k=1}^{M} N_k \times w_{ki}}{\sum_{k=1}^{M} N_k}, \;\; i = 1,2,3,...n\} \tag{1}$$

where N_k denotes the word count in the document piece corresponding to the kth child node, w_{ki} is the probability of the kth child node with respect to the ith topic.

4 Visual Design

Varied views to a topic hypergraph can be generated by choosing different hierarchies of the nodes. For instance, Figure 3 illustrates a selected hierarchy of a topic hypergraph which consists of three topics and seven document pieces. The shown hierarchy includes four nodes, three hyperedges and four edges. The visualization indicates that two topics are contained in Theme3. Theme2 has multiple document pieces (namely, 2,3,4), which can be further explored by switching to a deeper hierarchy.

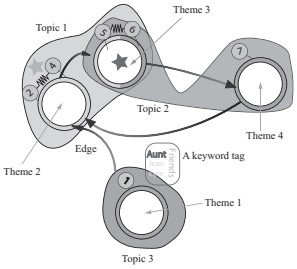


Figure 3: Conceptual illustration of a topic hypergraph.

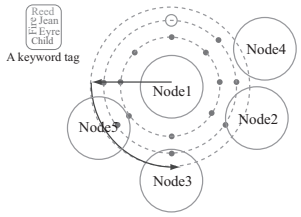


Figure 4: To locate a keyword tag of Node1, a sequence of locations is tested against four nodes, in the order of a radial concentric layout around Node1 (outwards and count-clockwise). The solid points are detected as invalid, and the empty circle is an appropriate location.

4.1 The Layout

The layout of a topic hypergraph is guided by the thematic structure, which can be formed in different views. We employ three stages to layout the hypergraph for the purposes of showing the correlation of different topics, and the interference among topics and themes.

- **Positioning the Hyperedges** The 2D locations $\{tp_i, i = 1,2,3,...,n\}$ of the hyperedges (topics) are determined by embedding their corresponding high-dimensional vectors $\{t_{ik}, k = 1,2,3,...,N_z\}$ (see Subsection 3.2.2) into a 2D space using the nonlinear mapping based MDS (multi-dimensional scaling) algorithm [32].

- **Positioning the leaf nodes** The position of each theme Tp_j in the finest level (i.e., the leaf nodes) are computed as $Tp_j = \sum_{i=1}^{n} w_{ji} tp_i$, where tp_i denotes the 2D position of the ith topic, and w_{ji} (see Subsection 3.2.2) is the weight of the jth theme with respect to the ith topic.

- **Positioning other nodes** The location of an intermediate node is set to be the center of the locations of its child nodes.

4.2 The Visual Encoding

Nodes A theme is represented with a ring-like circled pie graph which encodes the probability distributions of related topics in the theme. The example (marked with a red star) shown in Figure 3 indicates that two hyperedges cover Theme3 while Theme2 only relates to one hyperedge. Each color-filled small circle identifies a document piece contained in the theme. For clarity, document pieces (e.g., labeled as 2,3,4) which belong to certain node and are adjacent in the document are encoded with two circles with a two-ending spring (see the example marked with a blue star). Each node can be independently expanded to show the detailed thematic information if it is not a leaf node.

Keyword Tags The keywords of a node (theme) are computed by selecting the words whose term frequencies in the corresponding document piece are high. The Wordle technique [33] is used to fill a keyword tag within a small rectangle, where the word size encodes the frequency. To minimize the overlap between a keyword tag and each

node, an appropriate location is determined by heuristically searching the neighboring positions around the nodes and choosing the one with the minimal coverage with all surrounding nodes. Figure 4 illustrates this process.

Hyperedges A topic is encoded with a color-filled iso-contour that illustrates the set relationship of the involved themes. The contours are generated by employing the Bubble Sets technique [34] to the set relationship among the topics and themes.

Edges The edges connecting two adjacent nodes effectively depict the thematic evolution information. For an edge, a Catmull-Rom spline is constructed based on the 2D positions of four consecutive nodes. Two techniques are employed to avoid possible visual clutter. First, if two nodes are very close, their edge is not drawn. Second, if an edge intersects existing edges, its shape is heuristically adjusted to minimize the intersection.

5 Interaction Design

We design an integrated system with a suite of interaction toolkits to support effective exploration of a long document. As shown in Figure 5, the interface includes three views: a main panel showing the topic hypergraph (left), a topic wall showing the lists of topics (top right), and a view showing the underlying document (bottom right). All views are linked: whenever there is a selection of an item in a view, corresponding or associated items in other views are triggered. For instance, when a node is chosen, its keyword tag appears around the node, and the corresponding document piece is automatically shown in the bottom right view.

The system provides the following interaction features to favor comprehensive understanding of the underlying document:

Text Annotation A topic hypergraph denotes a topological structure, and is abstract for the user to understand the thematic information. Two types of text annotations are employed. The first one is the keyword tag that shows the important words in a theme with the Wordle technique [33]. It is displayed on the main panel when a theme is selected by the user. The other is the topic wall that shows topics in a block. Each block representing a topic has the same color as the corresponding hyperedge. by clicking a topic block.

Highlight A very long may contain a lot of themes (nodes), which may cause heavy visual clutter. The adaptive visualization of a topic hypergraph cannot completely address this problem. Highlighting the ones with a large significance is a natural choice to help the user focus on the most interesting items. For instance, when a node is chosen, its linked edges and hyperedges are enhanced with a high luminance, and are arranged on front of other drawing primitives. With this simple scheme the local thematic variation around the underlying document piece is depicted.

Hierarchical Theme Exploration A useful feature is the interactive exploration of a topic hypergraph by specifying individual hierarchies for different nodes. The interaction is natural: when a node is chosen, its child nodes are automatically expanded. Likewise, a set of nodes can collapse into their father node. The expansion or the collapse is smoothly visualized to show visually pleasing transition. The switching between different hierarchies allow the user observe the theme evolution and topic interference among different hierarchies.

Topic Significance The probability distribution of a topic is changing along the thematic evolution in a long document, and then themes in which the probabilities are non-zero are nodes for constructing a hyperedge for the underlying topic (see Subsection 3.1). In terms of the visualization, the iso-contouring technique used in the Bubble Set representation [34] can show the set relationships of a hyperedge, but it is inefficient to show the significance of a topic with respect to a theme (a specific document piece). Our system allows the user to interactively specify a threshold to modulate the set relationship of a hyperedge (topic) with the associated nodes: only the ones in which the probabilities of the topic are higher than the given threshold are regarded as valid and be included in the iso-contouring process. Progressively changing the threshold can reveal how a topic is related to different document pieces, as shown in Figure 6 and the video demonstration.

6 Case Study

We have designed and implemented an integrated system with Java and Prefuse toolkit [35]. To demonstrate the capabilities and usefulness of our approach, we perform two case studies on three books: *Jane Eyre* by Charlotte

Brontë [26], *Living History* by Hillary Rodham Clinton [36] and *My Life* by Bill Clinton [37]. Further, we examined the efficiency with a pivot user study.

6.1 Jane Eyre

Jane Eyre describes the life of the protagonist, Jane Eyre, from her childhood to a mature young woman. The word count of the entire document is 185165. This number becomes 31273 after removing the stop words and words that only occur for one time. Two kinds of thematic structures are examined: chapter-based organization determined by the author; the structure obtained with the TextTiling algorithm [30].

Chapter-based Structure

There are 38 chapters in the document, each of which is regarded as a leaf node. Seven topics are extracted by using the LDA algorithm [3] (see Figure 6 (a)). The topics that are involved in the first half of the document, like Topic 5, Topic 6 and Topic 7, have minor influence on the thematic evolution. Instead, Topic 1, Topic 2, Topic 3 and Topic 4 that describe the life after Jane's growth, play a major role.

The generated topic hypergraph has five hierarchies. Figure 6 (b) shows the root node of the topic hypergraph. It simply shows the distribution of seven topics, and lists the set of all document pieces, i.e., 38 chapters. A finer hierarchy depicted in Figure 6 (c) summarizes the distributions of seven topics in the entire document. Figure 6 (d) demonstrates the finest hierarchy with the topic significance threshold of 0.2, and is more informative in terms of the thematic evolution, compared with Figure 5. For instance, the thematic information in the region marked with the red star is expanded from the first node of Figure 5. From the visualization in Figure 5, we can find that Topic 5 and Topic 6 construct the first four document pieces. The finer visualization in Figure 6 (d) reveals that Topic 6 is dominant in the first three pieces, while Topic 5 mostly exists in the fourth document piece.

Figure 6 (e) is generated from Figure 6 (d) by setting a topic significance threshold as 0.4. A higher threshold offers the opportunity to filter less significant themes with respect to a topic. For instance, by comparing Figure 6 (d) and Figure 6 (e), it is clear that the main components of Topic 7 are the themes marked with the blue stars, not the ones marked with the green star in Figure 6 (e). Likewise, Topic 1 plays a more important role in Chapter 25 than Topic 7.

Figure 5 depicts a middle hierarchy of the constructed topic hypergraph. It is easy to find that Topic 6 describing Jane's childhood dominates the first four chapters. The following five chapters talk about the life at Lowood school, and largely relate to Topic 5. Then, Topic 1, Topic 2, Topic 4 and Topic 7 contribute to the content about the life in Thornfield. In particular, Topic 2 about the love mostly appear in the lower half of this part, indicating that the love between Jane and Rochester becomes clear in that period. The next chapters describe the living with her cousins. Here, a long edge connecting to the node of Chapter 21 indicates a dramatic semantic variation (marked with the red star). This is consistent with the fact that in Chapter 21 Jane left Thornfield to visit her dying aunt who is an important character in the first four chapters, while the themes adjacent to Chapter 21 talk about Jane's life in Thornfield.

The hyperedges represented with iso-contours imply how topics correlate with each other in different document parts. For instance, the node of Chapter 10 (marked by the blue star) is covered by two hyperedges of Topic 5 and Topic 6 (see Figure 5). The ring-like circled pie graph records the distribution of Topic 5 and Topic 6 in Chapter 10. Indeed, two parts of Chapter 10 describe the life at Lowood school, and the visit of Jane's relatives, which relate to Topic 5 and Topic 6, respectively. Chapter 10 is actually a transition to next chapters, where Topic 5 exists. A similar analysis to Chapter 38 (the end of the novel, marked with the green star) finds two topics: the marriage and love between Jane and Rochester (Topic 2), and her cousins (Topic 7).

Thematic Partition

By using the TextTiling algorithm [30], the novel is segmented into 46 pieces. Based on the partition, 9 topics are extracted with the LDA algorithm [3], as shown in Figure 7 (a). Basically these topics are similar to 7 topics extracted from the chapter-enabled structure. Figure 7 (b) shows the coarsest hierarchy with a topic significance threshold of 0.1. The proximities and sequentiality of themes, and the relationships between topics and themes are clearly depicted, like the correlations among Topic 2, Topic 7 and Topic 9 in the 46th node. Meanwhile, the 35th and 42th document pieces are thematically similar, and mainly relate to Topic 11. Figure 7 (c) shows the finest result compared with Figure 7 (b), and hence is more expressive to depict subtle thematic evolution and the interferences between topics and themes (e.g., the region marked with the green star). For instance, the first 34

document pieces (1-34) in Figure 7 (b) is represented by one theme that touches Topic 1,2,5,6 and 9. In contrast, the distributions of these topics are clearly displayed in Figure 7 (c). Alternatively, Figure 7 (d) visualizes the same hierarchy as Figure 7 (c), but employs a higher topic significance threshold of 0.3. As a result, less important topics are culled, and the topic interference is clarified, as demonstrated in the red circled regions.

6.2 Living History and My Life

We examined the capability of our approach for comparing multiple long documents by analyzing two books *My Life* and *Living History*. *My life* has 56 chapters, and 468490 words and 80471 keywords after removing the stop words. *Living History* has 38 chapters, and 221744 words and 25176 keywords after removing the stop words.

My Life describes the story of Bill Clinton from his childhood to the president relieving. Likewise, *Living History* spans the period of the childhood and the ending of the first-spouse of Hillary Clinton. It is apparent that two books are different in various aspects, e.g., the writing style, the used words, and the thematic organization. For the sake of comparison, for each book we remove the words whose total counts are lower than 5. In addition, we segment *My life* into 41 pieces by means of the TextTiling algorithm [3] to make the node numbers of both books approximately identical.

By applying our approach to the composition of two books, 14 topics are extracted, as shown in Figure 8 (a). In the visualization, the document pieces (small circles) from *My Life* are encoded with the blue color, and the ones from *Living History* are in light green.

Figure 8 (b) shows the coarsest hierarchy of the constructed topic hypergraph. Studying the visualizations help discover the similarity and difference between two books. First of all, the topic list indicates that two books focus on politics, and their thematic structures are similar. In addition, the thematic evolutions of two books share some common progress, like the transition among the themes indicated by three circles in red. The topics that cover these themes are commonly used in both books: Topic 7 about the personal growth is mentioned in the beginning of both books; Topic 12 covers the themes in the lower half parts of both books.

Figure 8 (c) shows a middle hierarchy with a topic significance threshold of 0.1. It demonstrates more topic co-occurrence in both books, like Topic 7 and Topic 12, and other topics in varying degrees. Simply speaking, the topics can be classified into two categories: the ones that are shared by both books (marked with the green stars), and the ones that specialize in one book. For instance, Topic 9 frequently appears in *My Life*, while the topic about the Women's rights is mostly mentioned in *Living History*.

6.3 User Evaluation

A user evaluation was performed by 12 volunteers. The test document is the novel *Jane Eyre* which contains 38 chapters. A topic hypergraph with 7 hyperedges, 38 nodes, and 5 hierarchies was constructed. Each subject was asked to finish the test in half an hour. Before the test, the subjects were trained in one hour to understand the task and get familiar with the system.

All subjects are non-native English speaker, but their English skills are in the middle level. All subjects major in computer science, and are classified into two groups. Group A consists of five male and one female graduate students which have some knowledge about information visualization. Group B includes two male and four female graduate students which have no experience on visualization. Most subjects have no knowledge on the tested book.

Based on the observations that sequential themes in a long document may contain large semantic change, three questions (Q1, Q2, Q3) were designed to validate whether our system can help the user detect the themes with large thematic variations, segment the document with respect to the topics, and list the transitional themes, respectively. Additional two questions (Q4, Q5) concerning the distributions and co-occurrence of topics were asked: ordering seven topics according to their importance in the document; finding the document pieces that describe the love between Jane Eyre and Rochester.

The overall feedback from the subjects is quite positive. Some user experiences were additionally recorded. First, the node locations play an important role in the document understanding. Second, navigating the hierarchical topic hypergraph from bottom to top greatly enhances the investigation of the thematic organization. Concerning the detection of the thematic variation, the subjects generally choose the document pieces which connect to long edges, or the themes that belong to at least two topics. Moreover, searching the text with the keywords of certain

topic can help the user quickly detect the themes to which the topic relates, and consequently locate the document pieces. Finally, 83% subjects thought that edges and hyperedges are equivalently important.

The average percentages of accuracy listed in Table 2 indicate that two groups achieve similar scores that range from 51.19% to 82.5%. In addition, the average percentages of accuracy for the first three questions concerning the themes are higher than those of the last two questions on the topics. This indicates that the topic hypergraph model can assist the user quickly understand the thematic structure of documents in non-native language. It also shows that our approach can be easily mastered by non-professional users.

Question	Group A	Group B	Average
Q1	55%	56.7%	55.83%
Q2	85%	78.3%	82.5%
Q3	58.3%	60%	59.2%
Q4	50%	52.4%	51.2%
Q5	56.7%	43.3%	51.7%

Table 2: The accuracy of answering five questions for two groups of subjects.

7 Conclusion and Future Work

A long document (e.g., a book) usually contains diverse thematic information and complex thematic structure. This paper describes our efforts on the representation, construction, visualization and exploration of a long document with a novel topic hyerpgraph model. A topic hypergraph is inherently compatible with document analysis approaches because its primitives are extracted with the established text mining techniques. By employing a hierarchical structure, a topic hypergraph provides a topological description to the thematic information in an adaptive mode. Compared with the well-studied streamgraph representation which is a flattened and linear version of the topic model, the topic hypergraph is more comprehensive, and allows for complex analysis and exploration tasks. The integrated document exploration system together with a suite of visualization and interaction toolkits, does effectively reveal the theme evolution, theme diversities, and topic interleaving, and meanwhile allows for comparing multiple long documents.

Exploring a document or a document corpus can be multi-faceted. The proposed approach provides a new means in the viewpoint of the topological thematic organization. To make it more perceivable, we expect to enhance the visualization with illustrative techniques, e.g., equipping a topic with descriptive pictures. We also plan to apply our approach to text streams (e.g., email) to help discover useful information. It is also of great interests to investigate specific visualization methods for diverse document types.

Acknowledgements

This work is supported by National Natural Science Foundation of China (No. 61232012), Doctoral Fund of Ministry of Education of China (No.20120101110134), and National High Technology Research and Development Program of China (2012AA12090).

References

1 Šilić A, Bašić B. Visualization of Text Streams: A Survey. Proceedings of the 14th international conference on Knowledge-based and intelligent information and engineering systems: Part II, Cardiff, UK, 2010, 31-43.

2 Miller N, Wong P, Brewster M, Foote H. TOPIC ISLANDS — A Wavelet-Based Text Visualization System. In Proceedings of IEEE Visualization '98, Los Alamitos, CA, USA, 1998, 189-196.

3 Blei D, Ng A., Jordan M. Latent dirichlet allocation. The Journal of Machine Learning Research, 2003, 3:993-1022.

4 Blei D, Griffiths T, Jordan M, Tenenbaum J. Hierarchical topic models and the nested Chinese restaurant process. In Proceeding of Advances in Neural Information Processing Systems 16, Vancouver, Bristish Columbia, Canada, 2003.

5 Havre S, Hetzler B, Nowell L. ThemeRiver: Visualizing theme changes over time. In Proceeding of IEEE Symposium on Information Visualization, Boston, MA, USA, 2002, 115-123.

6 Wei F, Liu S, Song Y, Pan S, Zhou M, Qian W, Shi L,Tan L, Zhang Q. TIARA: a visual exploratory text analytic system. In Proceedings of ACM SIGKDD international conference on Knowledge discovery and data mining, Washington, DC, USA, 2010, 153-162.

7 Fisher D, Hoff A, Robertson G, Hurst M. Narratives: A visualization to track narrative events as they develop. In IEEE Symposium on Visual Analytics Science and Technology, Columbus Ohio, USA, 2008, 115-122.

8 Obstfeld R. Fiction First Aid: Instant Remedies for Novels, Stories, and Scripts. Writers Digest Books, 2001.

9 Cao N, Sun J, Lin Y, Gotz D, Liu S, Qu H. FacetAtlas: Multifaceted Visualization for Rich Text Corpora. IEEE Transactions on Visualization and Computer Graphics, 2010, 15(6): 1172-1181.

10 Lebanon G, Mao Y, Dillon J. The locally weighted bag of words framework for document representation. Journal of Machine Learning Research, 2007, 8(10): 2405-2441.

11 Deerwester S. Improving Information Retrieval with Latent Semantic Indexing. In Proceedings of Annual Meeting of the American Society for Information Science, 1988.

12 Hofmann T. Probabilistic Latent Semantic Indexing. In Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval, New York, NY, USA, 1999, 50-57.

13 Keller M, Bengio S. Theme topic mixture model: A graphical model for document representation. In Workshop on Learning Methods for Text Understanding and Mining, 2004.

14 King S. On Writing: A Memoir of the Craft. 1st edition, Scribner, 2000.

15 Segel E, Heer J. Narrative visualization: Telling stories with data. IEEE Transactions on Visualization and Computer Graphics, 2010, 16(6): 1139-1148.

16 Hassan-Montero Y, Herrero-Solana V. Improving tag-clouds as visual information retrieval interfaces. In International Conference on Multidisciplinary Information Sciences and Technologies, 2006.

17 Lee B, Riche N H, Karlson A K, Carpendale S. SparkClouds: Visualizing trends in tag clouds. IEEE Transactions on Visualization and Computer Graphics, 2010, 16(6), 1182-1189.

18 Karam G. Visualization using timelines. In Proceedings of ACM SIGSOFT international symposium on Software testing and analysis, New Orleans, Louisiana, USA, 1994, 125-137.

19 Zhang J, Song Y, Zhang C, Liu S. Evolutionary hierarchical dirichlet processes for multiple correlated time-varying corpora. In Proceedings of ACM SIGKDD international conference on Knowledge discovery and data mining, Washington, DC, USA, 2010, 1079-1088.

20 Ishikawa Y, Hasegawa M. T-Scroll: Visualizing trends in a time-series of documents for interactive user exploration. Research and Advanced Technology for Digital Libraries, 2007, 4675: 235-246.

21 Mao Y, Dillon J, Lebanon G. Sequential document visualization. IEEE transactions on visualization and computer graphics, 2007, 13(6):1208-1215.

22 Wattenberg M, Viégas F B. The word tree, an interactive visual concordance. IEEE Transactions on Visualization and Computer Graphics, 2008, 14(6):1221-1228.

23 Ham F, Wattenberg M, Viégas F B. F B Mapping text with phrase nets. IEEE Transactions on Visualization and Computer Graphics, 2009, 15(6):1169-1176.

24 Don A, Zheleva E, Gregory M, Tarkan S, Auvil L, Clement T, Shneiderman B, Plaisant C. Discovering interesting usage patterns in text collections: integrating text mining with visualization. In Proceedings of the Sixteenth ACM Conference on Conference on information and Knowledge Management, Lisboa, Portugal, 2007, 213-222.

25 Chen Y, Wang L, Dong M, Hua J. Exemplar-based Visualization of Large Document Corpus. IEEE Transactions on Visualization and Computer Graphics, 2009, 15(6):1169-1176.

26 Brontë C. Jane Eyre. Simith, Elder Co., 1847.

27 Eades P, Feng Q. Multilevel Visualization of Clustered Graphs. In Proceedings of the Symposium on Graph Drawing. Berkeley, California, USA, 1996, 101 - 112.

28 Voloshin V I. Introduction to Graph and Hypergraph Theory. Nova Science Publishers, 2009.

29 Choi F. Advances in domain independent linear text segmentation. In Proceedings of the 1st North American chapter of the Association for Computational Linguistics conference. Morgan Kaufmann Publishers Inc., Seattle, Washington, USA, 2000, 26-33.

30 Hearst M A. TextTiling: segmenting text into multi-paragraph subtopic passages. Comput. Linguist.,1997, 23:33-64.

31 Ponte J, Croft W. Text segmentation by topic. Research and Advanced Technology for Digital Libraries, 1997, 1324:113-125.

32 Sammon J W. A nonlinear mapping for data structure analysis. IEEE Transactions on Computers, 1969, C-18(5):401-409.

33 Viégas F B, Wattenberg M, Feinberg J. Participatory visualization with Wordle. IEEE Transactions on Visualization and Computer Graphics, 2009, 15(6):1137-1144.

34 Collins C, Penn G, Carpendale S. Bubble Sets: Revealing Set Relations over Existing Visualizations. IEEE Transactions on Visualization and Computer Graphics, 2009, 15(6):1177-1185.

35 The Prefuse: An Information Visualization Toolkit. http://prefuse.org/.

36 Clinton H R. Living History. New York, NY, USA, Simon & Schuster, 2003.

37 Clinton B. My Life. Random House Digital, Inc, 2005.

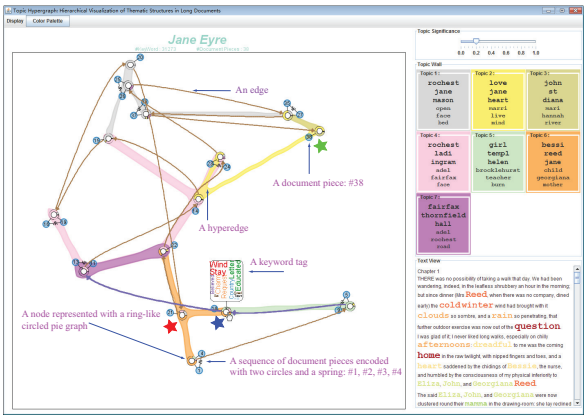


Figure 5: Representing and visualizing the thematic structure of the novel *Jane Eyre* [26] with a hierarchical topic hypergraph. The interface includes three main panels. The right panels show the topic list and the underlying document, respectively. A selected hierarchy of the topic hypergraph is shown in the left view: a colored iso-contour encodes a hyperedge (topic); a node (a theme) is represented with a ring-like circled pie graph that encodes the probability distribution of each topic in the node; a curved arrow connecting two nodes represents an edge. The document pieces contained in each node are represented as small circles in blue, and a keyword tag is used to show its keywords of the node.

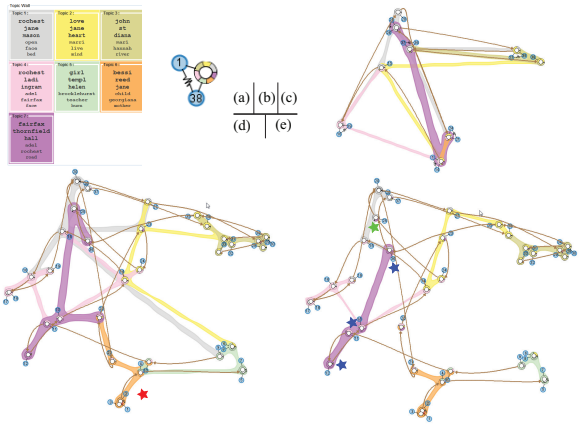


Figure 6: The topic hypergraph illustration of the novel *Jane Eyre* with different hierarchies and thresholds for the topic significance. (a) The list of seven topics; (b) Root node that includes 38 document pieces (chapters); (c) Distributions of topics in the entire document. (d) Finest hierarchy with the topic significance threshold 0.2; (e) Finest hierarchy with the topic significance threshold 0.4.

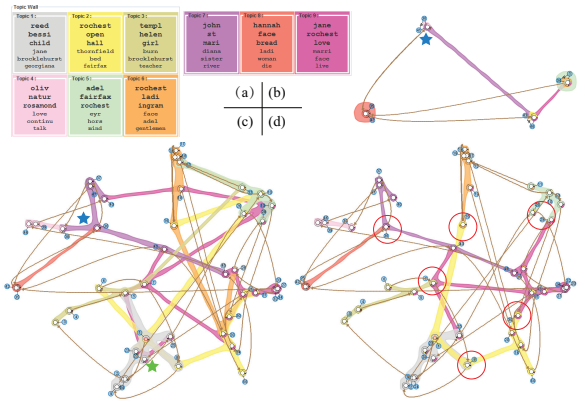


Figure 7: Generating a topic hypergraph by employing a different document segmentation scheme from the one used for Figure 6. (a) The topic wall showing 14 topics. (b-d) The visualizations with different hierarchies and thresholds.

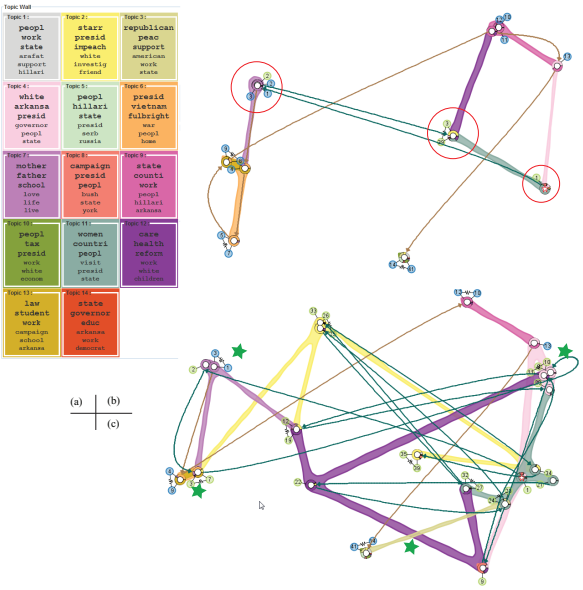


Figure 8: Thematic comparison between two books *My Life* and *Living History*.