

Understanding Text Corpora with Multiple Facets

Lei Shi Furu Wei Shixia Liu Li Tan Xiaoxiao Lian*

IBM Research - China
19 Zhongguancun Software Park
Beijing 100193, China

Michelle X. Zhou†

IBM Research - Almaden
650 Harry Road
San Jose, CA 95120, USA

ABSTRACT

Text visualization becomes an increasingly more important research topic as the need to understand massive-scale textual information is proven to be imperative for many people and businesses. However, it is still very challenging to design effective visual metaphors to represent large corpora of text due to the unstructured and high-dimensional nature of text. In this paper, we propose a data model that can be used to represent most of the text corpora. Such a data model contains four basic types of facets: time, category, content (unstructured), and structured facet. To understand the corpus with such a data model, we develop a hybrid visualization by combining the trend graph with tag-clouds. We encode the four types of data facets with four separate visual dimensions. To help people discover evolutionary and correlation patterns, we also develop several visual interaction methods that allow people to interactively analyze text by one or more facets. Finally, we present two case studies to demonstrate the effectiveness of our solution in support of multi-faceted visual analysis of text corpora.

Keywords: text visualization, multi-facet data visualization.

1 INTRODUCTION

Recently, there has been a great interest in analyzing complex text documents, each of which contains multiple data fields [17]. Although existing text mining techniques have been successful [8][13][4][17], average users cannot easily consume the analytic results, let alone leveraging them in their decision making processes.

In real-world applications, human analysts are often required to examine evolutionary and correlation patterns among multiple facets across a text corpus. For example, the patient records of an emergency room contain three free-text fields: cause of injury, reason for visit, and diagnosis, along with multiple value (structured) fields, such as patient gender and age. An interne may want to learn the co-occurrences between the “reason for visit” (symptom) and the final “diagnosis” (between two unstructured facets); a health-care insurance analyst on the other hand may want to find out the correlations between the patient’s “cause of injury” and his/her “age group” (between two unstructured and structured facets) to better assess insurance terms. Moreover, a government officer may need to study the seasonal patterns between “diagnosis” and the “date of visit” (between unstructured and time facets) for disease control purpose.

However, it is nontrivial to build a visual analytics tool that supports all the above analyses especially in aiding users in their pattern-findings within the multi-facet text corpora: First, it is challenging to combine the raw text corpus and its various text analytics results for effective visualization; Second, we cannot find existing

visual metaphors to effectively illustrate the multi-faceted textual data and their evolutionary or correlation patterns; Third, rich interactions are also needed to help users in their explorative, iterative visual analysis.

Existing text visualization either focuses on revealing the patterns discovered from the extracted entities [16][9][20][28][25], or aiming at visually summarizing the unstructured text content [24][21][27][6]. Few works [11][7] combine the extracted data facets with the unstructured text content in support of multi-faceted analysis. To the best of our knowledge, our work is the first that tightly integrates interactive visualization with a multi-faceted data model of the text corpora for effective visual representation, navigation, and analytics.

In this paper, we first define our general multi-faceted data model for text corpora, classifying its data facets into four categories: the time facet, the category facet, the text content (unstructured) facet and the associated structured facet. We discuss the extraction and synthesis of data facets from a raw text corpus, including topic extraction, sentiment analysis, and the time-sensitive keyword summarization. We then introduce a visualization design based on TIARA [15] that combines a trend graph with tag clouds. Each of the four data facet is mapped to one visual dimension respectively. We further present our interaction methods that offers users great flexibility to manipulate and customize their visual analysis processes by their own needs.

We demonstrate the success of our solution in two real-world case studies. The first is with the US health care survey data. Through guided navigation of the data, we find both seasonal patterns hidden in the emergency room records and correlations between structured facets (e.g., “patient sex”) and unstructured text content (e.g., “cause of injury”). The second case study is to analyze hotel customer reviews. By juxtaposing the entity keywords extracted from the hotel reviews (unstructured facet) over the time line and further brushing them by overall ratings (structured facet), users can obtain a quick comparative overview across the hotels. After a candidate hotel is selected, users could also drill down to its categorized sentiment reviews as a reference for confirmation.

The remainder of this paper is organized as follows. Section 2 summarizes related work. Section 3 introduces our multi-faceted data model for text corpora and the techniques applied to extract the facets. Section 4 shows our visualization design. We describe two case studies in Section 5 and conclude the paper in Section 6.

2 RELATED WORK

Generally speaking, our work consumes the *text analytics* results with *interactive visualization*, hence is related to the both areas.

Text analytics has been positioned as the major techniques to mine and model text corpus. Latent Semantic Indexing (LSI) [8] summarizes the corpus by modeling both the documents and keywords as vectors in a latent linear subspace. Recently, the generative probabilistic models extended from LSI, such as pLSI [13] and Latent Dirichlet Allocation (LDA) [4], become popular due to their success in explicitly modeling a document as a mixture of topics and each word as an instance of the topic appeared stochastically. Motivated by the success of state-of-the-art text analytics technolo-

*e-mail: {shllsh,weifuru,ltan,xxlian}@cn.ibm.com, shixia@gmail.com

†e-mail: mzhou@us.ibm.com. This work was performed when Michelle X. Zhou was in IBM Research - China.

gies, our solution is to leverage both the text analytics and visualization to help detect and validate the patterns and correlations within a time-evolving text corpus. Essentially our visualization is designed to illustrate the analytical results generated by LDA. We also leverage Natural Language Processing (NLP) techniques such as named entity recognition, Part-of-Speech (POS) tagging, and sentiment analysis to extract data facets in order to better illustrate the text corpus.

Meanwhile, various text visualization techniques have also been proposed in the recent years. They can be classified into two categories according to their motivation. The first kind aims at visualizing the patterns, features and relationships of the entities in a corpus. The works done by Wong et al. [29][28] visualize the association rules and sequential patterns mined from the text. TileBars [12] shows the distribution of terms or queries in search results. Takmi [16] and FeatureLens [9] systems discover frequent entities in the text content and provide multiple coordinated views to illustrate the distribution and evolutionary patterns attached with the entities. Jigsaw [20] system composes visualizations to illustrate the connections between the extracted entities to facilitate the investigative analysis. Recently, WordTree [27] augments the traditional keyword in context visualization with a tree-based graph for visual query and navigation.

The second category of text visualization techniques focus on visually summarizing the content of a text corpus. Tag cloud [23] and its artistic version, Wordle [24], illustrate raw text using compact word cloud where font size encodes the importance of a word (e.g. its TF-IDF score). These techniques only depict limited aspects of a text corpus, they are mostly used for overview purpose instead of interactive analysis. To study syntactic or lexical relationships, other techniques then perhaps are more effective, including: PhraseNet [21], WordTree [27] and DocuBurst [6]. However, due to the complexity of the internal document structure at the sentence and paragraph level, the existing techniques are limited to show only certain aspects of a text corpus without providing users a comprehensive understanding of the corpus..

Our proposed solution differs from the existing techniques since we leverage the extracted facets from a text corpus to help illustrate its entire content. Several pieces of work have attempted to achieve the similar goal. ThemeRiver [11] and its derivatives [26][5][10] use a river metaphor to depict the thematic variation over time within a text corpus. It visualizes the combination of the time and category facets. The Parallel Tag Cloud [7] divides a text corpus by a selected structured data facet and then places the keywords extracted in each facet value category as a vertical tag cloud. The separate tag clouds are further glued together in a parallel coordinate. It essentially leverages the extra structured facet to facilitate the content comparison. While our design is inspired by these works, we augment the scope by mapping all of the four types of data facet to the visualization. Furthermore, we support user interactions to customize the mapping of each facet to the visual dimension to allow flexible navigation, which in turn greatly facilitates text analytics.

There are also several other works targeting visually analyzing the text corpus in specific application domains, such as Theme-mail [22] for email analysis, NewsLab [10] for video news analysis, and TileBars [12] for examining search results. In contrast, we aim at building a general framework to visually analyze text corpus of multiple application domains.

3 FACETED DATA MODEL FOR VISUAL TEXT ANALYTICS

We aim to provide a general data visualization framework for text corpus over visual analytics tasks where 1) the content evolution over time is captured; 2) the correlations among different facets are revealed. The data model for text corpora with multiple facets is defined as: for each single document within the corpus, we decouple the information into four types of data facet, namely the time

facet, the category facet, the content facet and the adjunct structured facet. These facets could be either inherent as meta-data or automatically/manually tagged as labels.

Formally, the text corpus is defined as $\mathcal{D} = \{\Gamma, F_u, F_s, T\}$, where Γ denotes the category facet derived from the topic modeling, clustering, classification algorithms, or even directly categorized by pre-defined labels. F_u and F_s denote the unstructured (content) and structured facets respectively. F_u generally includes the raw text fields in each document, such as “*cause of injury*” in the patient records. It could be further summarized by natural language processing techniques. F_s includes the numerical or nominal data fields in each document used to better illustrate the unstructured text content, such as the “*patient sex*” of a record. F_s can also be derived from analytical results, e.g. the sentiment orientation of hotel reviews in our second case study. Finally T denotes the time stamp of each document, for most cases this is available in the meta-data.

In the following parts, we present the technical details for mining and extracting the data facets defined above as well as describe the data interface implementation between the back-end data processing component and the front-end visualization widget.

3.1 Category Facet Mining

Some text collections have built-in category information such as the hotel name in the TripAdvisor case. For the majority of others without an inherent and reasonable classification, we employ the topic based content categorizing method. The text content of the corpus is organized as a collection of topics, each represented by a set of keywords (up to 50). The topic categories then form the category facet of the corpus.

In detail, category facets can be obtained in many ways. Traditionally by document classification, it requires a large number of annotated training samples which are costly to obtain. As a result, it is often desirable to use unsupervised learning methods which automatically discover hidden themes in the document collection. Clustering [14] and latent topic models are a few such methods.

In this work, Latent Dirichlet Allocation [4] is used to extract topics from a document collection. A unified topic model is trained on the integrated content by combining the multiple text fields within each document together. Given a document collection $D = \{d_1, d_2, \dots, d_N\}$ where N denotes the document number, each document d_i is assigned with a distribution over K topics learned from the document collection where K denotes the pre-defined topic number. Meanwhile, each topic t_j is also assigned a distribution over the word vocabulary $W = \{w_1, w_2, \dots, w_M\}$ collected from the whole text corpus, where M denotes the size of the vocabulary.

3.2 Unstructured/Content Facet Extraction

The content of a document usually consists of multiple text fields which are directly mapped to the unstructured facets in our data model. In other cases where there is no such text field, we leverage natural language processing techniques to divide the text content into separate groups (a.k.a. fields). For example, we can divide the text content into separate unstructured facets according to POS. We can also extract named entities (such as location, organization, people, etc.) and action words (i.e. verbs) to establish new content facets. In our case study over hotel review corpus, we classify the content keywords into two groups: the sentiment keywords and the entity keywords (see Table 3), thus two unstructured facets, “sentiment” and “entity”, are extracted.

The content of each text field, denoted as f_i , is summarized by several lists of weighted keywords, each for one topic category. These lists are generated in two steps. In the first step, each keyword appearance in the f_i field of the document collection is tagged with a topic label by LDA. Then in the second step, for each topic category of f_i field, all the keywords belonging to it are ranked by TF-IDF [18], where a weight is assigned for each keyword.

Table 1: Definition of *topic* Objects

Attribute	Description
<i>keywordnames</i>	a set of keywords describing this object
<i>keywordweights</i>	keyword weights of <i>keywordnames</i>
<i>timepoints</i>	a set of time points
<i>topicheights</i>	the topic height (i.e. number of documents) at <i>timepoints</i>

Table 2: Definition of *timesensitivekeyword* Objects

Attribute	Description
<i>starttime</i>	the start time of this object
<i>endtime</i>	the end time of this object
<i>keywordnames</i>	a set of keywords describing this object
<i>keywordweights</i>	the keyword weights of <i>keywordnames</i>
<i>keywordfacets</i>	the categories of the keyword in <i>keywordnames</i>

Further, to reveal the evolutionary patterns of the content facets, the topic keyword summarization is generated in a time-sensitive way. Given a text corpus made up of a document set tagged with time, we divide the documents into sub-collections by pre-defined time frames, and the time-sensitive keywords are extracted from each sub-collection to synthesize as a temporal summarization.

3.3 Structured Facet Mining

The structured facets are often embedded in the text documents, such as the “patient sex” in the emergency room records and “overall ratings” in the hotel reviews. Meanwhile, we are also interested in extracting other meaningful structured facets. In the trip-advisor case study, we develop text analytics techniques to infer the sentiment orientation of each review. They are divided into three categorical values, namely the “Positive”, “Negative” and “Neutral” sentiment orientations. Also, we manually tag a dictionary of entity keywords classified into four feature types (see Table 3). The feature classification behaves as another structured facet.

3.4 Data Interface

We have designed a unified data interface between the analytics engine and the visualization component in JSON standard [2]. The analysis results are organized in terms of a set of *topic* objects defined in Table 1. For each content or structured facet, the *topic* object is further associated with a set of *timesensitivekeyword* objects defined in Table 2.

4 VISUALIZATION

4.1 Overview

The visualization proposed in this paper builds on TIARA [15]. A typical view is given in Figure 1. The main 2D graph in the right depicts the visual representation of the time-evolving text corpus. Essentially all the four facets extracted from the data are mapped to the visual dimensions in the representation: 1) The X axis is mapped to the time facet. 2) The stacks along the Y axis are mapped to the category (a.k.a. topic in this section) facet with Y axis values of each stack indicating the number of documents in each topic. Hence each stacked strip in the graph corresponds to one topic trend in the data. 3) The keywords on each stacked strip are summarized from the content facet belonging to the corresponding topic, in a time-sensitive manner to reveal the content evolution pattern over time. The size of each keyword is mapped to the re-ranked keyword weight. The more frequent a keyword appears in the corresponding time frame compared to other frames, the larger the keyword will be drawn. 4) The visual dimensions of the keyword other than size,

such as keyword color and font, are mapped to the adjunct structured facet within the text corpus. In the case of Figure 5(a), the keyword color indicates the tendency for the keyword to appear in male or female records.

Apart from the trend visualization, there also lies a floating legend listing the label of topics shown in the graph. This legend is linked to the topic trend through synchronized highlighting effects. To the left of the visualization, there is another panel listing the unstructured and structured facets eligible for navigation. Other visual components such as the interaction panel and search box are also available as shown in Figure 4(a).

This visualization supports several basic interactions: trend highlight upon mouse-hover (Figure 4(a)); trend expansion for more content keywords upon mouse-click (Figure 4(b)); show trend height through the toggled button on top of the trend visualization (Figure 4(a)); drill in to the associated text snippets upon clicks over one content keyword (Figure 6(c)); reset to the initial view.

4.2 Data Navigation Interactions

Apart from the basic interactions, we further highlight the four advanced data navigation techniques used for manipulating the four types of data facet. They are designed to facilitate the user to analyze the evolution and correlation among the data facets of text corpus.

Temporal Zooming: In our visualization, the top and bottom contour of each topic trend are drawn in Bezier curve with fixed-interval control points. To smooth the topic contour, the number of control points selected can not be sufficiently large, in most case, about 20 points for each trend. Therefore the fine-grained evolution pattern can not be traced in the default view. To compensate for this, we introduce the temporal zooming interaction. As in Figure 2, the user drags the side bars to select a new time range and clicks to zoom. Then the topic trends are animated to the view with the new time range. Both the trend curves and the time-sensitive keyword summarizations are updated to the finer granularity so that more details are provided.

Topic Editing: Generally the number of topic trends shown in one view is capped for the ease of perception, which could be much smaller than the number of topics trained from the text corpus. The topic editing interaction is then incorporated to allow the user to customize the list of topic trends in the visualization. The user could access this edit mode by clicking the button in the interactive legend. After that, the full topic list is unfolded for selection, as given in Figure 2. Upon the submission, the topic trends are animated to the new layout.

Unstructured Facet Navigation: By default, the text content shown in this visualization includes all the unstructured data facets in the corpus. To help the user focus on one single facet or compare multiple ones, we introduce an extra facet navigation panel to the left of the visualization. Upon changes to the selected unstructured facets, the keywords on the topic trend are updated to those summarizing the selected facets. The keyword color is mapped to the unstructured facet the keyword comes from. As in Figure 4(a), the “Diagnosis” keywords are drawn in blue and the “Reason for Visit” keywords are drawn in green. A color indicator is inserted into the legend for visual guidance.

Structured Facet Mapping: To analyze the correlation of structured and unstructured facet, we allow the user to customize the mapping between structured facet value and the visual dimensions of content keyword. This is also achieved by selecting the mapped structured facet in the facet navigation panel. The configurable visual dimensions of the keyword include the keyword color hue, transparency and font type. Figure 5(a) shows an example of mapping the “patient Sex” value into the keyword color.

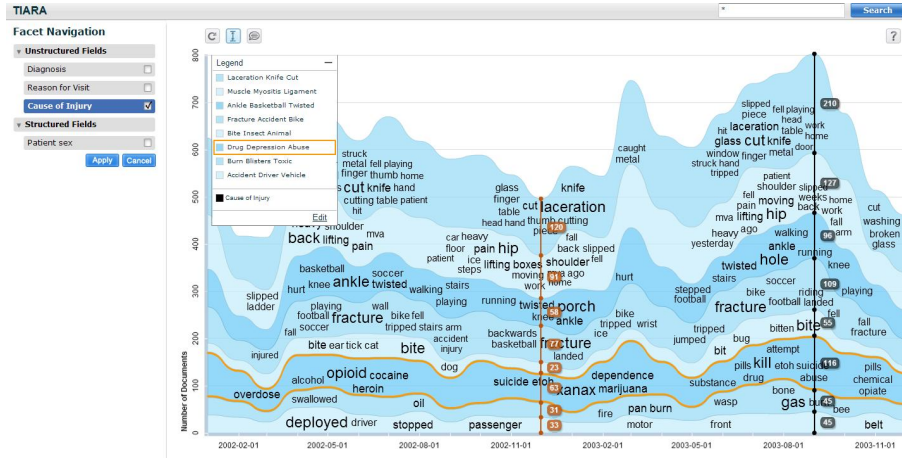


Figure 1: Visualization of text corpus with multiple facets.

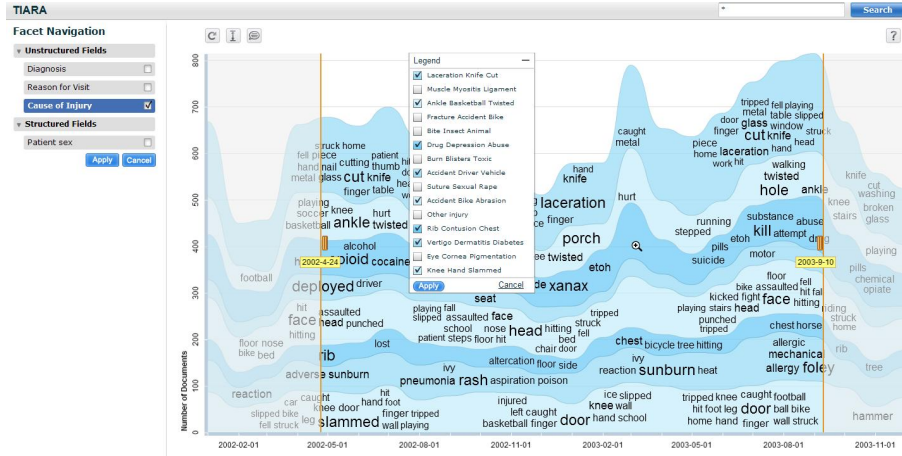


Figure 2: Data navigation interactions.

4.3 Fisheye Distortion and Topic Keyword Layout

More technical details related to the interaction and keyword layout are talked about in this part.

Fisheye Distortion:

Upon the mouse click, the topic trends in this visualization will expand to show more keywords or correlations if multiple unstructured facets are selected, as shown in Figure 4(b). There are three basic requirements for such trend distortion, from the most important one to the least:

- (i): The contour of the whole graph should remain intact so that the overall trend pattern is preserved.
- (ii): The number of time-sensitive keywords shown in each time frame of the selected topic trend should be sufficient for details purpose after the trend expansion.
- (iii): The temporal shape of the selected topic trend should be kept as undistorted as possible.

To achieve that, we design the algorithm based on the standard 1D fisheye distortion [19]. The advantage of the fisheye solution includes the undisturbed overall contour, and also the distortion happens in a natural way, the largest at the selected foci and lower as the distance to foci increases. However, the classical 1D fisheye adopts the uniform expansion parameter along the axis, so that the trend segment with small heights can not be expanded too much, due to the failure of expansion of the trend segment with the larger height. The non-uniform fisheye distortion could resolve this issue, but in return, will break the relative height order of the selected topic and further destroy the visual momentum of the trend pattern. We developed a three-step algorithm to meet the requirements

raised triply.

(I): In the first step, the desired height after expansion for the selected topic trend is estimated. We currently support two methods to calculate that. The first one is to expand the selected trend to a upper-bound ratio of the overall trend display, say 50%. The other way is to expand the trend so that the trend segment in each time frame can at least places a pre-defined number of keywords, say 20 in each segment.

(II): In the second step, the desired height is adjusted to meet shape preserving requirement for the selected topic. The original height of all the segments in the selected topic are sorted in ascending order and we scan the segments in this order. If the desired height of the current segment is smaller than the adjusted height of the previous one, we adjust the current one's height to at least the value larger than the previous one. The delta height could be a pre-defined ratio of the delta height of the original topic segments.

(III): In the last step, the distortion of each topic trend is calculated by 1D fisheye algorithm. The center line of the selected topic trend is chosen as the baseline from which the trend is expanded both upward and downward to reach the adjusted height.

For each specific segment, the fisheye distortion factor d is calculated by solving

$$\frac{(d+1) \cdot \frac{h}{2H_i}}{d \cdot \frac{h}{2H_i} + 1} \cdot H_i + \frac{(d+1) \cdot \frac{h}{2H_b}}{d \cdot \frac{h}{2H_b} + 1} \cdot H_b = h^* \quad (1)$$

where h and h^* denote the original and desired height (after expansion) of the topic trend, H_i and H_b denote the height of the overall

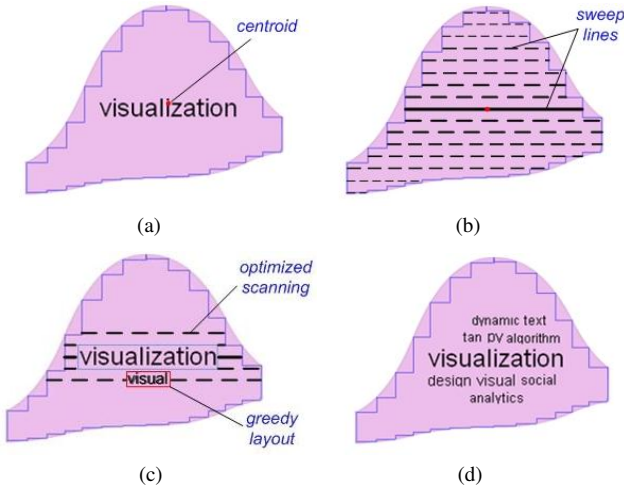


Figure 3: Sweepline layout algorithm over one trend segment: (a) Rectilinear polygon contour and its centroid; (b) Sweep lines; (c) Incremental scanning for the second keyword's layout; (d) Final layouts.

topic trend above and below the center line.

Then the final height of the i th topic trend (except the central one) in this segment is computed by

$$h_i^* = \begin{cases} \frac{(d+1) \cdot \frac{t_i-c}{H_t}}{d \cdot \frac{t_i-c}{H_t} + 1} \cdot H_t - \frac{(d+1) \cdot \frac{b_i-c}{H_t}}{d \cdot \frac{b_i-c}{H_t} + 1} \cdot H_t & (\text{above center}) \\ \frac{(d+1) \cdot \frac{c-t_i}{H_b}}{d \cdot \frac{c-t_i}{H_b} + 1} \cdot H_b - \frac{(d+1) \cdot \frac{c-b_i}{H_b}}{d \cdot \frac{c-b_i}{H_b} + 1} \cdot H_b & (\text{else}) \end{cases} \quad (2)$$

where c denotes the Y axis coordinate of the center of selected topic trend, t_i and b_i denote the Y axis coordinates of the top and bottom brim of the i th topic trend.

Keyword Layout Algorithm:

Here we describe our layout algorithm to place keywords on the topic trend. For each topic trend in one segmented time frame, e.g. the filled area in Figure 3(a), a bag of time-sensitive keywords are pre-computed for their layout within it. Our algorithm works iteratively in a greedy manner:

(I): The rectilinear polygon contour is computed from the curved contour of this segment, as shown in Figure 3(a). The centroid of the segment area is further located, where in most cases we will place the first keyword here. (Exceptions happen when it is not wide enough to accommodate the first keyword and the new location is found in the next step.)

(II): We find the layout of the next keyword by sweep line algorithm. The sweep lines are fixed-interval horizontal lines over the trend segment, as shown in Figure 3(b). We start to scan each line from the one across the centroid of the area in the order from the nearer one to the farther one. We locate the first line in which there is a gap to place the current keyword and select one feasible location on the line nearest to the centroid as candidate. Then we continue to scan the lines until the distance of the next line from the central line is already larger than the distance from the candidate to the centroid, as shown in Figure 3(c). During the process, if there is a better location with shorter distance to the centroid, we update it to the candidate.

(III): After we find the layout for the current keyword (Figure 3(c)), we update the contour of the keyword cloud already fixed, still in the form of rectilinear polygon. Then we go back to Step ii to handle the next keyword until all the keywords are assigned the layout.

Figure 3(d) gives the final layout results with our algorithm.

5 CASE STUDIES

5.1 Visual Analysis of Patient Records

Our first case study focuses on the healthcare domain. The dataset analyzed is the National Hospital Ambulatory Medical Care Survey (NHAMCS) from U.S. CDC [1], which archives the sampled patient cases in ambulatory care services provided by hospital emergency and outpatient departments across the America. In our experiment, we choose to use the patient records during the year of 2002 to 2003, about 20,000 records in total. For each record, there are three text fields ("Diagnosis", "Reason for Visit" and "Cause of Injury") and one typical categorical information ("Patient Sex"). A concrete usage scenario is described below.

Consider Alice, a government officer in disease control and prevention department, who is investigating the major causes and diagnosis for residential illnesses countrywide. After being trained to use our tool, she starts her analysis by selecting the "Diagnosis" and "Reason for Visit" fields and then use our tool to summarize the selected data. Upon Alice's request, a standard topic trend view is composed (Figure 4(a)). She quickly discovers that cutting, twisting and drug abuse are the dominant causal factors within the selected data corpus. Both temporal and content patterns within the topic trends are traceable: e.g. the number of cases falls in each winter and rebounds from the spring; there are correlations like {"open", "finger", "cut"} and {"bone", "fracture"} between the fields of some topics.

After that, Alice decides to drill down for more details on the topic she is most interested in. She proceeds by clicking the edit button on the interactive legend panel and selects the one indicating "vertigo" illnesses from the topic list. Through the zooming interaction on the timeline, she also selects a time range of Feb. 2002 to Jan. 2003 to receive only the content within a whole year. She then clicks the "vertigo" topic trend for details. Our tool herein provides an expanded view of this topic and further splits the content attached to the two selected fields into separate sub-trends. This correlation view is given in Figure 4(b), the content within the expanded topic is divided horizontally into four segments, each corresponds to one season of the year. Alice discovers that while the "vertigo" symptom is present throughout the year, however, the major "diagnosis" differs. In Spring, the patients suffered from adverse effect of drugs as well as some common diseases like diabetes and essential hypertension. While in the Summer, heat exhaustion turns out to be dominating. Further in the Winter, the same symptom may be ascribed to complications of common illnesses. The patterns in the Autumn are quite interesting, where more urticaria and fever are found to be the causes and symptoms.

Alice also plans to analyze the correlation of patient profiles (structured) with the text fields (unstructured). This time, she starts by selecting "Cause of Injury" and "Patient Sex" fields. Our tool then returns the topic trend as in Figure 5(a). The keywords associated more with male patients are shown in dark green color while the other keywords appearing more in female patients are displayed in brown. She realizes that the topmost "cut" injury happens mostly to men as nearly all the keywords there are in dark green. Alice continues looking for details by clicking on a topic trend labeled as "drug abuse". This trend is subsequently expanded and split into two sub-trends, the top one for male and the bottom one for female, as shown in Figure 5(b). Alice finds that the cause of injury for men mainly refers to hard drugs, such as cocaine, opioid, cannabis, as well as alcohol. While for women, drugs for therapy, such as xanax, aspirin and tylenol are the principle causes. It's also interesting to notice that the suicidal related keywords such as "suicide", "kill" and "attempt" are almost exclusively associated with women. Similar patterns are found when Alice turns to the "Ankle Twisted" topic, as shown in Figure 5(c). Men tend to twist their ankle during sports-related activities including basketball, football and soccer. Compared with those, women generally get their ankles hurt dur-

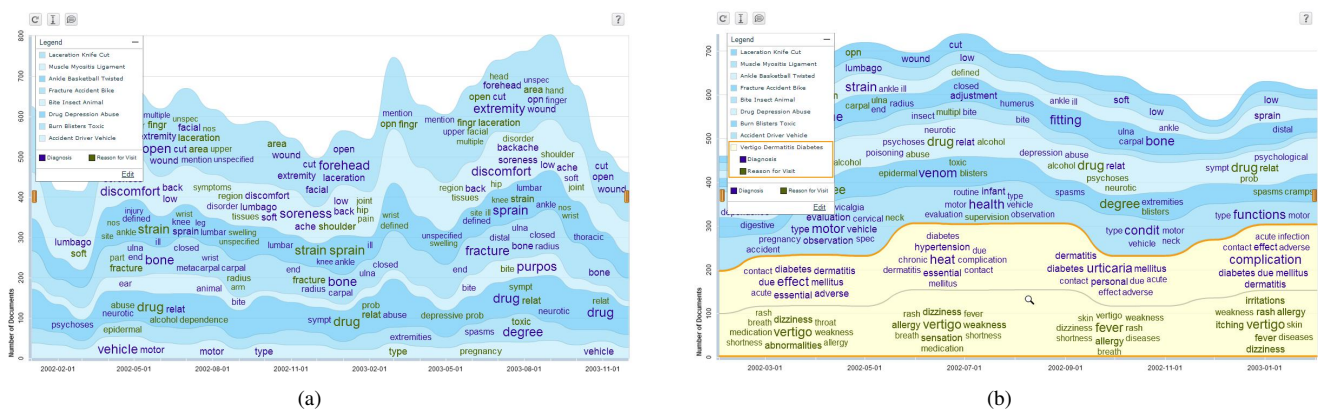


Figure 4: Unstructured data analysis for emergency room record: (a) Topic overview with “Diagnosis” and “Reason for Visit” fields selected; (b) Content correlations between the two fields within a whole year.

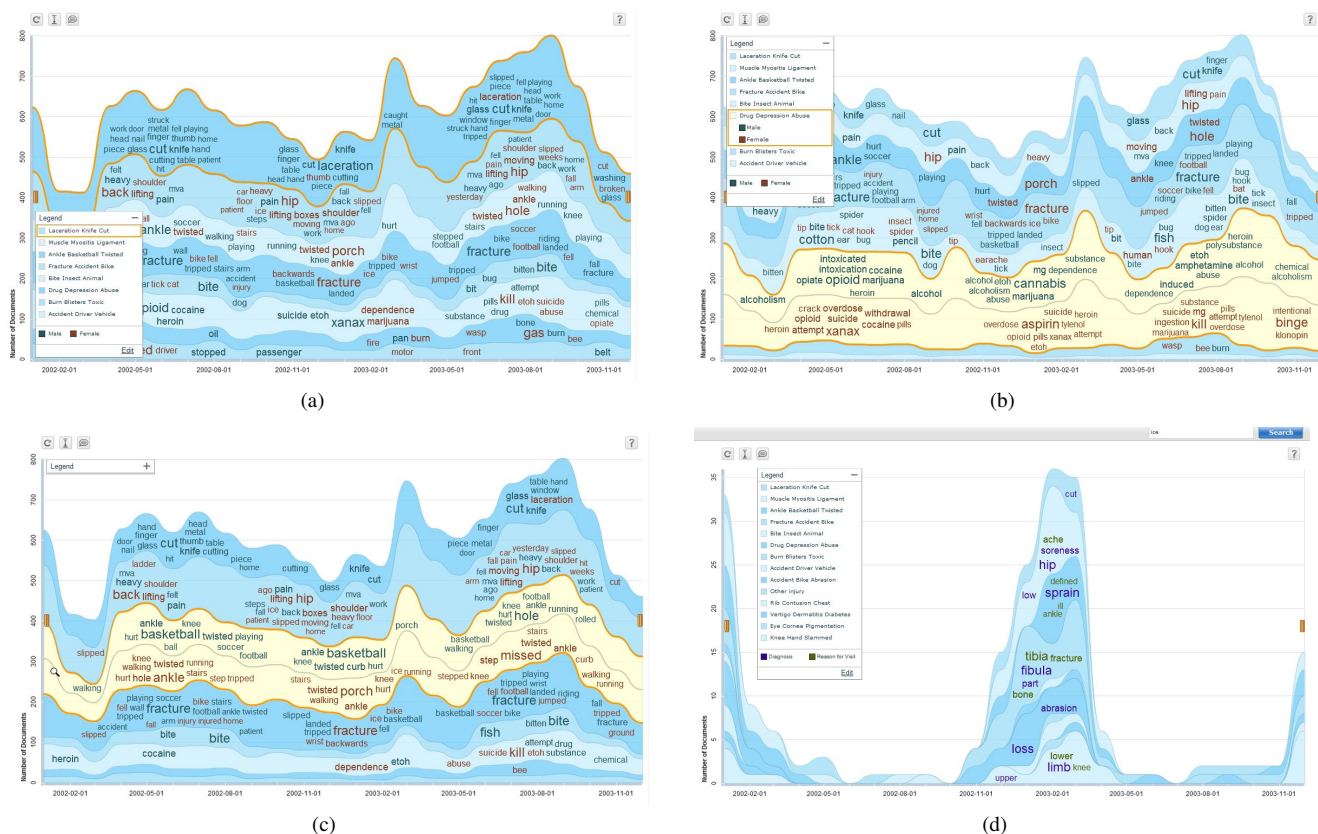


Figure 5: Structured data analysis for emergency room record: (a) “Patient Sex” distribution over “Cause of Injury” field; (b) “Patient Sex” correlations in “Drug Abuse” topic; (c) “Patient Sex” correlations in “Ankle Twisted” topic; (d) Topic view upon keyword “ice” searched.

ing walking at home (e.g., on the porch or stairs).

As Alice carries our her study in the snowy Winter, she is also interested in the injuries related to “ice”. She then types the word “ice” into the search box and submit. Our tool promptly returns the topic trends containing the word “ice”, as shown in Figure 5(d). The results confirm her hypothesis that the ice-related injuries mostly happen during the Winter time and seldom occur in other seasons.

5.2 Visual Analysis of Hotel Reviews

The second scenario is in the travel domain, where we apply our tool to the analysis of online hotel reviews in [3]. As the variance of the number of reviews for each hotel affects the analytical results, in this case study, we only select the top 10 hotels in Hong Kong that have received the largest number of reviews during the year of 2006 to 2009. The average review length for these hotels

ranges from 110 to 160 words. After the pre-processing stage, each review is formatted as a form consisting of a customer visit profile, including the date of visit and travel purpose, textual review comments and quantitative customer ratings of the hotel (between 1 to 5). Different from the medical care survey case, in this analysis we employ the hotel name as the category/topic facet for comparison, and partition the text content into two facets for navigation. In detail, we extract two types of keywords from the review content: the *entities* commented in the review (such as “floor” and “bed”) and the *sentiments* customer feels about the entity (such as “fantastic” and “painful”). A dictionary for the entity keywords is given in Table 3. They are further classified into four types of features, namely “room”, “cleanliness”, “service” and “location”. Similarly, the sentiment keyword dictionary contains 906 words in total (not listed due to space limitation) and is partitioned into three types accord-

ing to the sentiment orientation, namely “positive”, “neutral” and “negative”. The feature classification, sentiment orientation, along with the existing customer ratings in the review, work together as structured facets to help users understand the text content. A typical use case is described below.

Consider Cathy, a lady traveling to Japan on her vacation, stopped by Hong Kong with a 1-day stay. She plans to do some shopping there, but has not yet decided which hotel to stay since she is not familiar with the city. She uses our tool to help make up her mind. Cathy starts by comparing the online reviews of 7 popular hotels in Kowloon. She is interested in the key features of each hotel and we also assume that she trusts the previous customer’s assessment. She selects the “Entities” and “Overall Ratings” facets to get an overview, as shown in Figure 6(a). She finds in the graph that some hotels have received bad reviews. These are identified by red or yellow keywords indicating the comments with overall rating below average. After navigating into the details of associated comments (clicking on the keywords), she is more concerned about the story of inefficient employees, smelly rooms, grubby marks in carpet, and even bed bug bites. Finally she locates the Langham Place Hong Kong (topmost in the graph) as the top one. Although there are some negative comments on the slow elevators in early 2007, it has not been reported for the last two years. Also, during the past two years, this hotel has received more than twice as many reviews as for any other hotels. It also receives a top overall ratings (dark green keywords). More importantly, the keyword “market” shows up to depict the key characteristic of this hotel, compared with the other hotels featuring “harbourview” and “decorations”. After reading the comments related to “market”, she finds it the perfect hotel she would like to stay with, due to its most convenient access to shopping centers and businesses. After all, she comes for shopping, not for sight-seeing or relaxing.

To confirm her decision, Cathy drills down further to retrieve more information about the customer’s sentiment for each feature category of this hotel. As Figure 6(b) is shown, she is satisfied by the reviews praising the “valuable” location, “efficient” and “gladly helping” staff, “neatly” organized facility, and the “clever” use of room space. One fault if there is any, Cathy notices the keyword “disturbing” and is afraid to be troubled by potential noises. She clicks on this keyword and finds out the details as shown in Figure 6(c). There is a shared window between the bathroom and bedroom. Some customers raised a concern that one’s going to the bathroom at night may “disturb” his/her companion with light spilled through the shared window. Since Cathy travels alone, she quickly dismisses this concern of others.

6 CONCLUSION

In this paper, we propose a general framework to characterize a text corpus with multiple data facets. Based on this data model, we introduce a hybrid visual metaphor to reveal the evolutionary and correlation patterns among the data facets. Four navigation methods, one for manipulating each data facet, as well as several customized interactions, are developed to assist users in their data navigation and pattern-finding tasks. We show the effectiveness of our work through two real-world case studies. Our visual analytic process in each case provides an efficient approach to visual pattern finding within a large-scale text corpus.

ACKNOWLEDGEMENTS

The authors would like to thank the anonymous reviewers during the two-round review process for their valuable comments and careful edits, which help to improve the quality of this work. The authors also like to thank Xiaohua Sun from Tongji University, for her suggestions in the visualization and interaction design.

Table 3: Entity keywords for sentiment analysis of hotel reviews

Attribute	Entities
<i>room</i>	room*, house*, lobb*, bed*, building*, bathroom*, floor*, size*, ambience*, decor*, television*, iMac*, internet*, shower*, counter*, door*, schampoo*, conditioner*, TV, tub, thermostat, interior*, headboard, lamp, elevator*, furniture, window*, sofa, carpet, seat*, facilit*, chair, sound-proofing, amenitie*, wc, connector, air-condition*, washroom, curtain*, Towel*
<i>location</i>	view, views, location*, site*, park*, station*, subway*, proximit*, port*, airport*, train*, restaurant*, mountain, harbo*, shop*, MTR, mall, market*, transport, seafroont, ferry
<i>cleanliness</i>	bug*, bedbug*, trash*, toilet*, smell*, clean*
<i>service</i>	security*, safe*, desk*, service*, concierge*, personnel*, staff*, bar*, breakfast*, fitness*, coffee*, check-in*, lounge*, food*, guy*, housekeep*, bell-boy*, bellboy*, employee*, bartender*, fruit*, chocolate, snack*, meal, buffet, drink*, refreshment*, receptionist*, slipper*, bath-robe*, Express, menu, queue*, check-out, water, croissant*, juice*, sandwich*, shuttle, manager, spa

REFERENCES

- [1] American National Hospital Ambulatory Medical Care Survey, <http://www.cdc.gov/nchs/ahcd.htm>.
- [2] JSON Format, <http://www.json.org>.
- [3] TripAdvisor, <http://www.tripadvisor.com>.
- [4] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.
- [5] L. Byron and M. Wattenberg. Stacked graphs c geometry & aesthetics. In *Infovis '08*, 2008.
- [6] C. Collins, S. Carpendale, and G. Penn. Docuburst: visualizing document content using language structure. In *Eurovis '09*, 2009.
- [7] C. Collins, F. B. Viégas, and M. Wattenberg. Parallel tag clouds to explore and analyze faceted text corpora. In *VAST '09*, 2009.
- [8] S. Deerwester, S. Dumais, T. Landauer, G. Furnas, and R. Harshman. Indexing by latent semantic analysis. *Journal of the American Society of Information Science*, 41:391–407, 1990.
- [9] A. Don, E. Zheleva, M. Gregory, S. Tarkan, L. Auvil, T. Clement, B. Shneiderman, and C. Plaisant. Discovering interesting usage patterns in text collections: Integrating text mining with visualization. In *CIKM '07*, pages 213–222, 2007.
- [10] M. Ghoniem, D. Luo, J. Yang, and W. Ribarsky. Newslab: Exploratory broadcast news video analysis. In *VAST '07*, pages 123–130, 2007.
- [11] S. Havre, E. Hetzler, P. Whitney, and L. Nowell. Themeriver: Visualizing thematic changes in large document collections. *IEEE Transactions on Visualization and Computer Graphics*, 8(1):9–20, 2002.
- [12] M. A. Hearst. Tilebars: visualization of term distribution information in full text information access. In *CHI '95*, pages 59–66, 1995.
- [13] T. Hofmann. Probabilistic latent semantic indexing. In *SIGIR '99*, pages 50–57, 1999.
- [14] A. K. Jain, M. N. Murty, and P. J. Flynn. Data clustering: a review. *ACM Computing Surveys*, 31(3):264–323, 1999.
- [15] S. Liu, M. X. Zhou, S. Pan, W. Qian, W. Cai, and X. Lian. Interactive, topic-based visual text summarization and analysis. In *CIKM '09*, pages 543–552, 2009.
- [16] T. Nasukawa and T. Nagano. Text analysis and knowledge mining system. *IBM System Journal*, 40(4):967–984, 2001.
- [17] K. Salomatin, Y. Yang, and A. Lad. Multi-field correlated topic modeling. In *SDM '09*, pages 628–637, 2009.
- [18] G. Salton and M. J. McGill. *Introduction to Modern Information Retrieval*. McGraw-Hill, New York, NY, USA, 1986.
- [19] M. Sarkar and M. H. Brown. Graphical fisheye view. *Communications of the ACM*, 37(12):73–83, 1994.
- [20] J. Stasko, C. Görg, and Z. Liu. Jigsaw: Supporting investigative analysis through interactive visualization. *Information Visualization*,

