A Multimedia Interface for Facilitating Comparisons of Opinions

Giuseppe Carenini and Lucas Rizoli
Department of Computer Science
University of British Columbia
2366 Main Mall, Vancouver, BC, Canada V6T 1Z4
{carenini, lrizoli}@cs.ubc.ca

ABSTRACT

Written opinion on products and other entities can be important to consumers and researchers, but expensive and difficult to analyze. We present a multimedia interface designed to facilitate the analysis of opinions on multiple entities, which could be beneficial to many individuals and organizations. It integrates an information visualization and an intelligent system that selects notable comparisons in the data and summarizes them in text. This system applies a set of statistics for comparing opinions across entities. We conducted a study of our interface with 36 subjects. Subjects liked the visualization overall and our system's selections overlapped with those of subjects more than did the selections of baseline systems. Given the choice, subjects sometimes changed their selections to be more consistent with those of our system. This suggests that system selections were valuable to them.

Author Keywords

opinion mining, automatic summarization, information visualization, evaluative text, user study

ACM Classification Keywords

H.5.1 Information Interfaces and Presentation: Multimedia Information Systems

INTRODUCTION

There is a lot of written opinion on products, services, and other entities available in online reviews, blogs, and the like. In marketing literature, this is known as electronic word-of-mouth (**eWOM**) [22]; in computational linguistics, as **evaluative text**.

The opinions expressed in such text can be of great use to many people and organizations. Consumers use others' opinions to help make decisions as to what they should purchase or support [9, 24]; marketers, designers, and manufacturers can use similar information to study consumer opinion [12] and make forecasts. This valuable information is available, but analyzing it can be very difficult, costly, and time-consuming [13].

There have been many efforts to mine opinions from evaluative text automatically [21]. These have made opinion data available, but the means of understanding and analyzing it effectively are lacking. Tools capable of generating textual summaries and simple graphical representations exist (e.g. [17], [7]), but are not designed to support analysis, particularly the *comparison of opinions on multiple entities*. Such comparisons are key in tasks such as competitive analysis [10], or deciding among alternatives to make a purchase.

In this paper, we present a novel multimedia interface for the analysis of opinion data, particularly the comparison of opinions across two entities. We aim to facilitate the analysis and comparison of opinion data, to allow users to leverage large collections of opinion data and come to actionable conclusions. Our interface includes a visualization of the data as well as an intelligent textual summary of notable comparisons in the data.

The visualization of opinion data represents the first contribution of this paper. Our visual representation was designed by following a task-based approach [30] to facilitate the analysis of opinion data, as well as to simplify comparisons among such data within and across features, as well as across entities.

The second and key contribution of this paper is a content selection strategy that selects notable aspects of the opinion data to be presented textually. Our strategy is based on a set of statistics that we argue can effectively describe the (dis)similarity of opinions on a feature across entities (i.e., on the *Lens* of two or more cameras). In essence, our content selection strategy says that only features that are very (dis)similar with respect to these statistics should be included in the summary.

The third contribution of this paper is a user study, conducted with 36 subjects, to evaluate our visualization as well as our content selection strategy. The usability of the visualization was evaluated by questionnaire, while the content selection strategy was tested by comparing subjects' and system selections from the same sets of opinion data, as well as verifying whether revealing the system selections to the subjects

IUI '09, February 8–11, 2009, Sanibel Island, Florida, USA.

prompted them to revise their initial selections to make them consistent with the system's. The results of this study indicate that subjects found the visualization accurate, easy to learn and read, and satisfying, that subject and system selections overlap, and that when shown the system's selections, subjects often revised their own initial selections to make them more consistent with the system's selections.

We first describe our interface: the visual representation of opinion data, and the system that selects comparisons and summarizes them. We report on our user study and discuss the results. We then review related work in the field of natural language processing on opinion mining and summarization, in information visualization on visualizing opinion data, and in multimedia interfaces on the integration of text and graphics. We conclude by identifying the strengths of our method and areas for future work.

THE MULTIMEDIA INTERFACE

Extraction and organization of opinion data

A description of our method of mining opinions from text is given in [8]. From a corpus of documents expressing opinions on an entity (e.g. user-submitted reviews from Amazon.com of the Canon G3 digital camera), our method applies techniques presented in [16] to return a list of the entity's features on which opinions are expressed (a camera's *Flash*, its *Appearance*, etc.), as well as the opinions themselves. In this way, it is possible to extract sets of features and opinions from a number of corpora, each of which expresses opinions on a different entity (reviews of a Sony camera, of a Nikon camera, etc.).

The polarity—whether an opinion is positive or negative—and strength—the degree of sentiment—of each opinion can also be determined. In our method, three levels of strength are considered, thus the polarity/strength of an opinion can be represented by an integer in $\{-3, -2, -1, +1, +2, +3\}$, where $+3$ is the most positive opinion and -3 the most negative.

The features, and their associated opinions, are then organized according to a common hierarchy. For example, the features *disk capacity*, *storage*, and *memory-card size* can be mapped to a single feature: *Memory*. Semantically similar features from different entities can also be mapped to a single feature (the Sony's *Steady Shot* and the Nikon's *Vibration Reduction* can be mapped to *Image Stabilization*). This hierarchy is user-defined, so it can reduce redundancy as well as reflect a user's needs or interests.

The outcome of this process is a collection of sets opinions on features of a number of entities, as well as a common hierarchy of features across entities (each camera has a *Flash* feature, a *Memory* feature, etc.). We assume this is the input to our system.

Visualization of opinion data

Our interface employs graphics that represent the data in order to make it accessible to users, as well as to aid in their analysis and comparison of the data. We designed our visualization according to the tasks it should support. To do so we created a task model by integrating relevant task taxonomies and frameworks from previous work in information visualization. These task taxonomies describe visual [30], interactive [25], and analytic [2, 1] tasks. These include reading the data accurately, easily characterizing subsets of the data, identifying anomalies, and relating data to support hypotheses. These tasks are general, but common and important to a number of potential users of our system (consumers, market analysts, researchers etc.).

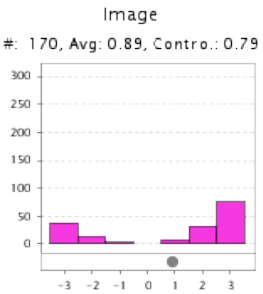

Figure 1. Example of a chart of the opinions on a camera's *Image* feature. The grey dot beneath the bars represents the mean of the opinions. Count (#), mean (Avg.), and controversiality (Contro.) are stated explicitly at the top of the chart.

The primary visual component of our interface is the bar chart representing opinions on a feature (Figure 1). Categories for the polarity/strength of opinions are represented on the horizontal axis, the number of opinions on the vertical axis. Each bar corresponds to a polarity/strength category, its height represents the number of that opinion. Bar charts are an good visual representation of the data because they are clear and familiar, thereby reducing learning times and potential misunderstanding [19].

The mean of the opinions on a feature is represented as a single grey dot plotted along the opinion axis, under the bars. Unlike the bars in the chart above it, to read the mean it is necessary to interpret the opinion axis as numeric, not categorical. The mean can be plotted anywhere between the tick marks representing the most negative and most positive opinions.

It is sometimes useful to have access to the exact values of key descriptive statistics. For this reason we state the count of opinions on a feature, the value of their mean, and the controversiality score above the chart, beneath the feature name.

The mean dot can both support and provide contrast to the information in the bars above it. For example, in Figure 1, the mean is near $+1$, though the actual number of $+1$ opinions in the data is low. This suggests that opinion on *Image* may be split (it is, in a J-shaped distribution). The proximity of the representations of these different, though related, descriptions of the data allow for such understanding to be reached more quickly than if they were apart or not available at a glance.

Camera A $[+1, +2, +2, +2, +2, +3, +3]$
- Lens $[+2]$
  - Aperture Modes $[+2, +2, +3, +3]$
  - Optical Zoom $[-2, -1, -1]$
  - ...
- Flash $[+3, +3, -1]$
- Image $[+3, +3, -1]$
- ...

Camera B $[-1, +1, +1, +1, +2, +3, +3]$
- Lens $[-1, -2]$
  - Aperture Modes $[-1, +1, +1, +1, +2, +2, +3]$
  - Optical Zoom $[-2, -1, -1]$
  - ...
- Flash $[+3, +3, -1, -1, -1]$
- Image $[+3, +2, -1]$
- ...

Figure 2. Partial view of the information extraction and organization process for two products.

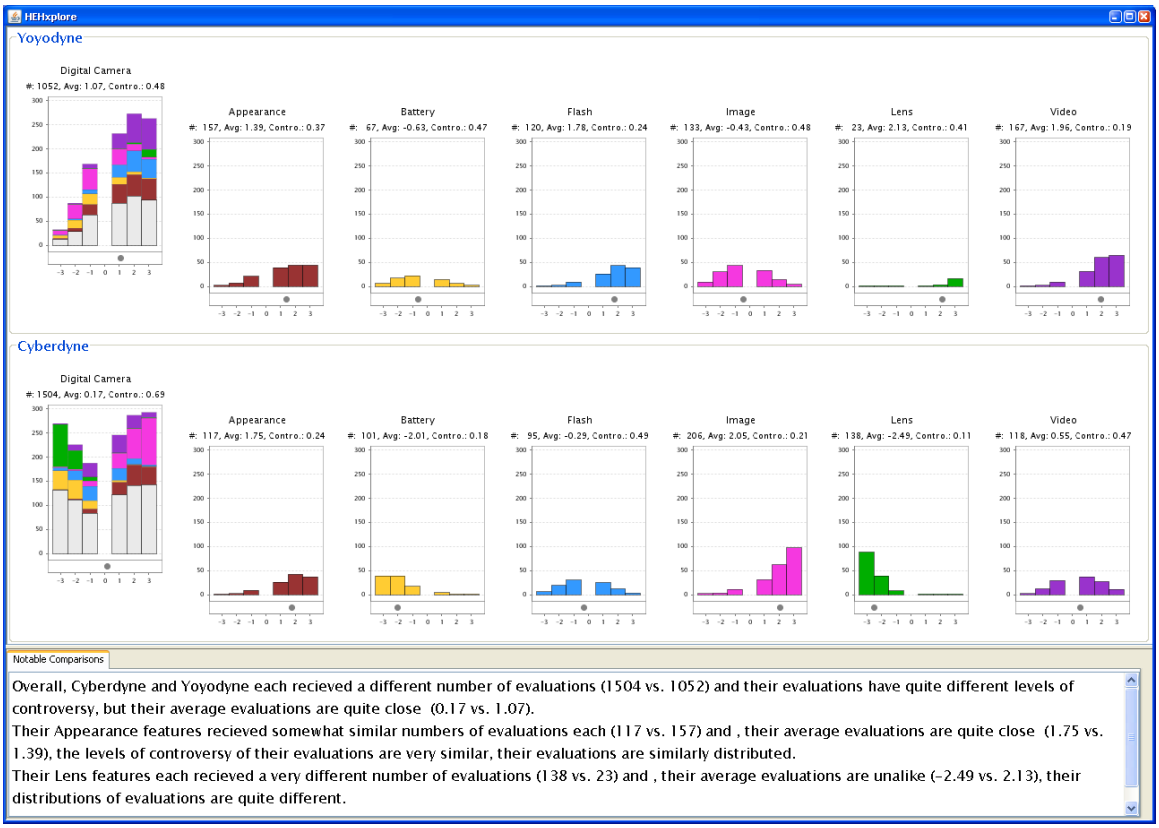

Figure 3. Screenshot of our interface displaying generated opinion data on digital cameras from two fictional manufacturers.

This extends to our representations of opinions on various features: a number of charts are arranged in rows and columns (Figure 4). Comparisons of opinions across features are made easier when many of these charts are arranged neatly and nearly [19, 26]. This follows Tufte's principle of small multiples, that many repeated representations can enforce "comparisons of changes, of the differences among objects, of the scope of alternatives." [27]

Representing the hierarchy of features

The features in our input data are organized in an hierarchy; the charts representing opinions on these features are arranged hierarchically in our interface. Features without children in the hierarchy contain opinions expressed on themselves (e.g. Figure 1). Features that have child features have opinions on themselves and also subsume the opinions on their children. We represent this by stacking the bars representing opinions of child features (Figure 4). Bars of opinions on child features are distinguished both by colour and by their stacking order: opinions expressed on a parent feature explicitly are the bottom-most set of bars in a chart.

Summarization of opinion comparisons

Our interface is designed to support the comparison of opinions across entities. Visualizing opinions on multiple entities allows a user to examine and compare opinions expressed in large corpora quickly and easily, but it may still be difficult and time-consuming to identify important similarities and contrasts of opinion across entities. Also, the nature of these important comparisons may be difficult to convey clearly and succinctly using graphics. To address this potential limitation and to further facilitate comparison, we propose a textual summary of notable comparisons of features across two entities.

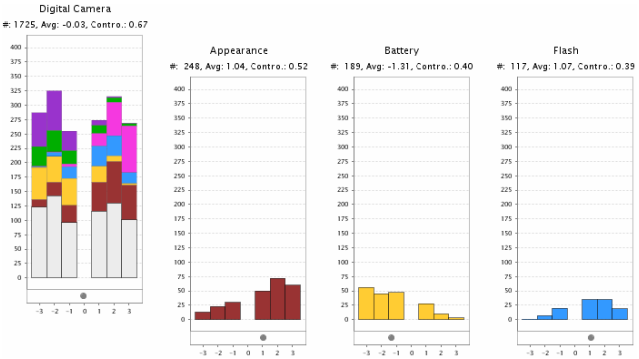

Figure 4. Example of a stacked bar chart, showing the relationship between charts of opinions on *Appearance*, *Battery*, *Flash* and their parent feature, *Digital Camera*. Notice that a) axes' scales are consistent, b) bar colours relate stacked bars to charts of child features, c) explicit opinions on *Digital Camera* are the bottom most bars in the chart (light grey), d) *Digital Camera* contains bars of other child features not shown (e.g. *Image*, *Video*).

We now present the method by which we describe the similarity of feature comparisons, select the most notable comparisons, and summarize them.

We argue that the (dis)similarities of feature comparisons can rely on a set of statistics that are adaptations of statistics previously developed for opinions on *a single entity*, to ones that describe opinions on *a pair of entities*. We first describe these single-entity statistics and then adapt them to the comparison of multiple entities.

Descriptive statistics for a single entity

There are a number of statistics that can be used to describe the opinions on a feature. These include the count, mean opinion, and controversiality (all of which are given above each chart in our visualization). Let $ps(f_a)$ be the set of opinions on the feature f of entity a. We can find the count of opinions, $count(f_a) = |ps(f_a)|$, as well as the mean (or average opinion):

$$mean(f_a) = \frac{1}{|ps(f_a)|} \sum_{ps_k \in ps(f_a)} ps_k$$

There is also the controversiality of the opinions: how split opinions are among positive and negative. An entropy-based controversiality score was introduced in [5]. This score ($contro(f_a)$) is a real value in the range $[0, 1]$. A feature controversiality of 0.0 occurs when all opinions on a feature are of the same polarity; controversiality of 1.0 occurs when they are strong and evenly split between positive and negative (see examples in Figure 5).

Similarities in feature comparisons across two entities

Just as we define statistics for describing opinions on a feature of a single entity, we define statistics for describing the similarity of opinions on a feature across two entities. We call these **aspects** of a comparison. Formally, the similarity of aspects are functions on opinion distributions on feature

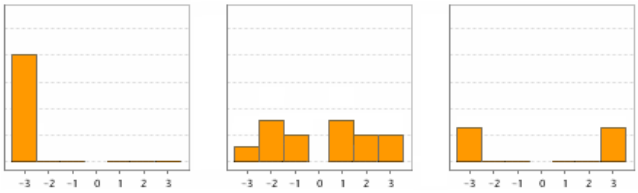

Figure 5. Three distributions with different controversiality scores: 0.0 on the left, 0.612 in the middle, and 1.0 on the right.

f for the pair of entities a and b: f_a and f_b. These functions return values in the range $[0, 1]$, where 1.0 indicates an extreme similarity and 0.0, extreme *dis*similarity.

As when considering opinions on a single entity, it can be important to know how many opinions are expressed on an entity and its features. We define the similarity of the **counts** of opinions on f_a and f_b as the ratio of the $count$ of each.

$$counts(f_a, f_b) = \frac{min(count(f_a), count(f_b))}{max(count(f_a), count(f_b))}$$

We define the similarity of the **means** of opinions as 1 minus the difference between the means proportionate to the maximum possible difference between means.

$$means(f_a, f_b) = 1 - \frac{|mean(f_a) - mean(f_b)|}{2 \times max\_strength}$$

In our system, $max\_strength = 3$, and the greatest possible difference between means is 6 (for the means -3 and $+3$). Note that the equation above is sensitive only to differences in strength between means, not to those of polarity. For example, it returns the same similarity value for the means -2.6 and -0.6 as for the means -1.0 and $+1.0$, though the first two means are both negative and the latter two are of different polarity. This is counterintuitive: the latter case should be less similar than the former. We therefore substitute the initial formulation with the following function when the two means are of different polarity (e.g. $mean(f_a) < 0 < mean(f_b)$) and sufficiently strong ($|mean(f_a)| > 0.5$ and $|mean(f_b)| > 0.5$) in order to capture the dissimilarity of means of different polarity:

$$means(f_a, f_b) = 1 - \sqrt[k]{\frac{|mean(f_a) - mean(f_b)|}{2 \times max\_strength}}$$

with $k > 1$. k was set to 3 in our study based on observing the effect of different values of k on several test cases during development.

The similarity of the aspect of **controversiality** is related to the difference between the controversiality scores of the two distributions of opinions.

$$contros(f_a, f_b) = 1 - |contro(f_a) - contro(f_b)|$$

This equation is not sensitive to whether the two sets of opinions are both (un)controversial (e.g. 0.6 and 1.0) or whether one is controversial and the other uncontroversial (e.g. 0.6 and 0.2). As with means, we exaggerate the difference between controversiality scores when they are different (e.g.

$contro(f_a) < 0.5 < contro(f_b)$) by substituting the equation above with the following:

$$contros(f_a, f_b) = 1 - \sqrt[k]{|contro(f_a) - contro(f_b)|}$$

In addition to the aspects above, we consider differences in the **distribution** of opinions. To do this, we employ Jensen–Shannon divergence (D_{JS}, also known as information radius) [18].

$$\begin{aligned} dists(f_a, f_b) &= 1 - D_{JS}(f_a \| f_b) \\ &= 1 - \left(\tfrac{1}{2} D_{KL}(f_a \| M) + \tfrac{1}{2} D_{KL}(f_b \| M)\right) \end{aligned}$$

where $M = \frac{1}{2}(f_a + f_b)$, and D_{KL} is Kullback–Leibler divergence. Jensen–Shannon divergence is the mean information loss between each distribution from their mean distribution and it is commonly used to measure a kind of distance between two distributions. Unlike D_{KL}, D_{JS} is bounded and symmetric.

Though all the similarity functions above return values between 0.0 and 1.0, this does not mean that they express comparable differences in similarity. For example, two opinion distributions with $means$ of 0.75 are not necessarily as similar as are two distributions with $dists$ of 0.75. To alleviate this problem, we simplified the statistics as follows. Each statistic was discretized into four categories: *very dissimilar* (VD), *dissimilar* (D), *similar* (S), and *very similar* (VS), but the thresholds that define these categories differ among the statistics (Figure 6). These thresholds were arrived at through an iterative evaluation of sample cases made by the authors. These, along with k, are parameters of our selection strategy that could be refined to better match human judgements in the future.

Figure 6. Visualization of thresholds used to discretize values returned by the system's various similarity functions into *very dissimilar* (VD), *dissimilar* (D), *similar* (S), or *very similar* (VS). For example, values of $counts()$ greater than 0.6 and less than 0.7 are *dissimilar*; values of $dists()$ greater than 0.95 are *very similar*.

Content selection

Our content selection strategy always includes an overall comparison of the two entities, which corresponds to comparing the distributions of explicit opinions about the entities combined with opinions about all their features (left-most feature in Figure 4). In addition to this, our strategy selects a subset of the feature comparisons (in our study, up to $\frac{1}{3}$ of the possible feature comparisons) worth mentioning. Then, for each selected feature comparison it determines the aspects that are worth mentioning. We now examine these two selection processes in order.

Selection of feature comparisons to be mentioned

Feature comparisons are first filtered by removing any comparison that covers too few opinions (in our study, 3% or less of the total count) as their statistics are likely not to be very meningful. Formally, a feature f is considered only if:

$$count(f_a) + count(f_b) > \frac{3}{100} \sum_k count(k_a) + count(k_b)$$

After filtering out low-count feature comparisons, the noteworthiness of each feature comparison is assessed by counting the number of very (dis)similar aspects of the comparison. That is,

$$nworthiness(f_a, f_b) = \sum_{g \in G} \begin{cases} 1 & \text{if } g(f_a, f_b) \in \{\text{VD}, \text{VS}\} \\ 0 & \text{otherwise} \end{cases}$$

where $G = \{counts, means, contros, dists\}$. Notice that this assigns greater noteworthiness to feature comparisons with the greatest number of *any kind* of strong (dis)similarity. For example, a comparison of *Lens* that has very similar $means$ and $contros$ has a $nworthiness$ of 2, and it is considered as noteworthy as a comparison of *Battery* that has very similar $counts$ but very *dis*similar $contros$.

Feature comparisons are ranked by $nworthiness$; comparisons are selected for mention in the summary until either $\frac{1}{3}$ of the features in the hierarchy have been selected, or the next most highest ranked comparison's $nworthiness$ is 0. When necessary, ties are broken by selecting the feature comparison with the most extreme $counts$ (that is, the comparison $y = \max_x |counts(f_x) - 0.5|$). This tie breaking strategy is justified by the assumption that $counts$ is the most critical aspect in a comparison.

Selection of comparison aspects to be mentioned

Whenever a feature comparison is included in the summary a statement on its $counts$ and a statement on its $means$ are always included as these two aspects are considered important in any comparison. A statement on $contros$ is included when it is very (dis)similar; likewise, $dists$.

Each feature comparison is summarized in a single sentence. Rhetorically, the statement on $counts$ is presented as the main claim. The statements on the other aspects are presented as **contrast or support for the main claim**, depending on whether they are consistent, with respect to similarity, with the statement on $counts$. For example, the overall comparison of Yoyodyne and Cyberdyne cameras (the feature *Digital Camera* in Figure 3) has dissimilar $counts$, very similar $means$, and very dissimilar $contros$. The statement on $contros$ supports the statement on $counts$ (they are both dissimilar); the statement on $means$ contrasts with the statement on $counts$. The sentence presented in the summary is realized using simple sentence templates.

USER STUDY

Goals

The primary goal of our study was to evaluate the quality of our system's selections of noteworthy comparisons by, firstly, comparing system and user selections, and secondly, finding whether and when users believe the system's selections are good. Our secondary goal was to ascertain the us-

ability of our visualization and to discover, more generally, how users interpret such opinion data.

To achieve these goals, we conducted a study in which we collected what human subjects select as the most noteworthy feature comparisons in a set of opinions, both before and after seeing selections made by our system. Subjects were asked to justify each of their selections by, first, noting whether they had selected a comparison because the opinions were either *similar*, *dissimilar*, or *notable for another reason*, and second, by writing a brief explanation. In addition to this, subjects classified each of our system's selections as *good* or *poor*. Lastly, subjects completed a questionnaire in which they rated the usability of our visualization.

Scenario

To encourage subjects to pay attention to the data as well as to make the task easy to understand, we developed a fictional scenario within which to present opinion data, our visualization, and selection strategy. Subjects were told that an unspecified camera manufacturer is conducting an analysis of the newest digital cameras released by its competitors. This company has hired the subject to analyze the opinions on pairs of digital cameras, and to identify interesting differences and similarities. Subjects were also told that they would be asked to double-check another analysts' work (in truth, the selections made by our summarizer).

Data generation

To our knowledge, there is no available corpus of evaluative text annotated with features in a hierarchy large and varied enough to serve as a basis on which to evaluate our interface. As such, we generated data that mimics what opinion data we do have [15], is in keeping with the study scenario, and is sufficiently varied to evaluate our interface.

We would like to evaluate our interface on the entire space of possible opinion data. This is, however, not practical. Instead, we generated a set of data that we believe represents the space of possible opinion data insofar as it alters the summaries generated by our system. Thus, we identified a number of feature comparison **types** (Table 1), each a set of constraints on the similarity of the aspects of a feature comparison. These constraints are the allowable categories of (dis)similarity—very dissimilar (VD), dissimilar (D), similar (S), or very similar (VS)—an aspect can take. These constraints cause each type to be mentioned with a different configuration of aspects as support or contrast.

For example, the constraints of configuration type 0 are met only when all aspects of a comparison are either similar or dissimilar (such a comparison is not noteworthy, and therefore will not be selected for mention in a summary). Notice that two configuration types between 5 and 6 can be specified, but their constraints are not met in practice[1].

[1]This is because $means$, $contros$, and $dists$ are related in such a way that no two of these aspects can be very similar while the third is dissimilar.

Type	$\|S\|$	$\|C\|$	$counts$	$means$	$contros$	$dists$
0	0	0	D∨S	D∨S	D∨S	D∨S
1	1	0	VS	S∨VS	D∨S	D∨S
2	0	1	S	VD	D∨S	D∨S
3	1	1	D	VD	VS	D∨S
4	2	0	VD	D∨VD	VS	D∨S
5	0	2	VD	S∨VS	VS	D∨S
	1	2	*Does not occur*			
	2	1	*Does not occur*			
6	3	0	VS	S∨VS	VS	VS
7	0	3	VD	S∨VS	VS	VS
?	?	?	VD∨D∨S∨VS			
M	?	?	*At least one* VD∨VS			

Table 1. The comparison types, the number of aspects mentioned in support ($|S|$) and contrast ($|C|$), and the constraints on the (dis)similarity of the aspects of a comparison.

Using these types, we identified a set of **summary cases** we believe to be representative of the larger space of possible summaries (Table 2). Each summary case uses types to constrain the data generated for one, two, and all other feature comparisons, as well as the overall comparison (comparison of the top-most feature in the hierarchy: *Digital Camera*, in our study). For example, case 0 uses type ? overall (which does not constrain how (dis)similar the opinions on the two entities overall must be) but specifies that all other comparisons must be of sentence type 0 (which constrains feature comparisons so that none are noteworthy). Case 15 differs in that it specifies that one comparison must satisfy the constraints of type 1, another must satisfy type M (that is, have at least one very (dis)similar aspect), and all other comparisons must be type 0.

For each of these cases, we generated data for two entities sharing a simple, two-level hierarchy of features. These features are *Appearance*, *Battery*, *Flash*, *Image*, *Lens*, *Software*, and *Video*; all of which are child features of *Digital Camera*.[2]

Subjects

36 subjects, 24 females, 12 males, aged 19–43 (median 23), participated in the study. Subjects were university students or graduates. They were recruited through an online subject pool. Each was paid $10 to participate.

Materials

All study materials were provided to subjects on paper. Charts were printed in colour on sheets of 8.5×11 inch paper, as were quizzes and primers.

Procedure

Subject sessions were designed to take no longer than an hour. Subjects were first briefed on the various parts of the session. They were then given time to read a six-page primer which introduced the scenario, explained the nature of the data, the charts, as well as the means and controversiality.

[2]Initially, we generated data for $8+1$ features, but we later reduced the number of features to $6+1$ after our initial pilot studies.

Case	Overall	C1	C2	C…
0	?	0	0	0
1	1	?	?	?
2	2	?	?	?
3	3	?	?	?
4	4	?	?	?
5	5	?	?	?
6	6	?	?	?
7	7	?	?	?
8	?	1	0	0
9	?	2	0	0
10	?	3	0	0
11	?	4	0	0
12	?	5	0	0
13	?	6	0	0
14	?	7	0	0
15	?	1	M	0
16	?	2	M	0
17	?	3	M	0
18	?	4	M	0
19	?	5	M	0
20	?	6	M	0
21	?	7	M	0

Table 2. The summary cases generated as example data for the study. Each specifies a configuration type for opinions overall, one comparison (C1), another comparison (C2), and all other possible comparisons (C…).

Once finished, they were given a brief quiz that asked them to rank opinions on three different charts according to count, mean, and controversiality. These charts had non-specific feature names and no explicit values for the statistics. Subjects were allowed to refer back to the primer while completing the quiz. The experimenter checked their answers, prompting them to reconsider them where they were incorrect. Subjects did not continue until they had the correct answers.

After completing the quiz, subjects were given a sheet with charts representing opinions of two digital cameras (similar to the top portion of the window in Figure 3) and a response sheet. The first phase of their response was to classify the opinions on the sheets as similar and dissimilar, select up to two notable features, and to give reasons for their selections. Once this phase was complete, they were given a second copy of the sheet of charts with the features selected by our system circled. The names of very (dis)similar aspects of the selected comparisons were also listed on the page, and whether they were similar or dissimilar.

In the second phase, subjects responded by classifying each of the system's selections as *good* or *poor*, their reasons for classifying them as they did, and were given a second opportunity to select up to two notable features. Subjects were not told that the selections were made by a computer system, only that they were "evaluating another analysts' selections."

Subjects were given as many sheets to respond to as could be done in approximately forty minutes. Ten minutes were allotted at the end of each session for subjects to complete the questionnaire.

Method

To evaluate the performance of our content selection strategy against the gold standard supplied by our study subjects, we calculated precision and recall of the system as well as the F-measure.

To determine subjects' perceptions of the usability of our visualization, we included in the final questionnaire a series of statements, and asked subjects to rate how strongly they agree or disagree with each statement. A number of these statements relate to Nielsen's quality components of usablity: learnability, efficiency, memorability, error, and overall satisfaction [20]. They are statements such as "It is easy to learn to read the charts" (learnability), and "I am confident that I read the charts correctly" (error).

Baseline systems

In order to set a baseline of performance, we found the expected performance of two simple alternative selection systems: a **naïve** system which selects 0–2 feature comparisons to mention randomly, and a **semi-informed** system. The semi-informed system is as likely to select 1, 2, or no comparisons as did the subjects in our study. Though it is likely to select the same number of comparisons, the comparisons it selects are picked randomly.

More formally, we can say that the probabilities of these systems selecting a certain number of comparisons are

$$\begin{aligned} \forall x,\ \ & Pr(Size = x | System = \text{naïve}) = \tfrac{1}{3} \\ & Pr(Size = x | System = \text{semi-info.}) \\ & = Pr(Size = x | System = \text{subjects}) \end{aligned}$$

Since we consider 6 selectable features in our study, for both systems, the probability of selecting feature comparison y is

$$\begin{aligned} Pr(Select = y | Size = 0) &= 0 \\ Pr(Select = y | Size = 1) &= \tfrac{1}{6} \\ Pr(Select = y | Size = 2) &= \tfrac{1}{6} + \tfrac{1}{5} \end{aligned}$$

By multiplying the probability of a system selecting features that overlap with selections made by subjects, we find the expected performance of each system (see Table 3). These are the baseline scores which our system must match or beat in order to be considered successful.

RESULTS

Subjects took approximately 10–15 minutes to finish reading and re-reading the primer. Some managed to respond to only a single sheet of charts, while others completed six in the same time. Though subjects were not carefully timed, the experimenter did notice that subjects typically responded faster to later sheets than they did to earlier ones.

Selection agreement

On average, over 98 sets of selections, selections made by our system agree with subject selections better than those we

System	Precision	Recall	F-measure
our system	**0.408**	**0.372**	**0.379**
naïve	0.209	0.168	0.186
semi-info.	0.305	0.305	0.305

Table 3. Mean precision, recall, and F-measure of our system, as well as the expected performance of the naïve and semi-informed alternative selection systems.

could expect from the baseline systems (Table 3): showing a 24% improvement over the semi-informed system, 104% over the naïve.

After seeing our system's selections, subjects selected the same features as our system more often than they did before seeing our system's choices (mean precision = 0.500, SD = 0.419; mean recall = 0.449, SD = 0.390; mean F-measure = 0.454, SD = 0.380). This is a change of roughly 20% in all measures. This change is statistically significant according to two-tailed paired t-tests (precision $t(97) = 2.84$, $p < 0.01$; recall $t(97) = 3.13$, $p < 0.01$).

Subjects, on average, rated 60% of system selections as *good* (SD = 0.418). This means that subjects were more likely rate system selections as *good* than they were to select them in their final selections. Subjects tended to go with their initial selections in the end (mean precision = 0.806, SD = 0.301; mean recall = 0.801, SD = 0.302; mean F-measure = 0.799, SD = 0.298).

Usability of visualization

Subjects rated each of the statements related to Nielsen's quality components (learnability, efficiency, memorability, error, and overall satisfaction), as well as to whether the charts were cluttered. The results are given in Table 4.

Charts are...	SD	D	N	A	SA	NR
Learnable		2	6	13	**15**	
Efficient		3	4	**20**	9	
Memorable		3	5	**17**	11	
Read correctly		1	6	**19**	10	
Satisfying	1	2	5	**22**	5	1
Cluttered		**18**	12	3	1	2

Table 4. Subjects' responses to statements related to components of usability. Subjects could strongly disagree (SD), disagree (D), agree (A), or strongly agree (SA) with a statement, or remain neutral (N), or not respond (NR). The most frequent response for each component is in boldface.

Discussion

Our system for selecting comparisons performs better than the baseline systems. Our system's selected comparisons were more alike those made by subjects; and the improvement over the naïve and semi-informed baseline systems is markable. However, the scores achieved by our system are not particularly high. This suggests that our system could yet well be improved, perhaps by tweaking the thresholds for categorizing comparison aspects.

Interestingly, *subjects often believed that our system made good selections.* It is possible that many cases in which subjects classified the system's selections as *good*, they believed them to be—but not as good as those they made themselves. Perhaps subjects are charitable when "double-checking another analysts' selections," but remain convinced of their initial selections.

Subjects were not always as certain of their initial selections after seeing those made by our system. *In some cases, subjects were convinced by the selections made by our system* and chose to include comparisons selected by the system instead of those they made themselves. This suggests that the selections made by our system were, in some cases, valuable and different from those that subjects were able to make from using the visualization alone.

From questionnaire responses, we find that *subjects consider our visualization to be usable.* The majority of subjects responded positively to the visualizations in general, though the experimenter's observations and subjects' comments made post-session and written in the questionnaires suggest there were differences in how quickly and confidently they read them.

RELATED WORK

Opinion mining and summarization

There has been a substantial amount of work on opinion mining [21]. Extraction systems such as and OPINE [23] are increasingly capable of correctly identifying evaluations in natural language text. Our own system relies on opinion mining detailed in [8]. Many of these systems not only extract opinions from text, but organize and summarize them. The form of these summaries varies, some being lists of pros and cons ([16]) or extractive ([6]), others, generated arguments ([5]). What they have in common is that, unlike our summarization system, they summarize opinions on a single entity.

Visualization of opinion data

There are a number of visualizations of opinions on a single entity. OpinionReader [11] arranges labels in a scatter plot, where the horizontal axis represents the combined polarity and strength of opinion, and the vertical axis, the frequency of those opinions. The purpose of OpinionReader is to summarize the expressed pros and cons of an entity or topic. Carenini, Ng, & Pauls [7] summarize opinions on a single entity by visualizing them in treemaps. The data visualized are different from those used by OpinionReader: opinions of varying strength and polarity on a hierarchy of features. These data are extracted by the same method we assume for our system. Though our system builds on their work, it differs in its visual representation, which does not require significant interaction in order to observe the distribution of opinions on a feature.

Opinion Observer [17] presents opinions on multiple entities. It displays the number of positive and negative evaluations as bars extending from a baseline, coding entity by colour. This allows for side-by-side visual comparison of

opinions on the same feature. This eases the comparison of opinions on a single feature across two or more entities. The data visualized by Opinion Observer is different from ours in that it does not include the strength of opinions, nor does it arrange features into a hierarchy. Opinion Observer also differs from our interface in that it does not attempt to identify and describe notable comparisons, nor does not allow multiple levels of analysis (single feature and the entity overall) to be visualized simultaneously.

Though not a visualization of evaluations extracted from evaluative text, SurveyVisualizer [4] is designed to display similar opinion data. It does so using a number of parallel-coordinate plots arranged in a hierarchy. Data from previous years' survey results—or, conceivably, of opinions on multiple entities—are represented by different lines, differentiated from each other by colour and thickness. Though the source and type the data visualized in SurveyVisualizer (surveys) is different from that of our system, we did attempt to adapt its hierarchical parallel-coordinates to our tasks and data [29]. We did not pursue this visual representation for various reasons, among them that the representation became cluttered, was not as accessible as bar charts, and that the incomplete nature of the mined opinion data made it difficult to maintain a consistent visual and mental representation of opinions on an entity.

ValueCharts+ [3] display the calculated evaluations on each entity according to valuation functions on their features. Despite being designed for a different set of tasks (preferential choice) many of the goals of ValueCharts+ are like those of our system: to facilitate the comparison and selection of entities based on evaluations of their features. The visual representation is also similar, a variation on stacked bar charts.

Multimedia interfaces

Though visualizations can represent large amounts of data in meaningful ways, it can be important to support them with text [28]. Multimedia interfaces have combined graphics and text in complementary ways, taking advantage of the strengths of one medium and compensating for the weaknesses of the other (e.g. [14]). Our interface is an extension earlier work on multimedia presentations of opinions mined from text [7]. This earlier interface presents the user both with a treemap visualization of opinions mined from a corpus of reviews, as well as a summary of those reviews. This summary contains links to the source text from which the summary's sentences are extracted. Our interface is similar in that it is also a multimedia interface, a complement of text and data graphics. It differs in that it summarizes feature comparisons across entities, rather than opinions on features of a single entity. Also, the sentences in our summaries are generated, not extracted from the same corpus of text as the opinion.

CONCLUSIONS

We have detailed a multimedia interface for facilitating the comparison of opinions on two entities. This interface includes two complementary presentations of opinion data: a visualization of the opinions on features of multiple entities, and a textual summary of the most noteworthy comparisons of opinions on features across two entities. We described the motivation for this interface, the design of the visualization, and the methods by which comparisons are ranked, selected, and summarized. The results of our user study show that our visualization is usable, and that our summarization system performs more like humans than do two baseline systems. The results also show that subjects often would reconsider their own analysis when shown that of our system, changing their conclusions about the data in order to be more in line with that of our system. While there is room to improve our interface, it does visualize opinion data in a useable way, and it selects feature comparisons that, in cases, humans find valuable and did not notice from inspecting the visualization alone.

Future Work

Overall, the results of our user study are encouraging. We hope to analyze the data collected in our user study in more detail. An examination of subjects' written reasoning for selecting feature comparisons seems likely to produce insight as to why system's selections differ from subjects', and, more generally, how subjects interpreted the data and used it to justify their selections. It may also be important to study the extent to which subjects agreed with each other. We would also like to evaluate our visualization method more precisely, and the benefits of interacting with it on a computer rather than in a static presentation. Lastly, it would be interesting to see if the data collected from subjects could be used to train a machine learning algorithm to select notable feature comparisons.

ACKNOWLEDGEMENTS

We thank Raymond Ng, Jackie CK Cheung, and Ivan Zhao for their help and suggestions. We also thank our reviewers for their constructive comments.

REFERENCES

1. R. Amar, J. Eagan, and J. Stasko. Low-level components of analytic activity in information visualization. In *INFOVIS '05: Proceedings of the 2005 IEEE Symposium on Information Visualization*, page 15. IEEE Computer Society, 2005.

2. R. Amar and J. Stasko. A knowledge task-based framework for design and evaluation of information visualizations. In *INFOVIS '04: Proceedings of the 2004 IEEE Symposium on Information Visualization*, pages 143–150. IEEE Computer Society, 2004.

3. J. Bautista and G. Carenini. An integrated task-based framework for the design and evaluation of visualizations to support preferential choice. In *AVI '06: Proceedings of the 2006 Working Conference on Advanced Visual Interfaces*, pages 217–224. ACM, 2006.

4. D. Brodbeck and L. Girardin. Visualization of large-scale customer satisfaction surveys using a parallel-coordinate tree. In *INFOVIS '03: Proceedings of the 2003 IEEE Symposium on Information*

Visualization, pages 197–201. IEEE Computer Society, 2003.

5. G. Carenini and J. C. K. Cheung. Extractive vs. NLG-based abstractive summarization of evaluative text: The effect of corpus controversiality. In *INLG '08: Proceedings of the 57th International Natural Language Generation Conference*. ACL, 2008.
6. G. Carenini, R. Ng, and A. Pauls. Multi-document summarization of evaluative text. In *EACL '06: Proceedings of the 11th Conference of the European Chapter of the ACL*, 2006.
7. G. Carenini, R. T. Ng, and A. Pauls. Interactive multimedia summaries of evaluative text. In *IUI '06: Proceedings of the 11th International Conference on Intelligent User Interfaces*, pages 124–131. ACM, 2006.
8. G. Carenini, R. T. Ng, and E. Zwart. Extracting knowledge from evaluative text. In *K-CAP '05: Proceedings of the 3rd International Conference on Knowledge Capture*, pages 11–18. ACM, 2005.
9. J. Chevalier and D. Mayzlin. The effect of word of mouth on sales: Online book reviews. Working Paper, 2003.
10. C. Dellarocas, N. Awad, and X. Zhang. Exploring the value of online reviews to organizations: Implications for revenue forecasting and planning. In *Proceedings of the 24th International Conference on Information Systems*, 2004.
11. A. Fujii and T. Ishikawa. A system for summarizing and visualizing arguments in subjective documents: Toward supporting decision making. In *COLING-ACL '06: Proceedings of the Workshop on Sentiment and Subjectivity in Text*, pages 15–22. ACL, 2006.
12. A. Ghose. *The Economic Impact of User-Generated and Firm-Published Online Content: Directions for Advancing the Frontiers in Electronic Commerce Research*. Wiley & Sons, 2007. Online chapter.
13. D. Godes and D. Mayzlin. Using online conversations to study word-of-mouth communication. *Marketing Science*, 23(4):545–560, 2004.
14. C. Hallett. Multi-modal presentation of medical histories. In *IUI '08: Proceedings of the 13th international conference on Intelligent User Interfaces*, pages 80–89, New York, NY, USA, 2008. ACM.
15. M. Hu and B. Liu. Feature based summary of customer reviews dataset. `http://www.cs.uic.edu/~liub/FBS/sentiment-analysis.html`, 2004.
16. M. Hu and B. Liu. Mining and summarizing customer reviews. In *KDD '04: Proceedings of the 2004 ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 168–177. ACM, 2004.
17. B. Liu, M. Hu, and J. Cheng. Opinion observer: Analyzing and comparing opinions on the web. In *WWW '05: Proceedings of the 14th International World Wide Web Conference*, pages 342–351. ACM, 2005.
18. C. Manning and H. Schütze. *Foundations of statistical natural language processing*. MIT Press Cambridge, MA, USA, 1999.
19. P. McLachlan, T. Munzner, E. Koutsofios, and S. North. Liverac: Interactive visual exploration of system management time-series data. In *CHI '08: Proceedings of the 26th Annual SIGCHI Conference on Human Factors in Computing Systems*, pages 1483–1492. ACM, 2008.
20. J. Nielsen. *Usability Engineering*. Morgan Kaufmann, 1993.
21. B. Pang and L. Lee. Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*, 2(1-2):1–135, 2008.
22. I. Pollach. Electronic word of mouth: A genre analysis of product reviews on consumer opinion web sites. In *HICSS '06: Proceedings of the 39th Annual Hawaii International Conference on System Sciences*, volume 3. IEEE Computer Society, 2006.
23. A.-M. Popescu and O. Etzioni. Extracting product features and opinions from reviews. In *HLT-EMNLP '05: Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing*, pages 339–346. ACL, 2005.
24. S. Senecal and J. Nantel. The influence of online product recommendations on consumers' online choices. *Journal of Retailing*, 80(2):159–169, 2004.
25. B. Shneiderman. The eyes have it: a task by data type taxonomy for informationvisualizations. In *VL '96: Proceedings of the 1996 IEEE Symposium on Visual Languages*, pages 336–343. IEEE Computer Society, 1996.
26. H. Siirtola. Interaction with the reorderable matrix. In *IV '99: Proceedings of the 1999 IEEE International Conference on Information Visualization*, pages 272–277. IEEE Computer Society, 1999.
27. E. R. Tufte. *Envisioning Information*. Graphics Press, 1990.
28. E. R. Tufte. *Visual Explanations: Images and Quantities, Evidence and Narrative*. Graphics Press, 1997.
29. I. Zhao, G. Carenini, and L. Rizoli. Visualizing feature-based customer review summarization system using p-node tree. Undergrad report, 2008.
30. M. Zhou and S. Feiner. Visual task characterization for automated visual discourse synthesis. In *CHI '98: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 392–399. ACM, 1998.