Exploring Controversy via Sentiment Divergences of Aspects in Reviews

Jin Xu* Yubo Tao† Hai Lin‡ Rongjie Zhu§ Yuyu Yan¶

State Key Laboratory of CAD&CG, Zhejiang University

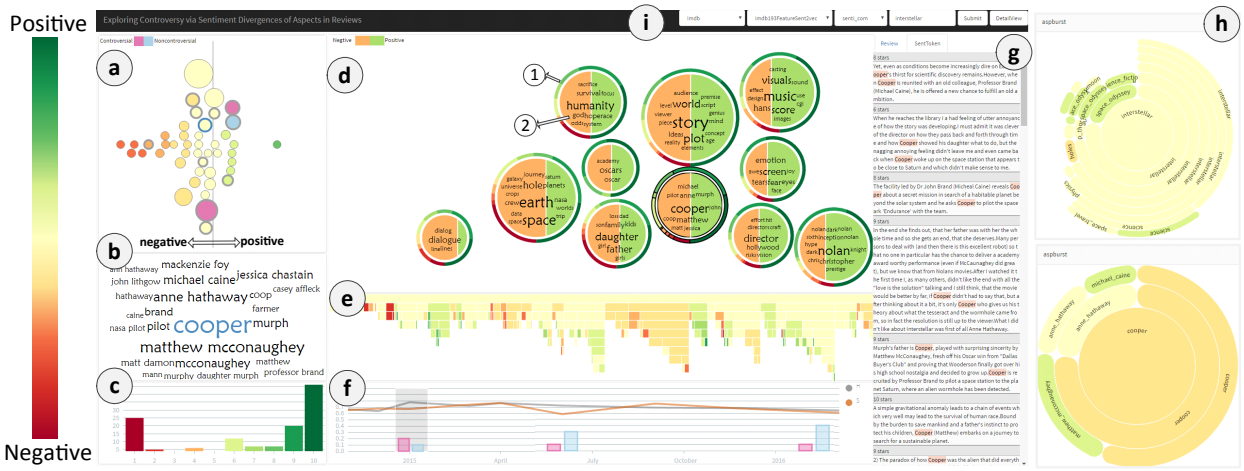

Figure 1: Our system explores the controversy in reviews with elaborated visual analytics components. a) the aspect bubble view shows aspects aligned with their sentiments, b) the word cloud view provides an overview of features in an aspect, c) the bar chart view depicts the rating distribution, d) the sentiment pie view displays the sentiment divergences of aspects
, e) the partition tree view presents the hierarchical structure of all aspects, f) the line chart view shows two controversial indexes over time, g) the text view provides the original reviews, h) the aspect burst view reveals the hierarchical structure of an aspect, and i) is the control panel.

Abstract

A visual summary of the controversial aspects of an item enables both customers and marketers to identify and address complaints and concerns about the item effectively. In this paper, we propose a novel visual analytics system, to visually explore when a controversy occurs and the causes behind the controversy via user-generated reviews with text and ratings in various domains, such as restaurants, home goods, and cultural products. Quantitative analysis of the ratings of an item is first applied to characterize the evolution of controversy over time. A novel aspect-extraction method based on hierarchical clustering is proposed to identify aspect-level reasons garnered from review texts that explain why a controversy occurs. Our system allows the user to interactively explore the time-evolving controversy trend, major aspects of reviews, and sentiment divergences of aspects to understand in depth the controversy in reviews. We evaluate the effectiveness of the proposed aspect-extraction method by means of accuracy of aspect identification, the usefulness of our system using three case studies in different domains, and a user study.

Index Terms: I.3.6 [Computer Graphics]: Methodology and Techniques—Interaction techniques

*e-mail: jinxcoder@gmail.com
†e-mail: taoyubo@cad.zju.edu.cn (Corresponding Author)
‡e-mail: lin@cad.zju.edu.cn
§e-mial: rongjiezhu@zju.edu.cn
¶e-mial: yanyuyu001@gmail.com

2017 IEEE Pacific Visualization Symposium (PacificVis)
18-21 April, Seoul, Korea
978-1-5090-5738-2/17/$31.00 ©2017 IEEE

1 Introduction

One of the most vital functions of social media is to allow users to express their opinions about issues, such as products, movies, and social events, over the Internet. These user-generated contents have been successfully applied to study many social phenomena, for example, peer influence [16, 36, 40], framing [13, 23], bias [28], and controversy [15]. Controversy refers to a phenomenon where people have divergent views or sentiments on a topic. For one topic or product, some people may express a positive attitude, while others may have a negative opinion. Such controversy occurs over the gun control, abortion, nuclear power, religion, the location of an airport, and the influence of a movie. By understanding how controversy develops within the court of public opinion, we can gain many new insights into how a controversy arises and evolves, learn more about various roles of different citizens in policy making, identify those factors which customers care most about in their decision making, and analyze how cultural or demographic differences affect people's views. Recently, in particular, large companies use controversial marketing tactics in order to spark interests in a wide population of customers and improve their business brands.

A large body of previous research has been devoted to the controversy in events and topics. The studies often leveraged data on political debates [9] or presidential elections [1] and the data on Twitter [12, 27] or other social media platforms, such as blogs [1] and opinion fora [2]. Most of this research constructed networks, such as citation networks [1], retweet and mention networks [12], and signed bipartite networks [2], to judge whether there is a controversy from the network structure. Other research studied the controversy by text mining and sentiment analysis. For instance, Livne et al. [22] employed Language Models to model user-generated contents and studied the differences between users in the usage patterns on Twitter. Cao et al. [9] estimated users' sentiments on Twitter and

clustered users according to their sentiment trends over time to show the sentiment divergence evolution in two communities. However, each event or topic may have several aspects, and previous research did not explain the causes of the controversy from the aspect level.

Relatively little research has focused on controversy in web reviews [3, 8]. The studies mostly examined controversy based on ratings of items, such as movies and restaurants, and allowed users to identify and address complaints. For instance, Amendola et al. [3] quantified polarized and uniform rating distributions to identify controversy over a movie. However, this research only depends on quantitative analysis of ratings to judge whether controversy occurs, and may require reading all the reviews one-by-one to summarize the causes of a controversy. This would be very time-consuming and prone to ignoring some potentially important aspects.

In order to overcome these limitations, we employ both ratings and text of reviews to explore when and why a controversy occurs from the aspect level. This can help marketers and customers better understand controversial aspects of an item in large-scale reviews. Rating analysis identifies when a controversy may occur, and text analysis further verifies whether a controversy really exists and extracts what the controversy resulted from. Instead of looking at review-level sentiment divergences [35], we analyze aspect-level sentiment divergences to identify the causes of a controversy with a greater degree of granularity. When divergence occurs in some aspects, i.e., some reviews have a positive attitude to these aspects, and others have a negative attitude, it is clear that people have divergent views on the item and these aspects are controversial aspects, which result in a controversy. Since the controversy of an item may change over time, we also need to characterize how the controversy evolves over time and when a controversy occurs.

In this paper, we propose a novel visual analytics system, integrating quantitative and qualitative analysis methods to explore a controversy over time in web reviews in text and ratings in various domains (e.g., restaurants, home goods, and cultural products). Our system is able to highlight two major factors: when — the time-evolving controversy trend and why — the sentiment divergences of aspects. We first apply quantitative analysis to the ratings of reviews to characterize the controversy evolution over time. Then, we propose a new method based on hierarchical clustering to accurately extract aspects from the text of reviews during the time period when the controversy occurs. After that, sentiment analysis is performed to compute the overall sentiments of the extracted aspects.

Finally, we design a visual analytics system to intuitively represent the analysis results and support comparison of sentiment divergences of aspects effectively. The sentiment divergence of an aspect is visually encoded by a divergence glyph, which integrates the sentiment, the topic, and the rating distribution. Together with other visualization components, our system supports interactively identifying controversial aspects and understanding the controversy behind the reviews effectively.

Our method offers three main contributions as follows:

- We present a framework leveraging rating and text analysis to explore when and why a controversy occurs in reviews.
- We propose a new aspect-extraction method based on sense embeddings and hierarchical clustering to identify the causes of a controversy based on the aspect-level sentiment divergence.
- We introduce a visual analytics system based on well-designed visualization components and intuitive interactions for supporting interactive exploration of a controversy. Our system is demonstrated in three case studies from different domains to show insights gained from controversial aspects of items.

2 Related Work

In this section, we review related works from two perspectives: controversy quantification and controversy visualization.

2.1 Controversy Quantification

The issue of controversy in social media and online news has been widely studied, and many methods to quantify controversies have been proposed. Most of them are based on network construction and social structure analysis. Adamic et al. [1] constructed a citation network and measured the degree of interaction between liberal and conservative blogs during the U.S. Presidential Election. Conover et al. [12] established retweet and mention networks and performed community detection using tweets during the U.S. congressional midterm elections. Akoglu et al. [2] constructed signed bipartite networks based on online forum data to represent opinions of individuals and quantified political polarity by the Loopy Belief Propagation algorithm. Garimella et al. [15] developed a random-walk-based measure on the retweet network to quantify the controversiality of a topic. Beyond a single topic, they can also compare multiple topics in any domain. These works could quantify the controversy over one single event such as the U.S. Presidential Election, and one topic with specific hashtags on Twitter. They focused on whether the event or topic was controversial and how controversial it was, but generally failed to explore the controversy from the aspect level.

Beyond network-based controversy quantification, there are also other methods to quantify controversies. Livne et al. [22] applied Language Models to model the data on Twitter by the candidates during the U.S. midterm elections, and used the Kullback-Leibler divergence to analyze differences between Democrats, Republicans and Tea Party candidates. Yasseri et al. [41] quantified the controversy of an article based on its editorial history by summing the weights of all mutually reverting editor pairs in Wikipedia. Brigadir et al. [7] studied language patterns in social and political contexts based on Distributional Semantic Models (DSMs). They considered changes over time and between communities with differing views. Amendola et al. [3] studied controversy in movie ratings in the Internet Movie database (IMDb) with two statistical indexes, dubbed hard and soft controversy, which quantified polarized and uniform rating distributions, respectively. They found that more recent movies were more controversial than older ones. We extend their statistical indexes to time-dependent indexes to characterize the controversy evolution over time based on the rating distributions. Besides the ratings of reviews, we also analyze the text of reviews by aspect extraction and sentiment analysis to explain the causes of the controversy from the aspect level.

2.2 Controversy Visualization

Many simple methods have been used to visualize the quantified degree of a controversy, such as bar charts [8], line charts [3], scatter plots [19], and box plots [24]. They mostly focus on the properties of a controversy. Constructed static networks are also displayed by node-link diagrams [15] and adjacency matric [2] in many controversy quantification papers.

There are a few visualization systems for controversy analysis. Brandes et al. [6] proposed a system based on a node-link diagram to depict the *who revises whom-network* to study controversy in encyclopedias. Yasseri et al. [41] applied the searchCrystal toolset to provide aspects of the overlap structure in multiple languages in Wikipedia to visualize highly contested pages. Cao et al. [9] grouped users based on their sentiment trends and designed a representation method based on DNA helices to visualize sentiment divergences between two groups. Frequent keywords are also presented to explore the reasons of the sentiment divergence. Our system focuses on controversy in web reviews via aspect-level sentiment divergences, which are visually encoded by a divergence glyph.

Similar to our aspect-level sentiment divergence visualization, there are several studies on sentiment analysis of reviews. Oelke et al. [29] interactively analyzed comments and ratings to determine the sentiments expressed by customers. Shi et al. [31] proposed a data model with multiple facets including topics and sentiments

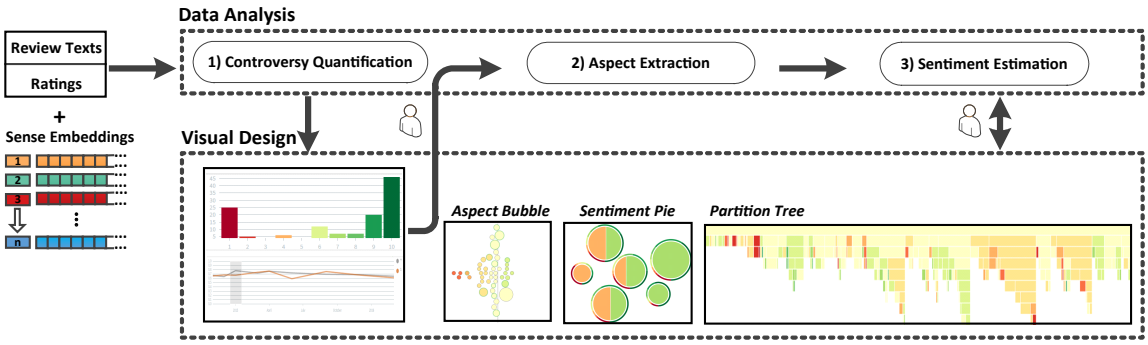

Figure 2: System overview. Data analysis module first quantifies a controversy via the ratings, extracts aspects from review text, and estimates the sentiment toward aspects. Visual design module shows the controversy index evolution over time, and designs three linked views to analyze the cause of a controversy.

and designed a hybrid visualization to facilitate users understanding of text corpora. Since a user may express both positive and negative sentiments on one topic, Wu et al. [37] studied uncertainty information in opinion extraction, combination, and visualization. Duan et al. [14] proposed a generic sentiment tuple to build a visual sentiment analysis system from aspects. However, they did not fully consider the relationships between controversy and sentiment divergences. We propose a new aspect-extraction method and we analyze controversy via aspect-level sentiment divergences.

3 Overview

We first formulate tasks for our system to explore controversy over time in web reviews. Considering the two major requirements of the system: when — the time-evolving controversy trend and why — the sentiment divergences of aspects, we propose four tasks (T):

T1: Characterize the time-evolving controversy trend. Since controversy in web reviews evolves over time, we need to characterize the time-evolving controversy trend. This trend enables users to quickly identify whether an item is controversial and when a controversy occurs. This is a fundamental task in our system.

T2: Extract aspects of reviews for theme summary. We aim to summarize the causes of controversy in terms of aspects. Since there may be hundreds of reviews, it would be time-consuming to read all reviews during controversy analysis. Aspects are usually considered as themes, and reviews may contain different aspects. For a product, people talk about not only general aspects, such as quality and price of a pair of headphones, but also particular aspects, such as suitable situations for using a pair of headphones. Thus, this method would be an effective means to extract aspects automatically and summarize the themes of reviews.

T3: Understand controversy from aspect-level sentiment divergences. Reviews may involve many aspects and the sentiments may vary with the aspects. If there are roughly half positive and half negative reviews for an aspect, this aspect can be considered to be one cause of a controversy. Thus, we need to identify all the aspects having sentiment divergence to fully understand a controversy.

T4: Relate aspects to the original reviews. It is essential for users to retrieve and analyze the original reviews at any stage of the previous tasks. An aspect can guide users to quickly identify related reviews and further understand the aspects by integrating their domain knowledge. Moreover, as it is generally difficult for aspect-level sentiment analysis to reach human-level accuracy, this task can also verify the analysis results of previous tasks.

These tasks frame the general requirements of our system. We need to determine when a controversy occurs based on the time-evolving controversy trend, and analyze the causes of the controversy based on the aspect-level sentiment divergence and the original reviews. Fig. 2 shows the pipeline of our system. The input of the system is the reviews of an item, consisting of both review text and ratings. It has two major modules: data analysis and visual design.

The data analysis module first computes the time-evolving controversy trend with quantitative analysis of the ratings of an item. The aspects of the reviews within a selected time period are extracted based on sense embedding and hierarchical clustering. The Sentiments of these aspects are estimated to explore the causes of the controversy.

The visual design module first shows the time-evolving controversy trend with the line chart and the bar chart. The aspect-level sentiments are presented in three linked views, namely, the aspect bubble view, the sentiment pie view, and the partition tree view. The aspect bubble view represents the sentiment distribution of aspects. The sentiment pie view helps users understand the detailed information of aspects. The partition tree view shows the hierarchical structure of aspects. The well-designed system incorporates intuitive interactions to support comparison of sentiment divergences of aspects and identification of the causes of a controversy effectively.

4 Data Analysis

The data analysis of reviews consists of three steps presented in the following sub-sections: controversy quantification, aspect extraction, and sentiment estimation.

4.1 Controversy Quantification

Controversy quantification is a fundamental step to identify whether an item is controversial and when the controversy occurs. Since the review dataset contains both ratings and text, it is different from the data on Twitter. Previous controversy quantification methods based on network structure analysis or with two explicit communities are not suitable for web reviews. Moreover, when controversy does not occur in reviews, it is time-consuming and ineffective to analyze controversy with text mining or other complicated methods. Thus, we first apply quantitative analysis to the ratings of reviews.

Ammendola et al. [3] considered a situation with many very-high ratings and many very-low ratings to be a hard controversy, and a situation with most ratings across a broad spectrum as a soft controversy. They proposed two indexes H and S as normalized measures of hard and soft controversies. The estimator similar to H was advocated by previous works [43], while the estimator of S was first introduced and achieved a good performance in controversy quantification. Since controversy may change over time, we extend the two indexes H and S in the time window with the width Δt, representing months, years, etc. Low H, high S, and high H characterize the three rating distributions: peaked, flat, and polarized, respectively. Thus, in the case of high S and low H, soft controversy tends to occur, and in the case of high H and low S, hard controversy tends to occur.

The hard controversy index H_i with the time window t_i is defined as a normalized standard deviation:

$$H_i = \frac{1}{c_H}[\sum_{m=1}^{M} p_m(r_m - \bar{r})^2]^{1/2}, \quad (1)$$

where M is the number of the rating level, such as the five-level rating from 1 to 5. In $p_m = v_m/N$, v_m is the number of votes of the rating $r_m = m$, and N is the total number of votes during the time window t_i. $\bar{r} = \sum_m p_m r_m$ is the average rating, and c_H^2 is the highest possible variance. $c_H = (M-1)/2$ is obtained if half the ratings are 1 and half are M (a completely polarized distribution). The higher the value of H_i is, the more polarized the rating distribution is.

The soft controversy index S_i with the time window t_i is defined as follows:

$$S_i = 1 - \frac{1}{c_S}[\sum_{m=1}^{M}(p_m - \frac{1}{M})^2]^{1/2}, \quad (2)$$

which is a square root of a χ^2 statics relative to the flat distribution with $p_m = 1/M$, normalized with $c_S = \sqrt{1-1/M}$. S_i vanishes if all votes are given to a single rating m. The higher the value of S_i is, the more even the rating distribution is.

Since a small number of votes generally leads to decreased performances when calculating H and S, we extend the time window until the number of votes is above a threshold when there are too few votes in the current time window (a month in our implementation). If there are not enough reviews about an item, it is biased to judge whether it is controversial from available reviews.

4.2 Aspect Extraction

Within the controversial time period, we need to extract aspects in reviews to analyze the causes of a controversy in a more understandable manner.

Features are components and attributes of an item, and they are mentioned as opinion targets. An aspect can be considered as a group of features. For example, features are terms like *pizza* and *pie*, and grouping them can obtain an aspect referring *food*.

Aspect extraction is a hot topic in natural language processing [4, 33]. Chen et al. [10] proposed a clustering method to simultaneously identify features and group them into aspects based on their domain-specific similarities and merging constraints. This clustering method has been proved to have a better performance than other methods.

Our aspect-extraction method consists of three steps: feature extraction and representation, partition generation, and partition merging. We first extract features and represent features via sense embeddings. The k-nearest-neighbor algorithm is employed to generate partitions, and these partitions are hierarchically merged to generate aspects. Compared to Chen et al. [10], our method employs an improved hierarchical clustering method that does not need the number of aspects and uses the similarity measure based on sense embeddings to better group features with similar semantic meanings.

4.2.1 Feature Extraction and Representation

As features are generally the attributes and components that describe the characteristics of an item, we consider nouns and consecutive noun phrases as features based on the Stanford Log-linear Part-Of-Speech Tagger. We filter out the stop words, which are mostly common but unimportant. The low frequency features are also filtered out, as the high frequency features are more likely the actual features of interest.

Although features are different, their semantic meanings may be the same. It is desirable to cluster features with similar semantic meanings, so that the terms referring to the same aspect are put into one group. The main question then is how to effectively capture the semantic meanings of features. Word embeddings [25, 26] have been proposed to generate dense, short and semantically-meaningful vectors to capture both syntactic and semantic information. They have been successfully applied to different natural language processing tasks including semantic similarity measurements [5, 11]. However, word embeddings inherit an important limitation in that they are unable to model distinct meanings of a word, despite the fact that a single word may have different parts of speech and each may have a different meaning or sense. For example, *duck* refers to the concept of a waterfowl, or the action of crouching. Therefore, we employ sense embeddings to represent multiple word senses per word [17,34,38]. Sense embeddings use different senses to represent different word types, and in particular, each sense is associated with a sense specific embedding.

As it is more accurate to capture the semantic meanings of words by using the related text corpus [39], we prepare the field-related text corpus and label the corpus with the Stanford Log-linear Part-Of-Speech Tagger. Then we train our sense embeddings by using the continuous bag-of-words (CBOW) model [26]. After training, word senses are represented by sense embeddings, and each feature corresponds to a vector in the low dimensional space, typically 50-500. Compared to the domain-specific similarities [10], we use sense embeddings instead of word embeddings to provide better semantic meanings of features.

4.2.2 Partition Generation

In order to group features to aspects, we first apply the k-nearest-neighbor algorithm to generate partitions, and then use a clustering method to hierarchically merge partitions.

The similarity between features is based on sense embeddings. There are many similarity metrics for two vectors, such as the Euclidean distance, the Jaccard similarity, and the cosine similarity. The most widely used similarity metric in word embeddings is the cosine similarity [20]. Therefore, we use the cosine similarity to measure the similarity between features. The higher the value of the cosine is, the more similar the two features are.

With the cosine similarity between features, each feature first finds its k most similar neighbors with the similarity above the user-specified threshold. For each neighbor feature in its k most similar features, we add the neighbor feature to the partition that the feature belongs to. In our experiments, we find that the meanings of features are very different when their similarities are less than 0.5. Hence, we set the similarity threshold to 0.5. In this case, when k is 2, the generated partition would contain almost all features. Thus, we set k to 1 to group the most similar features into a partition, and the number of partitions is large enough to separate different aspects.

4.2.3 Partition Merging

After all features are grouped into partitions, we need to hierarchically cluster partitions into aspects.

We first describe the similarity measurement between partitions, i.e., groups of features, as follows:

$$sim_{avg}(P_l, P_m) = \frac{\sum_{v_{i'} \in P_l} \sum_{v_{j'} \in P_m} sim(v_{i'}, v_{j'})}{|P_l| \times |P_m|}, \quad (3)$$

$$r(P_l) = argmax_{v_{i'} \in P_l} f(v_{i'}), \quad (4)$$

$$sim_{rep}(P_l, P_m) = sim(r(P_l), r(P_m)), \quad (5)$$

$$sim(P_l, P_m) = min(sim_{avg}(P_l, P_m), sim_{rep}(P_l, P_m)), \quad (6)$$

where $sim_{avg}(P_l, P_m)$ is the average similarity of features between partitions P_l and P_m, $r(P_l)$ is the most important feature measured by the TF-IDF value as the representative feature in partition P_l, and $sim_{rep}(P_l, P_m)$ is the similarity between the representative features of the two partitions. $sim(P_l, P_m)$ is the similarity between two partitions, which is the minimum value between the averaged similarity of features and the similarity of two representative features in two partitions. Two partitions are considered as similar when both the average similarity and representative similarity are high.

We hierarchically merge partitions until the minimum similarity between partitions is above the threshold, 0.3 based on our experiment.

4.3 Sentiment Estimation

We need to estimate the sentiments of aspects to understand a controversy from sentiment divergences of aspects.

An aspect is a group of features, and a feature can be represented by a group of sentences which contain the feature. Thus, the sentiment of an aspect can be estimated by the sentiments of sentences, which contain at least one feature in the aspect. There are many sentiment analysis techniques to identify whether a sentence is positive or negative, and the accuracies of these methods are between 80% and 85% [32]. We leverage Apache OpenNLP Document Categorizer to classify sentences into pre-defined categories. To improve the accuracy of sentiment classification, we prepare the corpus-related training datasets to train the Document Categorizer model. We define the sentiment of an aspect as the ratio of the number of positive sentences to the number of all sentence in the aspect. When the sentiment approaches 1 or 0, the aspect tends to be positive or negative. When the sentiment is around 0.5, the aspect tends to be controversial.

5 Visual Design

In order to visually explore a controversy in reviews, we design a visual analytics system to enable users to gain insights into when and why a controversy occurs. In response to the aforementioned tasks, we derive four design requirements to guide our design process:

R1: Controversy evolution characterization. The system should support users well in understanding the time-evolving controversy trend, which is a fundamental requirement of our system. Users can identify whether the item is controversial and when the controversy occurs.

R2: Aspect presentation. The system should provide a good overview of aspects and their sentiments for summarization. It should also make it easy to identify which aspects are mainly positive or negative and the main content of each aspect. The original reviews of the aspect should be accessible on demand.

R3: Sentiment divergence visualization. The system should enable users to visually compare the sentiments of different aspects and highlight controversial aspects with sentiment divergences. This is vital for users to understand the causes of a controversy.

R4: Usability. Since the system is designed to be used by marketers and customers without much explanation and training, it should be easy to learn and use. We use familiar representation methods with intuitive interactions to improve the usefulness of our system.

These design requirements are reflected in the visual design of our system. Fig. 1 shows the overview of our system with eight views, which are connected by brushing and linking to allow for flexibly exploring controversy from the aspect level. In the following subsections, we discuss our system according to the design requirements. The line chart and bar chart present an overview of a controversy evolution over time (R1). The aspect bubble view, the partition tree view, and the aspect burst view provide three complementary visual representations of aspects (R2). The sentiment pie view displays sentiment divergences of aspects (R3). We favor usability over interactions (R4).

5.1 Controversy Evolution Characterization

A controversy is described by two time-dependent indexes, H and S, and we employ the widely used line chart to present the controversy evolution information. The horizontal time axis is scaled automatically to the range of time that reviews were posted. The time-evolving trends of H and S are shown in the gray and chocolate colored lines, respectively. The line chart can be used to select the time period when a controversy occurs.

Although H and S reveal the controversy evolution over time, the original rating distribution should also be provided to verify the controversy assumption. The bars are encoded by the colors from red to green corresponding to the ratings from low to high. As shown in Fig. 1(f), H is higher than S and approaches 0.8 by the end of 2014. This indicates that a hard controversy (polarization) possibly occurred, which can be further verified by the rating distribution in Fig. 1(c). While in September 2015, S is a little higher than H and approaches to 0.8. This indicates that a soft controversy (even voting) possibly occurred.

5.2 Aspect Presentation

We extract aspects from the reviews within the user selected time period, and estimate the sentiments of aspects. These aspects with their sentiments are visualized in the aspect bubble view, the partition tree view, and the aspect burst view. The size in these three views is proportional to the number of sentences associated with the aspect or partition, as discussed in Section 4.3. The colors from red to green are used to encode the sentiment from negative to positive in these three views, as shown in the left part of Fig. 1.

In the aspect bubble view, a bubble represents an aspect. The sentiment from negative to positive is encoded by the horizontal position from left to right, as shown in Fig. 1(a). When two aspects have similar sentiments, they are arranged from the center to both sides to avoid overlap. The aspect bubble view can be used to provide an overview of the aspect-level sentiment distribution.

The partition tree view shows the process of the agglomerative hierarchical clustering from bottom to top, as shown in Fig. 1(e). Each rectangle is a partition. This view provides an overview of sentiments of multi-level aspects to analyze the sentiment differences of the partitions.

Besides an overview of all aspects in the partition tree view, we also provide the aspect burst view to display the hierarchical structure of an aspect for detailed analysis, including sentiments and features of its child aspects. The aspect burst view is a radial tree, and two aspect bursts are shown in Fig. 1(h). Traversing the tree from periphery to center follows a merging process of the root aspect. The text label on each sector is the most frequent feature in the aspect. The root aspect in the aspect burst view can be selected by clicking on a sector of interest. Once an aspect is selected, the aspect burst view displays the merging process of this aspect. This view facilitates users examining the sentiments of the child aspects.

5.3 Sentiment Divergence Visualization

The aspect-level sentiment divergence should be visually presented to identify the causes of a controversy within the user selected time period. This also further verifies whether the controversy actually occurs within this time period. We design a divergence glyph to support a visual comparison between aspects in the sentiment pie view.

As shown in Fig. 1(d), the divergence glyph encodes the sentiment, topic, and rating distribution of an aspect in an unified design, and consists of two parts: the inner pie diagram and the outer ring. The traditional pie diagram expresses the quality of each section through the size of the central angle or area. For multiple pie charts, however, slice sizes are very difficult to compare side-by-side. Here, the inner pie diagram is vertically split into two parts, the width of which is proportional to the number of the sentences with negative (orange) and positive (green) sentiments in an aspect. The most important features of the aspect are displayed inside the pie diagram. The font size is proportional to the importance value of the feature. A richer word cloud is provided in the word cloud view, as shown in Fig. 1(b), since the space in the pie diagram is limited to show all important features.

The outer ring around the pie diagram is a circular surrounding bar chart, which summarizes the rating distribution of the aspect. As mentioned in Section 4.3, an aspect is associated with sentences. Thus, the rating distribution of an aspect can be considered as the rating distribution of the reviews that contain the sentence. The ring is colored from red to green to indicate the rating from low to high, and the arc length of each sector is proportional to the number of reviews in this rating. We arrange the sectors of the outer ring from

Figure 3: Controversy analysis for *Interstellar* on IMDb. Both H and S have high values, and H is higher than S just after the movie is released, and then H decreases over time. The aspect bursts of the *cooper* and *matt* aspect are represented in (a) and (b), respectively.

-90 degrees to 270 degrees, which facilitates a comparison between the sentiment distribution and the rating distribution.

Similar to the aspect bubble view, all divergence glyphs are arranged from left to right according to their sentiments in the sentiment pie view. If an overlap or collision occurs, the glyph searches for a suitable position based on the Archimedean spiral. The sentiment pie view facilitates the visual summarization and comparison of sentiment divergences of aspects.

5.4 Interaction

Our visualization system supports well-designed interactions to explore a controversy effectively.

Selecting a time period. Since a controversy may change over time, users can select an interested time period via brushing in the line chart with H and S for guidance. After selecting a time period, the rating distribution and aspects are both updated based on the reviews within the time period.

Selecting aspects. As there may be many aspects in large-scale reviews, it is desirable to select aspects of interest for detailed analysis to reduce visual clutter. Users can select aspects in the aspect bubble view together with the sentiments of multi-level aspects in the partition tree view, and the selected aspects are stroked with dark gray and their detailed information is shown in the sentiment pie view.

Selecting an aspect. When users click a divergence glyph in the sentiment pie view, the divergence glyph is stroked with black, its aspect is stroked with blue in the aspect bubble view, and its aspect burst view is popped out to show the detailed sentiment and hierarchical structure of the selected aspect. The word cloud view is also updated to present more features of the aspect. When users select the left/right part of the pie diagram, the corresponding part is stroked with black and reviews are listed in the text view.

Selecting a feature. Since word clouds can guide users to quickly understand reviews, the system supports feature exploration. When users select an interested feature in the sentiment pie view or the word cloud view, it is colored blue, and the sentences are sorted by its frequency in the text view.

Labeling an aspect. During exploration, the aspects can be labeled as controversial with pink or noncontroversial with blue in the aspect bubble view, as shown in Fig. 1(a). The number of labeled aspects is visualized by the bar chart in the line chart, as shown in Fig. 1(f). This enables users to find how many controversial aspects are in a given time period. The aspects with the controversial label help users summarize their finds to the causes of the controversy.

6 Evaluation

In this section, we first evaluate the effectiveness of the proposed aspect-extraction method. After that, we show the usefulness of our controversy analysis method based on three case studies in different domains. Finally, we present a user study to evaluate our system.

6.1 Evaluation on Aspect Extraction

Recently, Chen et al. [10] proposed a clustering method to identify product aspects from web reviews, and they released a dataset for the evaluation on aspect extraction. Since our method can also be used in the product domain, we use the same dataset to compare the effectiveness of our method with other methods. Table 1 describes the dataset of two products on Amazon and one product on the online shop of a cell phone company. The training corpus of sense embeddings is Amazon Electronics data with 3,663,769 reviews.

	Cell-phone	GPS	TV
#Reviews	500	500	500
#Aspects	46	37	34
#Features	419	637	485

Table 1: Data sets and gold standards.

We compare our method against four approaches on aspect extraction, namely, MuReinf [33], L-EM [42], L-LDA [42], and CAFE [10]. MuReinf uses the mutual reinforcement association between features and opinion words to iteratively group them. L-EM applies the Naive Bayesian-based EM algorithm to group synonym features into categories. L-LDA is based on LDA to group features. CAFE groups the features into clusters based on their domain-specific similarities and merging constraints. Since MuReinf, L-EM, L-LDA only focus on aspect extraction, features are extracted by CAFE. Our method also uses CAFE to extract features.

We evaluate the results via the widely used Rand Index [30], since it is considered as a standard measure of the similarity between clustering results. The Rand Index is simply $2(a+b)/(n\times(n-1))$, where n is the number of the objects, a is the number of the pairs that belong to the same cluster in both partitions, and b is the number of the pairs that belong to different clusters in both partitions. The Rand Index lies between 0 and 1. When two partitions agree perfectly, the Rand Index is 1.

	Cell-phone	GPS	TV
CAFE+MuReinf	0.7973	0.8212	0.8334
CAFE + L-EM	0.7581	0.7772	0.7879
CAFE + L-LDA	0.7904	0.8144	0.8247
CAFE	0.8041	0.8238	0.8326
Our method	**0.8624**	**0.8477**	**0.8826**

Table 2: Rand Index of aspect identification.

Table 2 lists the Rand Index of different methods, and our method has the best performance on aspect identification. The improvement on aspect extraction may be mainly due to sense embeddings, which better capture the semantic meanings of features. In addition, our

method does not require the number of aspects, as it automatically stops when the minimum similarity between partitions is large enough.

Since aspect extraction is performed after the user selects a time period, the computational efficiency should be high enough to support interactive exploration. For instance, 731 features are extracted from the reviews of a pair of headphones, and the final number of aspects is 35 after the hierarchical clustering. The computational times are 550 ms and 6669 ms for the hierarchical clustering with and without the k-nearest-neighbor algorithm, i.e., partition generation, respectively. Thus, the efficiency of our method is acceptable for users to interactively explore a controversy.

6.2 Case Studies

We have applied our system to analyze the items in three domains as follows: movies, products, and restaurants.

6.2.1 Characterizing the time-evolving controversy trend

We first describe how our system characterizes the time-evolving controversy trend. We leverage the 2616 reviews on IMDb for *Interstellar*, released in November 2014, as shown in Fig. 3. The sense embedding set is trained from the reviews of 193 movies randomly collected from IMDb. The line chart shows two indexes from October 2014 to March 2016. Both hard and soft controversy indexes are high and their ranges are from 0.6 to 0.8. H reaches its maximum value at 0.8, much higher than S, just after the movie is released. H then decreases over time.

According to the evolution of H, we can assume that the hard controversy possibly occurs just after the movie is released. We first select this time period. The left rating distribution verifies our hypothesis that many reviewers give the lowest rating, and many other reviewers give the highest rating. Then, we select the recent time period. H decreases and its range is between 0.6 and 0.7. The right rating distribution also indicates that the polarization is weakened, as the number of the highest rating is clearly larger than the number of the lowest rating. Previous research has found a similar result where movie audiences are more critical just after a movie is first screened and more likely to post extreme views [18, 21], and then the variance steadily decreases over time [43]. This demonstrates the usefulness of our system on characterizing the time-evolving controversy trend.

We further examine the causes of a controversy. In the left aspect bubble view, most bubbles, such as the *story plot*, *oscars*, and *cooper* aspects, are located near the middle line (the neural sentiment), which indicates that the sentiments of these aspects may be divergent. One main aspect is the *story plot* aspect with many comments. Some comments are "*the story is just incredibly absorbing and really satisfying*", while other comments are "*the basis of the sentimental story does not sit well in the overall story*". We label this aspect as controversial. While the right aspect bubble view indicates that many aspects have a large percentage of positive reviews. The *characters* aspect is located near the middle line with many reviews. We can conclude that reviewers mention about the *story plot* more in the early period, while they talk more about the *characters* in the later period. The *cooper* aspect, nearing the middle line, is also controversial and its aspect burst is represented in Fig. 3(a), where the *matthew mcconaughey* and the *michael caine* aspects have more positive reviews than the *anne hathaway* aspect. The *matt damon* aspect in the right sentiment pie view is very noticeable as the aspect with the most negative sentiments in Fig. 3(b), and this may be because Matt Damon plays a villain in *Interstellar*.

6.2.2 Interpreting sentiment divergences of aspects

In this case study, we demonstrate how our system can effectively interpret sentiment divergences of aspects. The 567 reviews of a pair of Bose headphones on Amazon are shown in Fig. 4. The sense

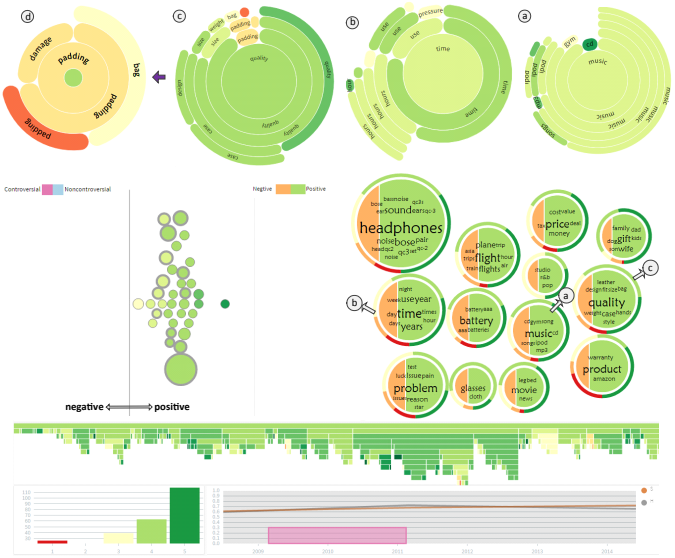

Figure 4: Controversy analysis for a pair of *Bose* headphones on Amazon. The aspect bursts of the *music*, *time*, and *quality* aspects are shown in (a), (b) and (c), respectively. The child aspects of the *padding* aspect is presented in (d).

embedding set is trained from the Amazon Electronics data. In the line chart, both H and S basically remain stable over time, and this is a typical trend for a product. Thus, we select the whole time period. The rating distribution shows that most reviewers give a high rating to the headphones. Moreover, all aspects are located in the right positive side in the aspect bubble view, and this agrees with the rating distribution.

Main aspects are displayed in the sentiment pie view. Reviewers are more interested in the *headphones* aspect, including *sound*, *noise*, and *volume* of a pair of headphones. The *flight* aspect with the features, *train*, *trip*, and *asia*, is about its usefulness when traveling, especially to a remote area. The *glasses* aspect may be strange in the reviews of a pair of headphones. By selecting this aspect, the text view shows that there are many people who care about whether the headphone design is comfortable for people with glasses.

Although all aspects tend to be positive, there are still some negative aspects, yellow and orange colored partitions, as shown in the partition tree view. We further explore sentiment divergences in three aspects, i.e., *quality*, *time*, and *music*. By clicking the *quality* aspect, its aspect burst in Fig. 4(c) shows that its child *padding* aspect is negative. When clicking the *padding* aspect, as can be seen clearly from its child aspects in Fig. 4(d), the negative sentiment of the *padding* aspect is mostly due to the *damage* aspect. The *time* aspect burst in Fig. 4(b) reveals the pressure issue when using the headphones for hours. The *music* aspect burst in Fig. 4(a) indicates that people like to use these headphones to listen to the music with an *ipod* or a mp3 player. The reviewers also mention the *gym* aspect, colored yellow, showing that people often use the headphones at the gym to drown out the noise, but their effect needs to be improved. We label these three aspects as controversial.

Through this case study, we show the usefulness of our system on interpreting the sentiment divergences of aspects. Marketers can identify and address those aspects with problems from aspect-level sentiment analysis, and improve negative and controversial aspects. Customers can rank the aspects according to their personal preferences and choose the product that meets their preferences.

6.2.3 Detecting and interpreting controversy

In this case study, we show how our system detects and interprets controversy through aspect-level sentiment divergence. We analyze

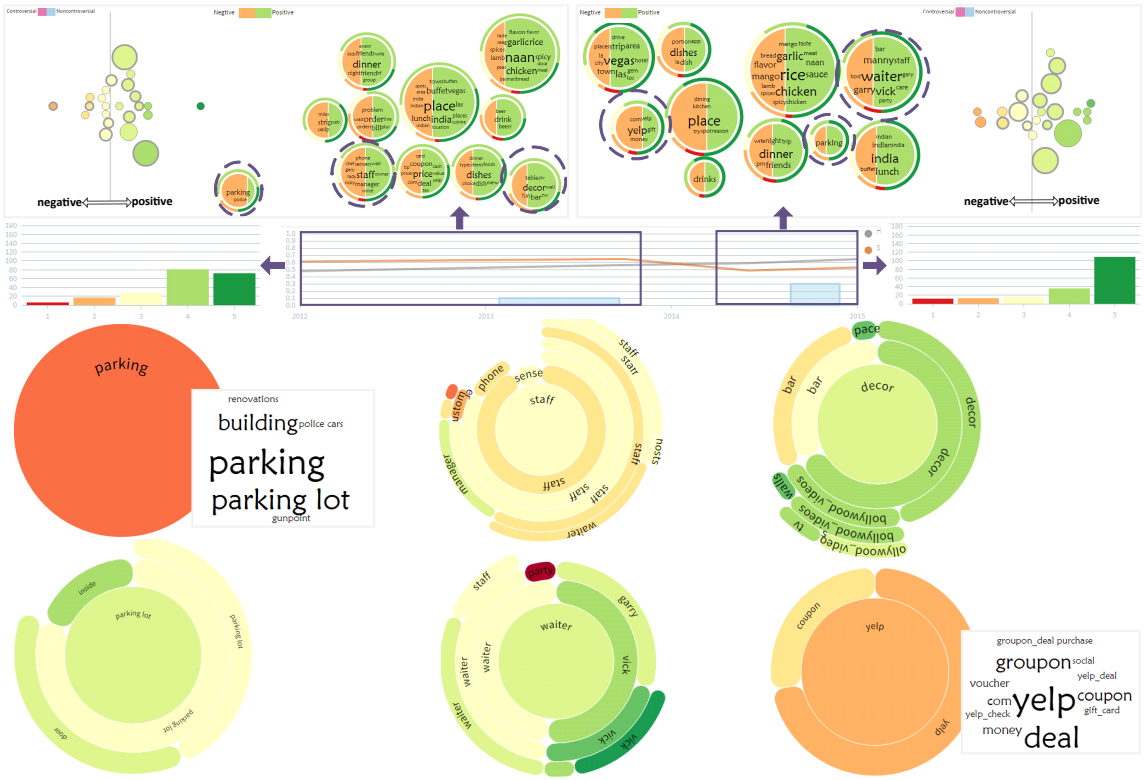

Figure 5: Controversy analysis for an Indian cuisine restaurant on Yelp. The rating distributions of two time periods, before 2014 and after 2014, are shown on the left and right of the line chart, respectively. The upper row of aspect bursts contains three aspect bursts before 2014, and the bottom row contains three aspect bursts after 2014.

the 790 reviews of an Indian cuisine restaurant on Yelp, as shown in Fig. 5. The sense embedding set is trained from the yelp academic dataset with 1,569,264 reviews. The line chart shows that S is higher than H before 2014, while S decreases from 0.7 to 0.6 after 2014, which indicates that the soft controversy has gradually been reduced. When we select two time periods, before 2014 and after 2014, the rating distributions are shown on the left and right of the line chart in Fig. 5, respectively. There are many ratings of 4 stars before 2014, while there are few ratings of 4 stars after 2014. How does this restaurant develop from a good restaurant to a great restaurant?

We further analyze the aspects and their sentiments generated from the reviews in the two time periods. We select some common aspects with different sentiments in the two time periods and distinctive aspects in each time period in the sentiment pie view. Their aspect bursts are provided in the last two rows in Fig. 5.

Generally, there are two cases that can raise the controversy index. One is that many aspects are controversial, and the other is that some aspects are positive and others are negative. Before 2014, the *nann*, *decor*, and *place* aspects tend to be positive, while the *parking* aspect has a large percentage of negative reviews and is located at the leftmost position in the aspect bubble view. Its aspect burst is close to red and its word cloud consists of features, such as *gunpoint*, *police cars*, and *renovations*. This means that people complained a great deal about the parking lot that may have needed major renovation. While after 2014, the *parking* aspect is located on the right of the middle line in the aspect bubble view, and its aspect burst is close to green. After reading related reviews in the text view, we can conclude that this restaurant received a new asphalt parking lot after 2014. In addition, the *decor* aspect only appears before 2014, so we can infer that this restaurant underwent a renovation before 2014.

Similarly, the *staff* aspect is relatively negative before 2014, but the *waiter* aspect, similar to the *staff* aspect, tends to be positive after 2014. As can be seen from the *staff* aspect burst, the staff is associated with phones. The reason for the negative sentiment may be that people complained about the staff using phones, which was proven in the text view when clicking the word *staff* in the word cloud view. The *waiter* aspect burst includes two child aspects, *vick* and *garry*, which are waiters' names who have mostly positive reviews when checking related reviews.

The *yelp* aspect is located at the leftmost position in the aspect bubble view and it only appears after 2014. As shown in its word cloud, we can infer that Yelp coupons become popular, while customers are unhappy with the coupons. After reading related reviews, we find that when customers use the coupon, they are informed of all kinds of restrictions that are not stated on the coupons. This should be improved.

This example demonstrates that our system can effectively detect and interpret a controversy, and gain insights into controversial aspects.

6.3 User Study

We performed a user study to evaluate the usability of our system, especially the visual design, on controversy analysis in web reviews for customers and marketers. This user study involves 30 participants (12 male and 8 female) aged 20 to 30 years from diverse majors (6 undergraduate and 14 graduate students). Nine of them have some knowledge of visualization. In particular, they all have online shopping experience and 7 of them have selling experience.

6.3.1 Study Design

The study consists of three sessions. We started with a 5-minute introduction and demonstration of our system. We then provided two datasets, *Yelp* dataset with the reviews of an Indian and a Japanese cuisine restaurant, and *Amazon* dataset with the reviews of a pair of *Bose* headphones and an *ipad*. We first allowed the participants

to choose one dataset and then four tasks corresponding to design requirements that were given to the participants. For each item in the dataset, they were asked to browse the corresponding website about the item guided by these tasks for 5 minutes, then use our system guided by these tasks for 20 minutes, and finally fill in a questionnaire. An optional 10-minute interview session was conducted, where 10 participants expressed their opinions on the functionality of our system. The study lasted about 1 hour.

We propose four design requirements, controversy evolution characterization, aspect presentation, sentiment divergence visualization, and usability in Section 5. The four tasks that include combinations of these requirements are the following: (1) when a controversy occurs in reviews (R1), (2) what aspects are the causes of the controversy (R3), (3) the sentiment divergences of the aspects that one cares about when deciding where to eat or what to purchase (R2, R3), and (4) what aspects need to be improved or can be considered as the controversial marketing tactics to target the right market (R2, R3). We recorded the task results of each participant and then asked them to fill in a questionnaire regarding the exploration experience. We organized the four design requirements into statement expressions in the questionnaire: (1) the system supports users well in understanding the time-evolving controversy trend (R1), (2) the system provides a good overview of the aspects and the sentiment of the aspects (R2), (3) The system enables users to identify why the sentiment divergences occur in aspects (R3), (4) The system is easy to learn and use (R4). Participants are asked whether they agree with the statement expression. We employed a Likert-type scale from 1 (strongly disagree) to 7 (strongly agree) on all statements.

The results of the questionnaire are the participants' subjective ratings on the system interface. Although we asked the participants to solve tasks, we did not evaluate the participants' performance of these tasks. There is no clearly delineated definition of controversy, which in turn makes the task results difficult to evaluate. We design these tasks to guide the participants to seriously explore the reviews with the system. However, participants are potentially biased toward positive answers that the authors wish to obtain. We need to interpret the results cautiously and it is still valid to compare the ratings to each other. Moreover, the opinions of the optional interview session are very helpful to improve our system.

6.3.2 Study Results

We now discuss the study results as summarized in Fig. 6. We also estimate the hard and soft controversy indexes of the user study.

The average ratings for all design requirements are 5.8 or above, which indicates that our system meets the expectations of the participants. All soft controversy indexes remain almost stable, while the hard controversy indexes are different between design requirements.

For R1 (Controversy evolution characterization), the average rating is 5.8, the lowest of the requirements, and it has the highest hard controversy index of 0.45. This indicates that participants are polarized for the performance of the controversy evolution presentation.

R2 (Aspect presentation) and R3 (Sentiment divergence visualization) receive the high rating of 6.2 and 6.3 with low hard controversy indexes of 0.29 and 0.30, respectively. Participants are satisfied with the visual design for aspects and their sentiment divergences. They are able to analyze sentiment divergences of aspects effectively using our system.

In the questionnaire, R4 (Usability) is divided into two parts, easy to learn and easy to use. The average ratings of these two parts are 5.8 and 5.9 with the hard controversy indexes 0.38 and 0.36, respectively. Most participants report that our system is easy to learn and use, since our design is based on well-known visual representations.

According to the record of the task results, all participants could find when a controversy occurs, the main causes of the controversy from the aspect level effectively, and their interested information. As

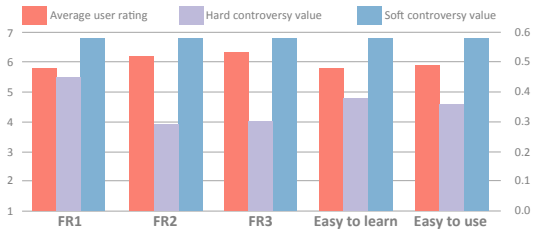

Figure 6: Average user rating, hard controversy index, and soft controversy index for R1 (Controversy evolution characterization), R2 (Aspect presentation), R3 (Sentiment divergence visualization), and R4 (Easy to learn, Easy to use).

for participants' opinions in the optional interview session, various suggestions for improvement were proposed. Two out of 10 participants point out that H and S controversy indexes facilitate users identifying controversy when there are sufficient reviews in each time window and they are not effective when the reviews are sparse. H and S are calculated by adding the next time window when there are too few votes in the time window, which may lead to a decrease in the time effectiveness for the controversy evolution presentation. This could confuse users and cause the hard controversy among users, which is consistent with the questionnaire results of R1. Two out of 10 participants report that it is hard for them to understand the hierarchical clustering process presented in the partition tree view, since they are not familiar with the clustering algorithm. While other participants favor the partition tree view that enables them to examine sentiments of multi-level aspects. One participant mentions that the word cloud would be better for placing features according to their sentiments and another one mentions that labeling an aspect would be more powerful to consider labeled aspects as input to interactively refine controversy analysis results. These will be our future work. Overall, most comments indicate that the system is attractive, useful, and enables them to find new information that is important for them but may be less important for others. One participant points out that the system is very helpful and should be released.

7 Conclusion

In this paper, we have proposed a novel visual analytics system leveraging rating and text analysis to interactively detect and interpret the controversy in web reviews. This framework could also inspire exploration of the datasets in which structured data and unstructured text co-exist. Two hard and soft controversy indexes based on the ratings of reviews are used to characterize the time-evolving controversy trend. We have proposed a new aspect-extraction method based on sense embeddings and hierarchical clustering. A new divergence glyph is designed to present aspects with their sentiments, topics, and rating distributions to facilitate visualization and comparison of sentiment divergences of aspects. Our evaluation, including the aspect extraction experiment, three case studies, and the user study, demonstrated the effectiveness and usefulness of our system.

Although useful and effective, our system still has some limitations, indicated in the user study. For the limitation concerning the sparsity of the reviews, the hard and soft controversy indexes may be ineffective when the reviews are sparse. We plan to incorporate text mining into the controversy quantification to improve the effectiveness of controversy qualification. We also intend to combine H and S to visualize controversy more intuitively. Aside from the product domain, we would like to evaluate our aspect-extraction method in other domains in the future.

Acknowledgments

The authors would like to thank the anonymous reviewers for their valuable comments and thank Yelp Data Challenge for making the data available. This work was partially supported by National Natural Science Foundation of China No. 61472354 and 61672452.

REFERENCES

[1] L. A. Adamic and N. Glance. The political blogosphere and the 2004 us election: divided they blog. In *Proceedings of the 3rd international workshop on Link discovery*, pp. 36–43. ACM, 2005.

[2] L. Akoglu. Quantifying political polarity based on bipartite opinion networks. In *ICWSM*, 2014.

[3] L. Amendola, V. Marra, and M. Quartin. The evolving perception of controversial movies. *Palgrave Communications*, 1, 2015.

[4] W. Bancken, D. Alfarone, and J. Davis. Automatically detecting and rating product aspects from textual customer reviews. In *Proceedings of the 1st International Workshop on Interactions between Data Mining and Natural Language Processing at ECML/PKDD*, pp. 1–16, 2014.

[5] M. Baroni, G. Dinu, and G. Kruszewski. Don't count, predict! a systematic comparison of context-counting vs. context-predicting semantic vectors. In *ACL (1)*, pp. 238–247, 2014.

[6] U. Brandes and J. Lerner. Visual analysis of controversy in user-generated encyclopedias. *Information Visualization*, 7(1):34–48, Mar. 2008. doi: 10.1145/1391107.1391111

[7] I. Brigadir, D. Greene, and P. Cunningham. Analyzing discourse communities with distributional semantic models. In *Proceedings of the ACM Web Science Conference*, p. 27. ACM, 2015.

[8] T. Cai, H. Cai, Y. Zhang, K. Huang, and Z. Xu. Polarized score distributions in music ratings and the emergence of popular artists. In *Science and Information Conference (SAI), 2013*, pp. 472–476. IEEE, 2013.

[9] N. Cao, L. Lu, Y.-R. Lin, F. Wang, and Z. Wen. Socialhelix: visual analysis of sentiment divergence in social media. *Journal of Visualization*, 18(2):221–235, 2015.

[10] L. Chen, J. Martineau, D. Cheng, and A. Sheth. Clustering for simultaneous extraction of aspects and features from reviews. In *NAACL*, pp. 789–799, 2016.

[11] X. Chen, Z. Liu, and M. Sun. A unified model for word sense representation and disambiguation. In *EMNLP*, pp. 1025–1035. Citeseer, 2014.

[12] M. Conover, J. Ratkiewicz, M. R. Francisco, B. Gonçalves, F. Menczer, and A. Flammini. Political polarization on twitter. *ICWSM*, 133:89–96, 2011.

[13] N. Diakopoulos, D. Elgesem, A. Salway, A. Zhang, and K. Hofland. Compare clouds: Visualizing text corpora to compare media frames. In *Proc. of IUI Workshop on Visual Text Analytics*, 2015.

[14] D. Duan, W. Qian, S. Pan, L. Shi, and C. Lin. Visa: a visual sentiment analysis system. In *Proceedings of the 5th International Symposium on Visual Information Communication and Interaction*, pp. 22–28. ACM, 2012.

[15] K. Garimella, G. De Francisci Morales, A. Gionis, and M. Mathioudakis. Quantifying controversy in social media. In *Proceedings of the Ninth ACM International Conference on Web Search and Data Mining*, pp. 33–42. ACM, 2016.

[16] M. Hu, S. Liu, F. Wei, Y. Wu, J. Stasko, and K.-L. Ma. Breaking news on twitter. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 2751–2754. ACM, 2012.

[17] I. Iacobacci, M. T. Pilehvar, and R. Navigli. Sensembed: learning sense embeddings for word and relational similarity. In *Proceedings of ACL*, pp. 95–105, 2015.

[18] N. S. Koh, N. Hu, and E. K. Clemons. Do online reviews reflect a products true perceived quality? an investigation of online movie reviews across cultures. *Electronic Commerce Research and Applications*, 9(5):374–385, 2010.

[19] D. Koutra, P. N. Bennett, and E. Horvitz. Events and controversies: Influences of a shocking news event on information seeking. In *Proceedings of the 24th International Conference on World Wide Web*, pp. 614–624. ACM, 2015.

[20] O. Levy, Y. Goldberg, and I. Dagan. Improving distributional similarity with lessons learned from word embeddings. *Transactions of the Association for Computational Linguistics*, 3:211–225, 2015.

[21] Y. Liu. Word of mouth for movies: Its dynamics and impact on box office revenue. *Journal of marketing*, 70(3):74–89, 2006.

[22] A. Livne, M. P. Simmons, E. Adar, and L. A. Adamic. The party is over here: Structure and content in the 2010 election. *ICWSM*, 11:17–21, 2011.

[23] Y. Lu, M. Steptoe, S. Burke, H. Wang, J.-Y. Tsai, H. Davulcu, D. Montgomery, S. R. Corman, and R. Maciejewski. Exploring evolving media discourse through event cueing. *IEEE transactions on visualization and computer graphics*, 22(1):220–229, 2016.

[24] Y. Mejova, A. X. Zhang, N. Diakopoulos, and C. Castillo. Controversy and sentiment in online news. *arXiv preprint arXiv:1409.8152*, 2014.

[25] T. Mikolov, K. Chen, G. Corrado, and J. Dean. Efficient estimation of word representations in vector space. In *ICLR*, 2013.

[26] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pp. 3111–3119, 2013.

[27] A. Morales, J. Borondo, J. C. Losada, and R. M. Benito. Measuring political polarization: Twitter shows the two sides of venezuela. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 25(3):033114, 2015.

[28] S. Mukherjee and G. Weikum. Leveraging joint interactions for credibility analysis in news communities. In *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*, pp. 353–362. ACM, 2015.

[29] D. Oelke, M. Hao, C. Rohrdantz, D. A. Keim, U. Dayal, L.-E. Haug, and H. Janetzko. Visual opinion analysis of customer feedback data. In *Visual Analytics Science and Technology, 2009. VAST 2009. IEEE Symposium on*, pp. 187–194. IEEE, 2009.

[30] W. M. Rand. Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical association*, 66(336):846–850, 1971.

[31] L. Shi, F. Wei, S. Liu, L. Tan, X. Lian, and M. X. Zhou. Understanding text corpora with multiple facets. In *Visual Analytics Science and Technology (VAST), 2010 IEEE Symposium on*, pp. 99–106. IEEE, 2010.

[32] R. Socher, A. Perelygin, J. Y. Wu, J. Chuang, C. D. Manning, A. Y. Ng, and C. Potts. Recursive deep models for semantic compositionality over a sentiment treebank. In *EMNLP*, vol. 1631, p. 1642, 2013.

[33] Q. Su, X. Xu, H. Guo, Z. Guo, X. Wu, X. Zhang, B. Swen, and Z. Su. Hidden sentiment association in chinese web opinion mining. In *Proceedings of the 17th international conference on World Wide Web*, pp. 959–968. ACM, 2008.

[34] A. Trask, P. Michalak, and J. Liu. sense2vec-a fast and accurate method for word sense disambiguation in neural word embeddings. *arXiv preprint arXiv:1511.06388*, 2015.

[35] H. Wachsmuth, J. Kiesel, and B. Stein. Sentiment flow–a general model of web review argumentation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pp. 601–611, 2015.

[36] Y. Wu, S. Liu, K. Yan, M. Liu, and F. Wu. Opinionflow: Visual analysis of opinion diffusion on social media. *IEEE transactions on visualization and computer graphics*, 20(12):1763–1772, 2014.

[37] Y. Wu, F. Wei, S. Liu, N. Au, W. Cui, H. Zhou, and H. Qu. Opinionseer: interactive visualization of hotel customer feedback. *IEEE transactions on visualization and computer graphics*, 16(6):1109–1118, 2010.

[38] Z. Wu and C. L. Giles. Sense-aware semantic analysis: A multi-prototype word representation model using wikipedia. In *AAAI*, pp. 2188–2194, 2015.

[39] J. Xu, Y. Tao, and H. Lin. Semantic word cloud generation based on word embeddings. In *2016 IEEE Pacific Visualization Symposium (PacificVis)*, pp. 239–243. IEEE, 2016.

[40] P. Xu, Y. Wu, E. Wei, T.-Q. Peng, S. Liu, J. J. Zhu, and H. Qu. Visual analysis of topic competition on social media. *IEEE Transactions on Visualization and Computer Graphics*, 19(12):2012–2021, 2013.

[41] T. Yasseri, A. Spoerri, M. Graham, and J. Kertész. *The most controversial topics in Wikipedia: A multilingual and geographical analysis*. Global Wikipedia: International and Cross-Cultural Issues in Online Collaboration, Fichman P., Hara N., eds., Scarecrow Press, 2014.

[42] Z. Zhai, B. Liu, H. Xu, and P. Jia. Clustering product features for opinion mining. In *Proceedings of the fourth ACM international conference on Web search and data mining*, pp. 347–354. ACM, 2011.

[43] Y. Zhang, T. Lappas, M. Crovella, and E. D. Kolaczyk. Online ratings: Convergence towards a positive perspective? In *2014 ICASSP*, pp. 4788–4792. IEEE, 2014.