Tapestry: Enhancing In-depth Text Analysis at Document Level

Xiaohua Sun
Tongji University
4800 Cao’An Road, Shanghai, China
xsun@tongji.edu.cn

ABSTRACT
Modeling text corpus into key topics is good for exploratory analysis. For analyses facilitating decision making, keeping information of individual documents intact is also important. In this paper, I propose the design of an interactive visual analysis interface that combines topic summarization with document information preservation. A series of example analysis activities and use scenarios show its strength in supporting decision making through in-depth analysis at document level.

Keywords
Topic document relation, decision making, interactive visual text analytics.

1. INTRODUCTION
Summarizing a large collection of documents into a set of topics or themes is a method used in many text analytics works starting from early research like ThemeScape by Wise, et al. [3], Galaxy of News by Rennison [4], Themeriver by Havre et al. [5] to recent explorations like EachMovie by Iwata, et al. [6] and TIARA by Liu, et al. [1],[2]. Besides topic or theme modeling, these systems also share another feature in common: plotting documents as dots on the visual representation (stripes in TIARA, clusters in Galaxy of New and other systems) of the topic to which they have the highest topic distribution value. This approach is good at giving an overall picture about the corpus under analysis. However, during the process of topic modeling and the process of assigning a document to only one topic, certain information about topics, documents, and topic document relations are chopped off. The topic-centered visualization also makes it difficult to carry out further analysis when systematic inspection of documents is needed.

This paper proposes an interface that both takes advantage of topic summarization in effectively dissecting a corpus and enhances in-depth analysis by preserving full-fledged information at document level. This is realized by switching the focal point from topics or themes to documents and document/topic relationships. In the exploratory analysis approach, topics or themes are the focal point with documents defining properties of them. For example, the number of documents that have high distribution value to a topic contributes to the rank of that topic and the size of it in the visualization; the keywords from those documents contribute to keywords used to label the topic. In the interface proposed in this paper, documents are the building blocks of the visualization with topic distribution as one of the properties of the documents and with topic development as patterns emerged out of the topic distribution pattern of all the documents in a collection.

The main part of this interface consists of rows of threads displaying the topic distribution values of documents to main topics modeled from the collection. It straightly shows the document/topic relationship without losing any information. On top of the distribution pattern, topics could be further refined in context to compensate the deficiency of current topic modeling technologies. The interface also includes an Interactive Timeline, a Topic Distribution Inspector and a Document Count Bar Chart. All these components work in synergy to reveal the properties of the data under analysis from different perspectives in one place. It can help users spot trends, patterns, anomalies, and provide necessary information for detailed analysis. Its simple and intuitive visualization and interaction metaphor makes it possible to serve as a workplace for conducting various types of in-depth analyses.

This paper is organized as follows. First, the design of the interface along with main functions supported by each part of it is introduced. Second, analysis activities supported and example use scenarios are presented to show the strength of this interface in facilitating decision making. Preliminary implementation of the design and next steps of this research are then stated along with the conclusion of the paper.

2. INTERFACE DESIGN
As illustrated in Figure 1, the interface is mainly composed of the Topic/Document “Tapestry”, the Interactive Timeline, the Topic Distribution Inspector, and the Document Count Bar Chart. There are also other UI components supporting searching, sorting, zooming, and topic merging/splitting. In this section, I will first introduce the metaphor from which I got the inspiration for this design. I will then describe different parts of the interface and their main functions.


2.1 The Visual Metaphor
The design of this interface is inspired by the making of tapestry, a weft-faced weaving. In this craft making, vertical threads (the warps) are set up as framing lines for the horizontal threads (the wefts) to weave back and forth. The warps are hidden in the completed work. It is the weft threads of different colors working together to form the design. Similarly, topics summarized can be considered as a framework, on top of which the topic distribution patterns of individual documents working together to express the rich document/topic relationship pattern and the temporal development of topics with all the nuances displayed.

2.2 Topic/Document “Tapestry”
This is the main part of the interface. It is composed of horizontal threads representing documents in a collection. The threads are placed from top to bottom along temporal sequence. They can represent individual document or a set of documents in a day, a week, a month, a year, or a period of time of arbitrary length. From left to right, the thread is divided into segments to display the topic distribution value of the document or document set to the top topics summarized. Saturation of color is used to encode the distribution value: the higher the distribution value to a topic, the higher the color saturation.

Users can work with segments on the document threads by topic. They can drag and change the horizontal ordering of segments corresponding to a topic across all the document threads as a group. The can also select topics to merge or split and update their corresponding thread segments across the “tapestry”. Both the number of topics selected to focus on and the screen space assigned to each topic is adjustable.

Besides distribution value, the thread segments can also display various information useful for analysis: 1) Displaying sentiment information by tainting the thread segments; 2) Displaying metadata-based querying results by highlighting the related segments; 3) Displaying keywords in a document or document set related to the topics under study in an expanded view of the thread.

The method of using interactive colored lines to represent various aspects of text materials and aid the detection of interesting patterns has been used in early research like Seesoft by Eick et al. [7] and FeatureLens by Don et al. [8]. Different from these existing works, instead of reflecting properties of documents or files in a linear manner, the colored line segments in this design work in a form similar to matrix to reveal the document/topic relationship patterns.

2.3 Interactive Timeline
The Interactive Timeline is another key component of the interface. It closely works together with the “tapestry” to support temporal exploration. It is not a timeline in the normal sense but rather a line encoding the temporal information of the document threads in the “tapestry”. Different colors are used to represent the different temporal granularity of the corresponding threads. As illustrated in Figure 1, green, brown, dark red, purple are used for the time of a day, a week, a month, and a year correspondingly. Dark blue represents a period of time of arbitrary length. Light orange indicates that the corresponding thread is an individual document no matter it is in the default view or the expanded view. Users can zoom in and out to change the point of interest on the timeline to different temporal granularity. The corresponding thread in the “tapestry” will then be set to represent document or document set belong to that length of time. Users can also reset temporal
granularity of all the threads to one of the standard time unit by clicking the reset button on top of the Interactive Timeline.

2.4 Topic Distribution Inspector
The visual hint about topic distribution displayed through color saturation in the Topic/Document “Tapestry” is good at revealing overall patterns. However, it is hard to tell from it quantitative topic distribution value of the document or document set of interest. A Topic Distribution Inspector is thus provided at the bottom of the interface to help with tasks when accurate topic distribution information is needed or when there are selections based on distribution values.

2.5 Topic Document Count Bar Chart
To the left of the Interactive Timeline is a bar chart showing the number of documents in the corresponding thread in the Topic/Document “Tapestry”. This helps to reveal the document distribution across time, which can be used as reference information for the temporal granularity adjustment. A reset button similar to the one on top of the Interactive Timeline is also provided. Clicking this button will set each thread to represent one document.

3. ANALYSIS ACTIVITIES SUPPORTED
After introducing the design of this interface, I will show in this section how the analysis activities supported by it could help with decision making.

Decision making and in-depth analysis
It is natural that before making decisions, people would like to carry out in-depth analysis on top of exploratory summarization of the corpus under inspection. The strength of this interface in keeping rich information of documents and displaying document/topic/temporal information in one place renders it an effective workplace for in-depth analysis.

Supported activities include: 1) Sorting documents or topics; 2) Searching/filtering documents using metadata or keywords; 3) Topic/document relationship detection; 4) Abnormal pattern detection; 5) Bursty pattern analysis with the aid of the Document Count Bar Chart; 6) Correlation analysis: i. Using the Topic Distribution Inspector to find topics and documents with high correlation; ii. Detecting topic correlation through shared documents; iii. Detecting document correlation through shared topics.

Decision making and interesting documents detection
The process of decision making sometimes involves with narrowing down the documents under inspection to a smaller and smaller set that are particularly relevant to the task under research.

Some use scenarios are: 1) Through selection on the Topic Distribution Inspector one could focus on documents with certain topic distribution values; 2) Selecting high-rise bars in the Document Count Bar Chart can bring to users the corresponding documents that contribute to the bursty pattern across time; 3) Contracting time period with few documents via the Interactive Timeline can help bring to focus contents spread out across time for easy scrutinization; 4) After discerning from the Topic/Document “Tapestry” documents with abnormal topic distribution pattern, further actions of including more topics or sorting topics by topic distribution value of this set of documents may bring to sight topics with low overall ranking but high correlation with the abnormal documents.

Decision making and hypothesis formulation
Formulating certain hypothesis and have it proved or rejected through further analysis is another way to bring people closer to decision making.

With this interface, users can move topics around or play with the Interactive Timeline to look for meaningful document/topic relationship evolvement patterns. Hypothesis about certain phenomena could then be formulated on top of this to guide further analysis. Some examples of such patterns are: 1) One topic starts after or around the ending of another, 2) Two or more topics share documents in common; 3) Documents contributing to some topics are periodical, whereas those for others are evenly spread out. In the second example, having common documents at the beginning may suggest that those topics are split out of one topic. Sharing documents towards the end suggests a merge into one topic. Having common documents at various times across the whole process indicates that those topics overlap with each other.

Decision making and topic refinement
As stated at the beginning of this paper, topic summarization is an effective way of digesting large collection of documents and has been used in many text analytics research. However, topics detected using different models are often based on low-level statistic values, which may not make sense under specific conditions. Human knowledge is often needed to refine the topic in context. This is especially important for analyses supporting decision making. Various means are provided by this interface for the fine-tuning of topics and their ordering. Some refinement examples are: 1) Putting small “beans” of similar categories into the same bucket, e.g. one may want to combine Google news and Sohu news if news is the category of interest; 2) Merging highly correlated topics (topics share a large amount of documents); 3) Merging topics talking about similar content (telling from their keywords); 4) Merging topics with similar or correlated temporal patterns, e.g. documents for one topic are mostly around July and August every year, documents for another topic appear around September and October. After inspection, the user may find it appropriate to merge them into one topic and consider it as one event which happens throughout those four months; 5) Splitting a topic when there is no meaningful pattern formed over time and when some pattern making sense emerges after topic partitioning.

4. PRELIMINARY IMPLEMENTATION
The main component of this interface - the Topic/Document “Tapestry” has been implemented and tested over NHAMCS data (National Hospital Ambulatory Medical Care Survey) with Latent Dirichlet Allocation (LDA) as the
model used for topic summarization (See Figure 2). During the process of experimenting different abstraction levels for the patterns revealed, alternative ways of setting the width of the threads proportional to the number of documents in a uniform period of time (Figure 2 (c) by day, Figure 2 (d) by week) are tested as well. Lines are used to mark the threads with non-uniform width. It is interesting to see in Figure 2, for the same set of documents and topics, (a) preserves the highest amount of details but shows the least visible pattern, whereas (d) reveals a very clear overall pattern but has many individual document information evened out.

5. FUTURE WORK AND CONCLUSION
In this paper, I have presented the design of an interactive visual analysis interface facilitating decision making by incorporating in-depth analysis at document level. A preliminary implementation and test of the interface over NHAMCS data shows certain promising strengths of this interface in revealing interesting topic/document relationship pattern and their evolution along time. For next step, I would like to develop and implement all the interface components and functions, test the use scenarios stated above, and run user studies over specific decision making tasks.

Fig.2: Test of the Topic/Document “Tapestry” over NHAMCS data. Threads in (a) represent individual document; threads in (b) represent documents in a day. Width of threads in (c) and (d) are proportional to the number of documents in a day and in a week correspondingly.

REFERENCES