名词，动词, 例子

only looking at subsection to avoid information repetition.

Text flow

**5 VISUAL DESIGN**

Although the mining results produced by the topic miner well summarize the topic evolutions in the input documents, they are abstract and difficult to understand. We thus developed a Texflow visualizer for presenting the three-level mining results comprehensively in a single view. Specifically, a topic flow graph inspired by the river metaphor is presented for visualizing the overall topic evolution patterns (Sec- tion 5.1), a set of glyphs is applied over the flow for encoding the critical events (Section 5.2), and a set of keyword threads is designed to reveal the keyword correlation details in each topic evolution flow (Section 5.3). Combined with a tag cloud view and a timeline view (Section 5.4), the Texflow system provides an efficient solution to various visual exploration and analysis tasks.

**5.1 Topic Evolution as Flow**

A naive solution to visualizing the topic evolution is to display all topic evolutions as a stacked graph. Although this approach can well illustrate the linear topic evolution along time [4, 9], it cannot well present complicated topic evolution patterns that include topic merging and splitting. To tackle this issue, we adopt a river flow metaphor (see Fig. 4) for visually conveying the topic evolution patterns in Textflow.

As shown in Fig. 4, we represent the evolution of topics along time as a topic flow graph. The varying flow height along the x-axis encodes the number of the documents that belong to this topic at each time point. Like a river flow in the real world, the topic flow can be either split into several branches when the corresponding topic splits, or merged with several other branches into one flow when their corresponding topics merge into one topic. After splitting (or merging), the branch with the content that is most similar to the topic before splitting (or after merging) becomes the main branch, with the rest as the secondary branches (Fig. 5). The heights of all branches are determined by the number of documents in the corresponding branch topics.

To visualize evolution patterns, we stack the topic flows of all topics vertically and align them based on their time stamps. To determine the order of the topic flows in the stack, we develop a flow layout algorithm, which can reduce the visual clutter caused by flow branch crossings and generate a smooth flow outline. The details about our flow layout algorithm are described in Section 6. To help users easily identify and track the specific topic flow(s), we uniformly pick colors from a rainbow color spectrum for original topic flows according to its vertical order in the stack. For the derived topics generated by merging and splitting, we adopt the following coloring strategy. The color of the merged topic flow is the blending of the colors of the corresponding branches. The blending weights are determined by the ratio of branch heights to the original flow height. The color of each branch after topic splitting is determined from the color of the original flow similarly. As a result, the color of the topic is actually influenced by the color of multiple topics in its history and thus provides a good supplementary visual cues for topic tracking. Our color scheme may cause misleading colors in some cases. For example, too many splitting/merging relationships likely lead to very similar colors among different flows. In such case, our color strategy may fail to differentiate related topics, users can then rely on the flow shape to track topics.

Our topic flow graph can be regarded as an enhancement of the traditional stacked graph. By hiding the secondary branches, the topic flow graph can be easily transformed into a stacked graph (see Fig. 5). In our implementation, we use the stacked graph as the initial visualization of the overall topic evolution patterns and then transform the visualization into a topic flow graph for visualizing the detailed topic merging and splitting patterns in which users are interested. This scheme provides an easy way for users to understand and explore the complicated topic evolution patterns and flatten their learning curve.

**5.2 Critical Event as Glyph**

Four intuitive glyphs are chosen to represent the birth, death, splitting, and merging events, respectively (see Fig. 6). We overlay these glyphs at the time points where they occur to illustrate the critical events of the corresponding topic flow. In our data model, topics are represented by a set of keywords. Therefore, those glyphs can also help users find intense keyword changes (see Fig. 4). The size of the glyphs is computed by the importance score of the critical events described in Section 4.2. The bigger a glyph is, the more important its corresponding critical event is.

**5.3 Keyword Correlation as Thread**

The keyword correlation is the third aspect that we consider. It is very important in topic analysis, because it allows users to understand why the topics develop and change at a finer granularity.

A straightforward way to visually encode keyword correlations is to overlay word clouds on the topic flow (see Fig. 7(a)), which is also probably the most common method to show a collection of keywords to users in the field of information visualization [8, 26, 25]. However, this representation is not effective and useful in our scenario. First, a single word cloud fails to show the content evolution of keyword correlations. Even if we can adopt a set of word clouds to encode the content evolution over time, this may introduce additional visual clutter to the topic flow. Second, it cannot illustrate the co-occurrence interactions between keywords over time, which characterizes the ma- jor critical events in topics.

Another option for capturing the dynamic content and co- occurrence interactions over time is to use a set of polylines to represent the keywords (Fig. 7(b)). Cooccurrence relations could be represented by distances between them. In addition, a force-based model could be adopted to lay out the polylines. However, we can not use it in our design either. The distance between the polylines encodes their closeness. On the other hand, the distance is also constrained by flow boundaries. For example, at the time point where the flow is thin, the polylines may not be close to each other. However, because of the lim- ited space, they have to be placed very close and may give users the wrong impression (see Fig. 7 (b)).

Instead, we design a novel visual primitive called *thread* to solve these problems. Fig. 4 shows a simple example of thread. Intuitively, if one keyword appears in a topic at a specific time point and disappears from the topic at another time point later, we draw a thread in that topic flow between the two time points. Meanwhile, we use the weaving effects (see Fig. 8) to represent the co-occurrence interactions between two or more keywords. In particular, we select one keyword as the primary keyword (the thick line in Fig. 8). In the thread representation, we only consider the co-occurrence interactions between the primary keyword and other non-primary keywords.

We also adopt various visual attributes in thread weaving to encode different information:

• Wave bundle: If several threads are weaving together, we use the weaving bundle to encode the co-occurrence interaction at that time point. The length of the wave bundle represents the time period length of the co-occurrence.

• Amplitude: We use the amplitude to encode the occurrence number of all involved keywords. The bigger the weaving amplitude, the more the corresponding weaving keywords appear at that time point. Usually the height of a topic at one time point is proportional to its related keyword number at that particular time point, so the topic flow can well accommodate the thread weaving with different amplitudes.

Opinion seer

**6 OPINION VISUALIZATION**

To assist users in visually analyzing the complex opinion data effectively, we developed an opinion visualization system that includes the opinion wheel, the tag cloud spreadsheet, and a set of tailored user interactions. Our design principles include effectiveness, intuitiveness, and attraction. Simplicity or intuitiveness is strongly required because our end users do not have much background on information technology, while the visualization should be aesthetically appealing because the users want to present their findings directly to a wider audience. By working closely with our target users, we developed a visualization system that could convey the results of the opinion mining, from simple to complex, while keeping its intuitiveness.

The system has two major views, an opinion wheel (Fig. 1) and tag clouds (Fig. 5). The opinion wheel seamlessly integrates a scatterplot (opinion triangle) with a radial visualization (opinion ring). The opinion triangle is primarily used for visualizing the extracted opinions, each of which is an opinion vector (*b*,*d*,*u*) with three elements: nega- tive, positive, and uncertainty values. The three vertices of the opinion triangle represent the most negative, positive, and uncertain opinions, respectively. Each customer opinion is plotted in the opinion triangle according to the distance from the three triangle vertices. For example in Fig. 7(a), an opinion shown in the lower left of the triangle means a negative opinion, in the lower right means a positive opinion, and in the top part means an opinion with high uncertainty. The opinion rings surrounding the triangle facilitate the visual exploration of correlations between the customer opinions and other data dimensions. The opinions in the triangle are projected onto the opinion rings to create circular histograms of different data dimensions. Furthermore, to help user examine the real reason of a certain opinion as well as to compare customer reviews, a diagram of tag clouds is synchronized with the opinion wheel. In this section, we will discuss our opinion visualiza- tion design and share our experience in collaboration with hospitality researchers for developing the opinion visualization system.

**6.1 Opinion Wheel: Integrated Visualization of Customer Opinion Data**

The major visualization of OpinionSeer is an opinion wheel, which is a tight integration of a scatterplot and a radial visualization. The opinion posts or features are represented by a scatterplot inside an opinion tri- angle. In the scatterplot, each point encodes an opinion post or feature.

The radial visualization is the bounding wheel of the opinion triangle. We adopt it to illustrate visually the correlations among multiple data dimensions (e.g., age, gender).

6.1.1 **Opinion Triangle**

Customer opinions are the center of customer feedback data, and play a key role in visual opinion analysis. In hospitality research, the gen- eral customer feedback analysis usually starts from customer opinions. Thus, the first step of our design is to determine a reasonable visual representation for the opinions. As described in Section 5.3, each extracted opinion is represented as an opinion vector (*b*,*d*,*u*), where *b* + *d* + *u* = 1. Proper visual encoding of the opinion vector is difficult using traditional information visualization techniques such as parallel coordinates because the important characteristic, *b* + *d* + *u* = 1, of the opinion vector cannot be clearly revealed. On the other hand, in an equilateral triangle, the sum of distances from any point in the interior of an equilateral triangle to all three sides is always equal to the height of the triangle. Thus, this triangle property can be used to visually encode the characteristic of the opinion vector (i.e., *b* + *d* + *u* = 1).

An opinion vector *x* = (*bx*,*dx*,*ux*) could be mapped to a point inside an equilateral triangle △*ABC* (Fig. 3) whose height is equal to 1. Vertices *A*, *B*, and *C* denote disbelief, uncertainty, and belief, re- spectively. To achieve this, we draw two lines *IJ* and *DE* which are parallel to *BC* and *AC*, respectively. Additionally, we make sure that the distance between *IJ* and *BC* is equal to *dx*. Similarly, the distance between *DE* and *AC* is equal to *ux*. The intersection point *P* of *IJ* and *DE* is the point that represents the opinion vector, *x*, inside the triangle. The distances from *P* to the three sides *BC*, *AB*, and *AC* are *dx*, *bx*, and *ux*, respectively. The sum of the distances is equal to the height of the triangle, that is, *bx* + *dx* + *ux* = 1. With the visual encoding method, all opinion vectors could be intuitively shown inside a triangle-style scatterplot, which is also called an opinion triangle in subjective logic [14]. For example, a strong negative opinion could be represented by a point toward the left disbelief vertex of the opinion triangle. Similarly, an opinion with a high degree of uncertainty could be represented by a point toward the top uncertainty vertex of the opinion triangle.

The opinion triangle used together with the subjective logic operators can greatly facilitate visual opinion comparison of different groups of customers. After separately applying the FUSION operator to the opinions of every selected group, we could obtain several fused opinion points inside the triangle; each point represents a fused opinion. By comparing these opinion points inside the opinion triangle, we could readily identify the differences of the customer opinion groups. This capability could then solve Q1 and Q2 described in Section 3.3. Compared with other visual metaphors, the opinion triangle could present the uncertainty information naturally; it is also a scatter- plot familiar to and used frequently by our target users. Thus, they can start with a familiar format.

Finding opinion patterns regarding categorical information is a fundamental task in hospitality research. In this section, we introduce our adapted visualization approach based on scatterplots, glyphs, and radial visualization layouts to facilitate this task.

Coordinated View versus Integrated View To find opinion patterns and correlations among different dimensions, the extracted opinions need to be analyzed in context, which requires simultaneous visualization of the multidimensional information. One straightforward solution is to provide users with multiple views coordinated with the opinion triangle. Each view focuses on one data dimension. Our initial prototype system includes multiple coordinated views: an opinion triangle view for extracted opinions, five bar chart views of related demographic information and temporal information, a parallel coordinates plot to reveal the relationship between temporal and geographic dimensions, and a map view for geographic information. After presenting and discussing the system to our target users, we did not adopt this approach as the users thought it was difficult for them to relate information scattered in multiple views to find interesting opinion patterns. To address the issue, we attempted to develop a comprehensive visual representation of the data capable of providing an integrated visualization of multidimensional data rather than multiple separate views. Although this would possibly introduce visual clutter when showing too much information simultaneously, we could keep the visual clutter at an acceptable level through proper design and user interactions.

Glyph-based Encoding We started our design from the opinion triangle, which is a triangle-style scatterplot. Each opinion point is associated with one opinion holder (i.e., the customer). Hence, we could simply utilize glyphs, geometric objects with different visual proper- ties, to encode multidimensional categorical information of the opinion holders inside the triangle. Some visual properties of glyphs such as color, shape, and size are available if we require rapid pre-attentive processing [30]. After discussing with our target users, however, we found it was not necessary to show too much information simultaneously in the scatterplot for the following reasons. First, regarding the general analysis tasks (Q3, Q4, and Q5) listed in Section 3.3, users only need to examine the relationship between customer opinions and another categorical dimension one by one, therefore unused dimensions are considered unnecessary. Second, with respect to the tasks related to temporal and geographic dimensions (Q6 and Q7), users may need to analyze multiple dimensions (opinions, demographic, temporal, and geographic information) simultaneously to find temporal and spatial opinion patterns, but the temporal and geographic dimensions cannot be encoded easily by glyphs. While many different locations and time ranges exist, the number of categories that each glyph property could encode is limited [30]. For example, no more than eight colors should be adopted if we want to understand data values quickly. Therefore, inside the opinion triangle, only two pre-attentive visual properties (color and shape) are employed for the glyphs. Color is used to encode the categories of a categorical dimension (e.g., age range), while shape is utilized to represent the groups of the opinions (e.g., room, service, and price).

**Categorical Ring** Scatterplot with glyphs can show an overall in- formation distribution of a certain dimension such as a distribution of age groups over opinions. However, in our application, a large number of customer opinions could be explored, which may introduce severe visual clutter. Consequently, it is difficult to find opinion relation- ships with respect to Q3 and Q5, not to mention the visual comparison regarding Q4. To alleviate the problem and improve the scatterplot readability, we incorporated a radial visualization layout into our opinion triangle. Radial visualization is an increasingly prevalent visual metaphor with a compact and aesthetically appealing layout in infor- mation visualization and visual analytics [7]. Compared with other existing radial visualization, our approach has two unique features: First, our radial layout supports the subjective logic and accounts for uncertainty. Second, we provide an integrated view of multiple impor- tant data dimensions specifically designed for opinion visualization. The basic idea of our approach is to project customer opinions in the interior of the opinion triangle to its circumscribed ring (called categorical ring), and then visualize the categories of the dimension to be examined on the sectors of the ring.

To ensure effective visualization, we first designed five different layouts using pre-attentive visual properties including color and size to display the category information on the sectors of the categorical ring, as illustrated in Fig. 4(a) - (e). These radial layouts were then presented to our two target users for user evaluation.

Both users rejected the design in Fig. 4(a) because it was difficult for them to associate depth of color with weighted average. They com- plained that it lost more information than other layouts. The layouts shown in Figs. 4(d) and (e) were received well by a user. He pointed out that size is visually more intuitive to associate with numbers or volumes than color depth, hence the layouts shown in Figs. 4(a), (b), and (c) were not preferred. He also felt that having different colors to represent different categories make it easier to identify than hav- ing similar or the same color schemes such as in Figs. 4(b) and (c). Additionally, he suggested that grouping the information neatly into sectors, like in Figs. 4(c) and (d), should be much better than Fig. 4(e). Another user especially like Fig. 4(e), as it is less complicated and the quantity information is width-oriented. In addition, it is easy to identify what information to be communicated in one glimpse. All the others are less preferred by the user because they are all required to read additional chart/table in order to find out what is going on and to understand. To conclude, it is better that the layout uses different sizes to indicate the number of customers in a particular category, together with different colors to represent various categories.

Based on the user feedback, we developed a new radial layout shown in Fig. 4(f) in which information of a particular dimension (e.g., age range) is projected to the circumscribed ring. Each sector is divided into multiple parts along the radius direction and each part corresponds to a specific category of customer ( an age group in this example). The size of each part is determined by the number of customers that belong to the corresponding category. Different colors are used to differentiate different age groups. This layout could be viewed as circular stacked bar charts. With this design, users can identify how the information dimension examined could affect customer opinions (Q3). If we project customer’ score ratings to the ring, we could also examine the relationship between the score ratings and the opinions extracted from the free-text comments (Q5). To enable a side-by-side visual comparison of the distributions (Q4), we first represent the opinion points using different shapes inside the opinion triangle for different groups of customers. Each sector on the ring is now equally divided into multiple subsectors, and each subsector is associated with one group of the customers. This allows users to visually compare the data distributions readily around the opinion triangle (Fig. 9).

Temporal and Geographic Rings

For Q6 and Q7, the temporal information (date of stay) and geographic information (customer location) should be presented to users for analysis. However, this information cannot be conveyed effectively by the categorical ring because they possess special features. The temporal and geographic dimensions usually contain more categories than others. In addition, the temporal information has unique multi-scale periodic patterns, and the geographic information has special directional patterns that cannot be revealed. Nevertheless, radial visualization is still well-suited for revealing both periodic and directional patterns [7]. Thus, we add a temporal ring and a geographic ring to the opinion wheel to visualize effectively the temporal and geographic information, respectively.

**The temporal rings** can be configured to different styles showing temporal information at different levels of detail based on user requirements, as illustrated in Fig. 1(a). The number of opinions expressed during a specific time range is encoded as the color in the sector associated with the related time range. Fig. 1(b) shows a geographic ring separated into a number of sectors; each sector corresponds to a location, such that the geographic direction of a location could be roughly revealed by the corresponding sector. The number of customer opinions from a location is encoded as a color in the sector associated with the related direction. The luminance (white-black) channel is used to encode the number in the sectors for both temporal and geographic rings because of its capacity to show data detail [26].

Although our design can address Q6, it is still difficult to find the relationships between temporal and geographic information (Q7). Inspired by Parallel Sets [18] which could effectively reveal relationships between category dimensions, we develop a technique to visually relate information between temporal and geographic dimensions. Fig. 1(b) shows the temporal ring and geographic ring simultaneously in the opinion wheel. The relationships could be revealed by connecting related categories using curved belts rather than parallelograms in Parallel Sets. Compared with Parallel Sets which show many-to-many relationships, our technique only shows a one-to-many relationship. Details are shown on demand using connections for only the selected sector on the temporal or geographic rings. This was motivated by explicit feedback from our target users on reducing information overload and visual clutter.

6.1.3 Multi-scale Exploration

The opinion wheel allows users to analyze customer opinions at different levels of detail. For instance, users could analyze customer opinions at the feature level when the opinions on a specific hotel feature or a set of hotel features are analyzed. With this visualization, users could visually compare the opinion distributions of two hotel attributes in- side the opinion triangle. The “AND” operator is exploited to combine customer opinions on different hotel attributes to facilitate the exploration at a higher level. If all feature opinions of each customer are combined using the “AND” operator, the overall customer opinions on hotels could be viewed and analyzed by users. Another operator “FUSION” could combine user opinions of different customers. Thus, users can fuse a group of opinions on a particular hotel feature of dif- ferent customers, or fuse a group of combined opinions (obtained by “AND” at the feature level) of different customers. This allows for visual analysis of customer opinions at multi-scale customer levels.