Review Spotlight: A User Interface for Summarizing User-generated Reviews Using Adjective-Noun Word Pairs

Koji Yatani, Michael Novati, Andrew Trusty, and Khai N. Truong
Department of Computer Science
University of Toronto
Toronto, ON M5S 3G4, Canada
{koji, atrusty, khai}@dgp.toronto.edu, michael@novati.ca

ABSTRACT

Many people read online reviews written by other users to learn more about a product or venue. However, the overwhelming amount of user-generated reviews and variance in length, detail and quality across the reviews make it difficult to glean useful information. In this paper, we present the iterative design of our system, called Review Spotlight. It provides a brief overview of reviews using adjective-noun word pairs, and allows the user to quickly explore the reviews in greater detail. Through a laboratory user study which required participants to perform decision making tasks, we showed that participants could form detailed impressions about restaurants and decide between two options significantly faster with Review Spotlight than with traditional review webpages.

Author Keywords

Summarization, user interface, user-generated reviews, natural language processing, word pairs

ACM Classification Keywords

H5.2. Information interfaces and presentation (e.g., HCI): User Interfaces.

General Terms

Human Factors

INTRODUCTION

Online review websites offer users a wealth of perspectives about products that can be purchased or locations that can be visited (which we will refer to as reviewed entities). However, the number of reviews for any given entity often extends well beyond what users can read quickly. User-generated reviews also vary greatly in length, detail and focus in comparison to ones written by professional editors. These issues make it difficult for users to quickly and easily glean useful details about the entity, as described in the following way by a participant in our study:

> *The problem with all these reviews is they put a lot of really useless information there. For example, this guy included a dialogue he had with a waitress... [That] makes it difficult when you actually try to quickly find something.*

There are several ways to provide a brief overview of user-generated reviews. For example, often reviewers are asked to rate their overall impression (usually from 1 to 5) of an entity. Although this rating gives the reader a quick understanding of how much the reviewer liked or disliked the entity, it does not offer information about why that rating was given. Several websites also allow readers to rate the usefulness of posted reviews, and order reviews by this usefulness rating. These features allow the user to browse reviews in a way that differs from needing to read them all. However, browsing in this manner may result in the user missing information that she could find to be important to her (*e.g.*, comments in recent reviews).

In this paper, we describe our iterative design of Review Spotlight—a system which provides a quick overview of user-generated reviews. Figure 1 shows the concept of Review Spotlight. The system displays a list of adjective and noun word pairs that appeared most frequently in the review text. It also performs a sentiment analysis on the extracted word pairs and uses different font colors to represent the level of positivity or negativity of each word

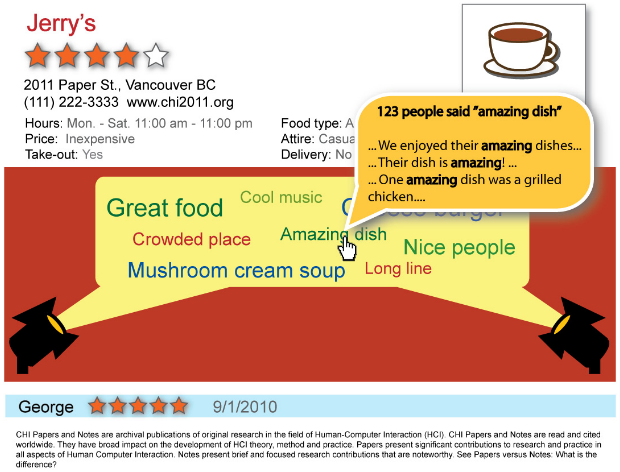

Jerry's
2011 Paper St., Vancouver BC
(111) 222-3333 www.chi2011.org
Hours: Mon. - Sat. 11:00 am - 11:00 pm
Price: Inexpensive
Take-out: Yes
Food type: A
Attire: Casua
Delivery: No
123 people said "amazing dish"
...We enjoyed their amazing dishes...
...Their dish is amazing! ...
...One amazing dish was a grilled chicken....
Great food
Cool music
Crowded place
Amazing dish
Nice people
Mushroom cream soup
Long line
George
9/1/2010
CHI Papers and Notes are archival publications of original research in the field of Human-Computer Interaction (HCI). CHI Papers and Notes are read and cited worldwide. They have broad impact on the development of HCI theory, method and practice. Papers present significant contributions to research and practice in all aspects of Human Computer Interaction. Notes present brief and focused research contributions that are noteworthy. See Papers versus Notes: What is the difference?

Figure 1. The Review Spotlight concept. Review Spotlight can be integrated into any review webpage, and shows some of the most frequently-mentioned word pairs in the reviews.

CHI 2011, May 7–12, 2011, Vancouver, BC, Canada.
Copyright 2011 ACM 978-1-4503-0267-8/11/05....$10.00.

pair. Review Spotlight also allows the user to click on a word pair to see additional contexts in which it is used in the original review text. The user interface does not arrange word pairs in a specific order so that the user can serendipitously acquire information even from word pairs with small fonts. Through our user study, we found that while using Review Spotlight, participants could form detailed impressions about reviewed entities and decide between two options significantly faster than when reading traditional review webpages.

RELATED WORK

The Review Spotlight interface is based on a tag cloud, a visualization using a set of words (or "tags"). The size and color of each tag are associated with the importance or relevance in the original text. The use of natural language processing can improve the efficiency and accuracy of extracting the tags that best represent the original document. However, this work primarily examines the user requirements for an interface that displays a summary of user-generated reviews, and not the process of efficiency of generating tag cloud. Thus, we mainly review research focused on interfaces for review summarization and the effects of a tag cloud on different tasks.

User Interfaces for User Review Summarization

Although user review summarization has been investigated well in computational linguistics [11], user interfaces employing it have not been studied extensively. One approach for user review summarization is to collect opinions on different features in an entity (*e.g.*, food or service in a restaurant review). Liu *et al*. developed a system to visualize how many positive or negative reviews were posted about features of an entity using bar graphs, but the system was not evaluated from the user interface perspective [17]. Carenini *et al*. used a Treemap visualization to show the information extracted from user-generated reviews organized in a tree structure based on the features [3]. However, their user study showed that the participants were often confused by the Treemap visualization, and preferred text-based summarization. They also developed a system similar to Liu *et al.*'s work [17], but they specifically designed it to allow the user to compare multiple entities [4]. Their user interface shows the distribution of positive and negative reviews along different features. However, they did not formally study the efficacy of their user interface.

A computational linguistics method often used for summarizing user-generated reviews is a sentiment analysis, which determines the semantic orientation of a given text. Turney [24] and Pong *et al*. [18] applied a sentiment analysis technique to analyze review text. Both systems used machine learning techniques to identify the semantic orientation of the phrases extracted using n-gram methods.

There are several systems that have applied sentiment analysis for tag cloud visualization. Dave *et al.* built a system to extract the tags from product reviews and display a sentiment score calculated based on those tags [6]. Lee *et al*. developed a system in which the user can manually add tags to an entity, and can rate whether the added tag contains a positive or negative sentiment [16]; the rated positivity/negativity of the tag is visualized using the font. Ganesan *et al*. incorporated emoticons into a tag cloud visualizing eBay seller feedback [9]. For example, a smiley face is automatically added to a tag that their system recognized as a positive tag. Although the effect of word sentiment visualization has not been studied in detail previously, we believe that incorporating a sentiment orientation into a tag cloud visualization could be useful for the user to quickly understand how positively or negatively the entity is reviewed.

Effects of a Tag Cloud on Different Tasks

Rivadeneira *et al*. categorized four user tasks for which a tag cloud could be useful [19]: *searching* (finding a particular term, often as a means to navigate to more detailed information about the term); *browsing* (casually exploring information without seeking any specific term); *impression formation* (building an impression about the entity that the tag cloud visualizes); and *recognizing* (providing additional information about the entity to support the user in identifying the entity she is seeking).

Several studies have been conducted on *searching* and *browsing*. In their study of the usefulness of tag clouds in information-seeking tasks, Sinclair *et al*. [22] concluded that such an interface is more useful for *browsing* tasks than for *searching* tasks. Kuo *et al*. demonstrated how a tag cloud displaying the search results of biomedical text at PubMed helped the user find the correct answers to simple medical questions [14]. They also found that a tag cloud was better for answering general questions (*e.g.*, whether a certain factor is transcriptive), but was not helpful for finding specific details (*e.g.*, the names of the genes involved in a particular biological process).

However, there still exist open research questions related to understanding how a tag cloud visualization could be useful and should be designed for beyond *searching* tasks. Viégas and Wattenberg [25] discussed the use of tag cloud visualizations for analytical purposes. These purposes are closely related to *impression formation* and *recognizing*, but the effect and design of a tag cloud visualization for these purposes have not been deeply studied. Although Review Spotlight can support *searching* or *browsing* for user-generated reviews, we are most interested in how Review Spotlight can support *impression formation*.

Effects of Tag Cloud Visual Features

The visual features of a tag cloud can impact the user's performance on *searching* tasks. For example, Rivadeneira *et al*. studied how many words people could recall after reading a tag cloud with different visual properties [19]. They discovered that words in larger font and those located at the upper-left corner of the tag cloud were easier to recall.

Bateman *et al*. examined the effect of nine visual properties of a tag for the task of determining the most important words based on visual appearance [2]. Their study revealed that the font size and font weight had strong effects on the word selection—but not the color.

The visual organization of a tag cloud also impacts user performance on *searching* tasks. Halvey and Keane studied the effect of three different layouts (a tag cloud, a linear list with vertical and horizontal alignments) with random and alphabetical ordering on *searching* tasks [10]. They found that the alphabetical ordering contributed to faster user performance than the random ordering in all layouts. In Schrammel *et al*.'s study of how a tag cloud word order affected the identification of specific tags and tags related to a particular topic [21], they showed that an alphabetical ordering was significantly faster than the other three ordering methods for searching for a specific tag. However, they did not find a significant difference in performance time when searching for tags related to a particular topic. These studies did not go beyond *searching* tasks whereas our focus in this study also includes *impression formation* tasks.

FORMATIVE STUDY

We conducted a formative study to learn about the types of information that users typically focus on and glean about a reviewed entity, and specific challenges they have with the user interface for user generated-reviews. We recruited 8 normal computer users who regularly did web browsing (4 males and 4 females between the ages of 20 and 50). They were likely to visit review websites in the past, but, like other normal computer users, it is unlikely that they frequently posted online reviews. In the formative study, we asked them to perform a think aloud as they read webpages with user-generated reviews (referred as "review pages" throughout the rest of this paper) to glean information about venues. We used four different review pages (two restaurant reviews from Yelp.com and two hotel reviews from TripAdvisor.com), each with more than 30 reviews at the time of the study. We instructed the participants to read the reviews as they normally would and to stop reading when they had drawn conclusions about that location. We then asked them to describe their impressions of the location as they would do for a friend who is considering visiting it. The entire session was audio-recorded and transcribed for analysis.

We found two important user behaviors from this study:

- **Formulating and adjusting an impression**: We found that the majority of the participants formed an initial impression about a location using the overall rating, the rating distribution, and photographs at the top of the page. The participants would then skim through the reviews to identify comments frequently repeated about the location by different reviewers, often noting the frequency with which those comments were expressed. Participants tended to notice and read reviews with remarks that differed from the commonly-expressed opinion, and then adjust their impression if necessary.

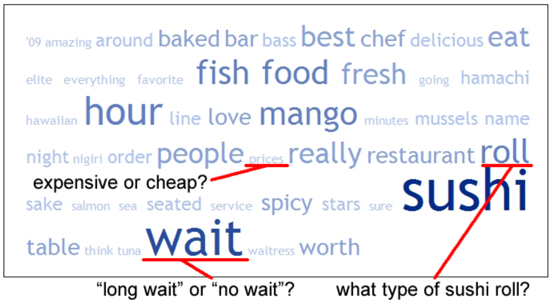

'09 amazing around baked bar bass best chef delicious eat
elite everything favorite fish food fresh going hamachi
hawaiian hour line love mango minutes mussels name
night nigiri order people prices really restaurant roll
expensive or cheap?
sake salmon sea seated service spicy stars sure sushi
table think tuna wait waitress worth
"long wait" or "no wait"?
what type of sushi roll?

Figure 2. A tag cloud of user-generated reviews using single words. It is possible to learn the general information about the restaurant, but details are hard to overview.

- **Verbalizing impressions with short phrases**: The participants also tended to verbalize their impression with short phrases, with descriptive information about the venue (*e.g.*, "Asian food") or subjective opinion statements (*e.g.*, "good steak").

These findings offer several design implications. First, our system should help the user gain a quick overview of the comments frequently mentioned about an entity. It is important for this overview to represent how many times reviewers commented on a particular aspect of the entity in their reviews (*e.g.*, whether "good food" is mentioned repeatedly by many reviewers or just once). The system should also allow the user to further evaluate the context in which those comments were made in order to support her in refining her impression about that entity. As participants skimmed the reviews, they often gleaned short phrases that captured the essence of each reviewer's comment. Thus, showing short phrases could accelerate the formation of impressions and decision making.

Based on the findings from our formative study and literature survey, we decided to use a tag cloud for our interface. Tag clouds already have been integrated into many websites, and have been familiar to many users. Thus, users are less likely to have problems with understanding the information displayed in the tag cloud and interacting with it. This is an important design aspect because we wanted our interface to be designed for casual users, and not for professionals who would examine the data deeply [5].

REVIEW SPOTLIGHT PROTOTYPE

We developed the Review Spotlight interface to help users quickly obtain information about a reviewed entity.

High-level Design

From our literature survey and observations made in the formative study, we decided to explore a tag-cloud-like interface. However, a standard tag cloud is not necessarily appropriate for our purposes. Figure 2 shows a tag cloud simply based on the frequency of words appearing in reviews for a restaurant. Although it is possible to learn general information about the restaurant from this tag cloud (*e.g.*, it is a Japanese restaurant, and the meals are probably good), the details that may be important for decision making are hard to overview. For instance, "roll" is a term mentioned frequently, but it is not clear what type of sushi rolls reviewers frequently mentioned. This finding motivated us to use a word pair as a meaningful chunk of information. Using word pairs is also in line with one of the design implications gained from our formative study (*i.e.*, using short phrases).

Figure 3 shows the interface of the Review Spotlight prototype. Although other tag cloud visualizations have also used word pairs extracted using n-gram methods [6], based on our findings in the formative study, we focused on adjective-noun pairs frequently mentioned in the reviews. The font size of each word pair is set to be proportional to the number of word pair occurrences. Review Spotlight also uses color to visualize the sentiment of each word, which will be discussed later.

When the user moves the cursor over an adjective-noun pair, Review Spotlight shows as many as four adjectives that are most frequently paired with that noun. In the example shown in Figure 3a, Review Spotlight shows "ridiculous," "worth," "bad," and "terrible" when the cursor is over the word pair "long wait." When the user clicks adjectives, the interface displays the number of occurrences of that adjective-noun word pair, and the sentences in which it appears in a separate textbox (Figure 3b). Thus, Review Spotlight supports quick examination of review details to enable the user to test her impressions.

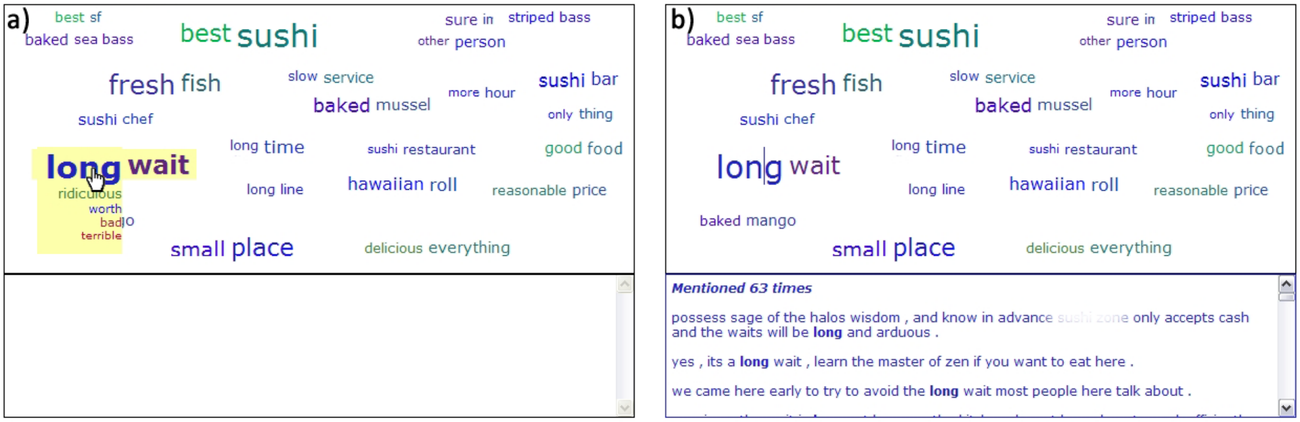
Implementation

To generate a summarization from the review text, Review Spotlight performs the following four steps: 1) extracts adjective-noun word pairs; 2) counts each word pair's occurrences; 3) performs a sentiment analysis of each word pair; and 4) displays the word pairs.

Review Spotlight first extracts adjective-noun word pairs from the review text. Using a part-of-speech (POS) tagger developed by Tsuruoka and Tsujii [23], our system labels the part of speech for the words in the review text. This allows Review Spotlight to identify the locations of the adjectives and nouns. Review Spotlight then pairs a noun with the closest adjective modifying it. Review Spotlight also extracts a word pair from a sentence with the "to be" verb. For instance, if Review Spotlight parses a sentence "The food is great," it extracts the word pair "great food." In addition, by focusing on adjective-noun pairs, Review Spotlight intrinsically removes noise introduced by common trivial words, such as articles and prepositions.

Review Spotlight then counts the number of occurrences of each word pair and groups word pairs by nouns. The system then eliminates the word pairs that only appear once in the original review text. It then calculates the font size for the extracted adjectives and nouns. The font size for a noun is determined linearly by its number of occurrences. The font size for an adjective is determined linearly by the number of occurrences of the word pair consisting of it and the associated noun. We set the minimum and maximum font sizes to 10 and 30 pixels, respectively.

Next, Review Spotlight performs a sentiment analysis on the word pairs using SentiWordNet [8], a context-free, word-based sentiment analysis tool. A sentiment value provided by SentiWordNet for each word consists of three scores (*i.e.*, positivity, negativity, and objectivity), and it is defined for each common use of the word. Review Spotlight first calculates the sentiment value for an adjective by taking the average of its sentiment values for all the use contexts. It then calculates the sentiment value for a noun by taking the average of the sentiment values of all paired adjectives weighted by the number of occurrences. It maps the sentiment value into the color scheme in which shades of green, red, and blue represent positive, negative, and neutral meaning, respectively; the darkness of the shade conveys the sentiment strength. Through a preliminary



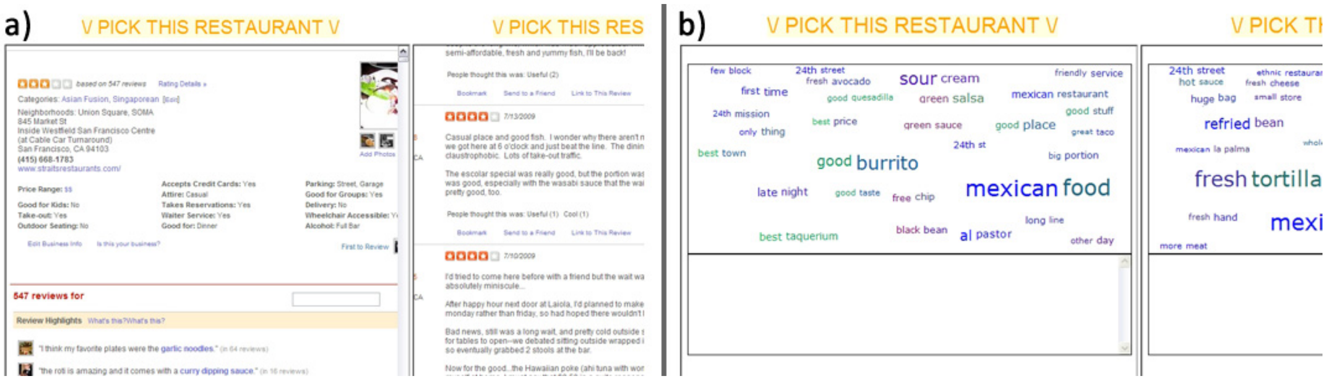
Mentioned 63 times

possess sage of the halos wisdom , and know in advance sushi zone only accepts cash and the waits will be **long** and arduous .

yes , its a **long** wait , learn the master of zen if you want to eat here .

we came here early to try to avoid the **long** wait most people here talk about .

Figure 3. Screenshots of Review Spotlight for reviews on the same restaurant used in Figure 2 (the name of the restaurant is masked in this figure): a) The user can see frequently-mentioned adjective-noun word pairs. She can also move the cursor over the word pair to see more adjectives that are paired with the noun; b) When she clicks an adjective (*e.g.*, "long"), Review Spotlight displays the original sentences from which the clicked word pair came.



based on 547 reviews Rating Details »
Categories: Asian Fusion, Singaporean [Edit]
Neighborhoods: Union Square, SOMA
845 Market St
Inside Westfield San Francisco Centre
(at Cable Car Turnaround)
San Francisco, CA 94103
(415) 668-1783
www.straitsrestaurants.com/

Price Range: $$
Good for Kids: No
Take-out: Yes
Outdoor Seating: No
Accepts Credit Cards: Yes
Attire: Casual
Takes Reservations: Yes
Waiter Service: Yes
Good for: Dinner
Parking: Street, Garage
Good for Groups: Yes
Delivery: No
Wheelchair Accessible: Y
Alcohol: Full Bar

Edit Business Info Is this your business? First to Review

547 reviews for

Review Highlights What's this?What's this?

"I think my favorite plates were the garlic noodles." (in 64 reviews)
"the roti is amazing and it comes with a curry dipping sauce." (in 16 reviews)

semi-affordable, fresh and yummy fish, I'll be back!

People thought this was: Useful (2)

Bookmark Send to a Friend Link to This Review

7/13/2009

Casual place and good fish. I wonder why there aren't n
we got here at 6 o'clock and just beat the line. The dinin
claustrophobic. Lots of take-out traffic.

The escolar special was really good, but the portion was
was good, especially with the wasabi sauce that the wai
pretty good, too.

People thought this was: Useful (1) Cool (1)

Bookmark Send to a Friend Link to This Review

7/10/2009

I'd tried to come here before with a friend but the wait wa
absolutely miniscule...

After happy hour next door at Laiola, I'd planned to make
monday rather than friday, so had hoped there wouldn't l

Bad news, still was a long wait, and pretty cold outside s
for tables to open--we debated sitting outside wrapped i
so eventually grabbed 2 stools at the bar.

Now for the good...the Hawaiian poke (ahi tuna with wor

Figure 4. Screenshots of the pages used in the laboratory user study. Two restaurant reviews were displayed side-by-side: a) Review pages (the name of the restaurant is masked in this figure); b) Review Spotlight.

experiment with the prototype system, we determined that users preferred the noun coloring based on this weighted average over the coloring based on the average of the sentiment values defined in SentiWordNet.

After the sentiment analysis, Review Spotlight performs a spatial allocation function to place the extracted word pairs within a given space (600 x 250 pixels by default, but the dimensions can be adjusted). Because Review Spotlight prioritizes larger word pairs in the allocation, an optimal layout (*i.e.*, the layout that accommodates as many word pairs as possible) can be accomplished by the greedy algorithm described by Kaser and Lemire [14]. However, we found that an optimal layout is visually complex, and large word pairs placed towards the top of the layout often attracted much more attention than other word pairs.

Review Spotlight instead places the word pairs randomly so that the user is not biased to any specific terms based on their placement position. Review Spotlight also adds padding around each word pair that is relative in size to the bounding box of the word pair. Although the resulting layout is much sparser than the optimal layout, our pilot studies showed that participants preferred it over the optimal layout because it was less visually complex. Review Spotlight performs this spatial allocation, starting with the largest word pair to the smallest, until it cannot find a location for a new word pair which does not cause an overlap on any other word pairs that have been placed already. Review Spotlight then combines as many four adjectives that are most frequently paired with the noun. These adjectives become visible when the user moves the cursor over the word pair.

The efficiency of the summarization process and the accuracy of the extracted word pairs are outside the scope of this paper. We have not experimented with these aspects of Review Spotlight in depth.

LABORATORY USER STUDY

We conducted a laboratory study to evaluate how Review Spotlight addresses user requirements for an interface summarizing user-generated reviews compared to traditional review pages. Although a comparative study against other interfaces is one possible type of user study, our intention here was to examine in depth how Review Spotlight supports *impression formation*, as defined by Rivadeneira *et al* [19].

Procedure

At the beginning of the study, we introduced the participant to a sample Review Spotlight interface and review page used in the study (Figure 4), and allowed the participants to explore both interfaces. After the participants became comfortable with the systems, we presented them the reviews for two restaurants side-by-side. Both reviews were displayed using either the Review Spotlight interface or the review pages. By separating the interfaces, we could differentiate the effects on the participants' decisions caused by the two different user interfaces. The participants were then asked to examine the reviews and decide which restaurant they would like to visit by clicking the link above each restaurant review. The system recorded all the mouse hover and click events on word pairs in Review Spotlight, and the time the participants spent making their decision between each restaurant pair. We also asked them to rate how strongly they preferred the selected restaurant over the other, and to provide the reasons for their decision verbally. The entire session was audio-recorded and transcribed.

We selected eight pairs of the restaurants that were located in a region unfamiliar to the participants and had more than 50 reviews on Yelp.com. Each restaurant within a pair offered similar cuisine, was of similar price range, and was from the same part of a city. Four pairs of restaurants had similar overall review ratings (PA1–PA4), and the rest consisted of pairs of restaurants in which one had a high overall rating and the second had a low overall rating (PB1–PB4). To examine the effect of Review Spotlight on impression formation in comparison to normal review pages, we presented four of the pairs (PA1, PA2, PB1, and PB2) to participants through both the Review Spotlight interface as well as the normal review page. To prevent participants from knowing that those four restaurant pairs were presented to them with both interfaces and from using predetermined decisions, we used the remaining four restaurant pairs (PA3, PA4, PB3, and PB4) as distracters. We presented PA3 and PB3 to participants using only the Review Spotlight interface; and we presented PA4 and PB4 to participants through the normal review pages. Thus, each participant viewed 6 Review Spotlight and 6 review page interfaces in the study. The presentation order of these twelve interfaces was randomized for each participant. The resulting Review Spotlight interfaces contained 26 word pairs with 66 hidden adjectives on average. We noticed that some word pairs were noises (*e.g.*, meaningless word pairs). However, we did not perform any manual filtering of the word pairs because we wanted to test our system in a natural manner.

Apparatus

All the Review Spotlight summarizations and review pages were prepared before the experiment, and stored on the computer used in the experiment. We cached all the pages on the computer before the study to minimize their loading time. We used a laptop running Windows Vista with a 30.5 x 19.0 cm, 1280 x 800 pixel display. The screen was sufficiently large for the participants to comfortably read the reviews for two restaurants side-by-side. We also provided the participants with a mouse.

Participants

Ten individuals (5 male and 5 female between the ages of 20 and 50) with a variety of backgrounds (such as students, system administrators, a retail manager, a home maker, and an accountant) participated in this study. None of them participated in our formative study. All participants regularly did web browsing, but none of them had any significant experience of writing and posting online reviews. The study lasted approximately 50 minutes. The participants were compensated with $20 cash.

LABORATORY STUDY RESULTS

Performance Time

We examined the data for the four restaurant pairs that were presented in both the Review Spotlight interface and the review pages (*i.e.,* the non-distracters) to compare the time that participants took to make a decision. Our unpaired Welch's t-test revealed that the participants spent less time with the Review Spotlight interface (122 seconds on average, SD=49) than the review pages (157 seconds on average, SD=63; t_{73}=2.75, p<.01). This finding is in line with the qualitative evidence we gathered from the study.

> *It's faster. Instead of like going through reading so much non-sense, [I can] just pick up important things right away.*

Forming Detailed Impressions Using Review Spotlight

In 30 out of the 40 cases, the participants chose the same restaurants with both user interfaces. The participants often cited the ratings and the number of reviews to explain their decision when viewing the review pages whereas they used specific details about the restaurants (*e.g.*, food or service) to explain their decisions with Review Spotlight. For example, one participant picked the same restaurant using different reasons as follows:

> ***With the review pages**: This one has more reviews and more positive [reviews]. Overall rating is higher too.*

Rating	Same Choice		Different Choice	
	RP	RS	RP	RS
Average (SD)	2.8 (0.75)	2.9 (1.0)	3.0 (1.0)	2.2 (1.0)

Table 1. The results of the self-rated preferences on the choices through the review pages (RP) and Review Spotlight (RS). A higher value (between 0 to 4) means a stronger preference. Same Choice represents the 30 cases in which the participants chose the same restaurants with both interfaces, and Different Choice represents the other 10 cases.

> ***With Review Spotlight**: I know that people say even though the wait is long, it's worth it. Usually I don't like waiting for food, but it must be good. And the price is reasonable. [The other restaurant] seems like ok too, but some people say the portion is really really small.*

Even in the ten instances where participants chose different restaurants between the two interfaces, they still provided a greater level of detail to support their decision when using Review Spotlight. In seven of these cases, the participants selected a restaurant with weak preference over the other restaurant. For example, one participant selected one restaurant with the review pages by focusing primarily on the ratings it received.

> *Everyone seems to think the restaurant on the left is very average, like there is nothing wrong there, but it wasn't fantastic. People on the right generally seem to think…it was average and some people thought above average, but the reviews generally seem to be more positive.*

However, he chose the other restaurant with the Review Spotlight interface because he thought that the restaurant he chose with the review pages may offer a small portion of food. In this case, Review Spotlight helped the participant uncover specific information that was important to his decision from the review text.

> *People on the left generally seem like they are happy with the food and the portion, and price. People on the right, …they mentioned what they ate, and seems like some people were happy with food, but some people thought the portion is small, so I was more sure that people thought more positively about the restaurant on the left.*

Quantitative Analysis of Review Spotlight Usage

Our system recorded the number of mouse hover events over word pairs in the Review Spotlight interfaces. Halvey and Keane's study indicated that the user typically scans the tag cloud rather than reads it while performing *searching* tasks [10]. To study whether the participants only scanned word pairs, we first removed the 3008 hover events that lasted 200 milliseconds or less from the 7240 hover events recorded in all 12 Review Spotlight interfaces (including the four distracters) as unintentional hover events. We determined this time threshold from the fact that the average reading speed is 200–240 milliseconds/word [13]. Although this cutoff does not mean a perfect separation of intentional and unintentional hover events (and furthermore, it might be too conservative), we believe that this is still a good approximation. As a result, we had 4232 intentional hover events, and the average number of intentional hover events per Review Spotlight page per participant was 35.3 (SD=2.5). This implies that the participants tended to read the word pairs in our experiment because they needed to examine what words appeared in Review Spotlight in order to form their impressions.

Our system also recorded the number of mouse click events on word pairs. We had a total of 1200 click events in the Review Spotlight interfaces. This means that the average number of click events per Review Spotlight page per participant was 10.0 (SD=1.3). 658 clicks (54.8%) occurred on the adjectives that appeared by default, and the rest of the clicks occurred on the hidden adjectives. Furthermore, we found that 28.4% of the intentional hover events resulted in a click. These numbers indicate that Review Spotlight encouraged the participants to explore more details about how the words were used in the actual reviews.

We were unable to analyze the types of word pairs which the participants tended to view. One reason for this is that some word pairs, such as "good food", were displayed by the Review Spotlight more frequently than other word pairs. The same word pairs also were displayed with different font sizes at different locations in the twelve Review Spotlight summarizations. We leave a quantitative analysis of the types of word pairs that the users typically look at to form their impressions about the restaurants for future work.

Qualitative Analysis of User Strategies

We did not provide the overall rating in Review Spotlight because we found that it does not always match word pairs that appeared in the interface. But participants successfully gained an idea of how positive or negative the reviews about the restaurant were. Although the word pair coloring based on the sentiment analysis was intended for this purpose, nine of the participants rather looked at the words themselves, their font sizes, and their exact numbers of occurrences to obtain their general impression.

> *I more used words that were used, and secondly the size of the words… and then I would read similar comments, and then I would look for [details of] where people say bad things, just to have a comparison of how strongly they say they were bad or good, and how often… I completely forgot about the color.*

Another common strategy was to focus on particular adjectives, such as "good", "great", or "poor". The participants also looked at how many times each word pair was mentioned in the original review page.

> *The first thing I would do usually is [look at] the biggest one. Then I checked things that seem relevant, like,… if I saw "best" or "worst," I would definitely check that. Also specific dishes…I see how many times they say that. Like if one side mentioned "best" twice, and the other side mentioned 20 times.*

The participants generally liked the ability to view a word pair in its original review. Half of the participants explicitly mentioned that this feature reduces the effort needed to read the reviews in order to find detailed information about what they were interested in.

> *I like that they are just short sentences that got right to the point of what I was looking for… It was neat to just click it and see what people had to say about. How many times it was mentioned, I also really liked.*

> *The thing I like the most was if I was only interested in food or service…, you can click that and see all comments about one particular aspect… Pick up one thing you're more interested in instead of reading through the reviews.*

DISCUSSION

One aspect we wanted to focus on in our study was how Review Spotlight could help users perform tasks which required what Rivadeneira *et al*. described as *impression formation* [19]. The participants often (30 out of 40 times) were able to choose the same restaurant with nearly the same level of preference using both the Review Spotlight interface and current review pages (see Table 1). However, our participants were able to make this decision significantly faster with Review Spotlight. Furthermore, their decisions were always guided by more detailed impressions than ones formulated when they used the traditional review webpage interface, even when the participants chose different restaurants.

Providing a More Consistent Presentation

Four participants explicitly commented on the presentation of the word pairs. Because the system places the word pairs without following any particular layout, it causes an inconsistency in the presentation. For example, one participant had trouble finding particular information that she wanted to know in Review Spotlight. Another participant described that the Review Spotlight interface was harder to find particular information that he was seeking because he often could not tell quickly whether that information was actually displayed.

> *Information may be there if you are looking…, but it's hard [to determine] if it's missing or not… You may be looking for price range, or attire, or whether they deliver. [On review pages,] it's easier [to find] just from looking at the top, but here (on Review Spotlight) [such information] may or may not be there.*

We decided to use a random ordering instead of a specific word ordering. This seemed to cause problems when the participants were looking for a particular word. However, the participants generally could identify words that are important or interesting to them quickly.

> *If you're looking for something, looking for wine, or…, it's easy to spot whereas [on the review page], you have to look for it, search for it.*

Random word ordering resulted in accidental learning and facilitated the *impression formation* tasks defined by Rivadeneira *et al* [19]. That is, presenting the user with different word orderings mitigates the problem of biasing them towards particular word pairs.

> *The good thing about [Review Spotlight] is that I came across the things like specific dishes that would've been harder to see in [the review pages]… Like people, if they mention mac 'n' cheese a lot or best burrito, you would be able to see that [in Review Spotlight].*

Participants also mentioned the inconsistency of the size of the word pair. Because the size of each word pair was determined based on its number of occurrences within the review text for one entity (in this case, a restaurant), the same font size across multiple Review Spotlight summarizations does not necessarily mean that the word pair has the same number of occurrences. This can be confusing when the user wants to compare multiple entities, as the participants did in this experiment.

In addition to user reviews, every review page provides a summary of basic information about the restaurant (*e.g.*, the hours, price range, and atmosphere) and visual elements (*e.g.*, a photo or a graph). The user may use these elements to quickly compare two restaurants. As seen in Figure 1, we envision integrating Review Spotlight into the original review page instead of replacing it completely. In this manner, the user would be able to obtain the general information about a restaurant that is typically found in a review page, but still be able to obtain an overview of the reviews with Review Spotlight.

Graceful Recovery from Linguistic Analysis Problems

As expected, the participants noticed and commented on errors caused by the natural language processing we used. Participants noticed that some word pairs did not make much sense and that the color of the fonts did not often match to what they thought (*e.g.*, "impeccable" has a high negativity value in SentiWordNet). This type of problem could be addressed by incorporating a more sophisticated method of determining the relevant information and sentiment in the review [12]. Another linguistic analysis problem is that the meanings of some word pairs were context-dependent. For example, if one reviewer commented "Last time we went, we had and loved the grilled chicken," and another commented "I will avoid their grilled chicken next time," the current Review Spotlight implementation would detect that "grilled chicken" is a common pair despite the contrasting reactions. Similarly, the current Review Spotlight implementation does not accurately extract the word pairs that appeared in negative sentences (*e.g.*, "This is not a good restaurant").

Review Spotlight allows the user to click on an adjective-noun pair to see the original review text containing that word pair so that she can learn additional information about the reviewed entity. But our user study showed that this feature also gracefully compensated for imperfect word pair extractions.

> *Something that I wouldn't notice is that someone says "this place was not very good," so they changed the positive word into the actual negative review about it, so that's why I found that it was important for that word, in particular, that I actually clicked on that in every review that I looked at… That's how it kind of differentiated if it's actually a good place to go to or NOT good place to go to.*

Improvements on the accuracy of the word pair extraction and sentiment analysis could contribute to a better review summarization. However, user-generated reviews, in general, are hard to parse automatically because they do not follow any particular format and they are not always grammatically correct. The findings from our study suggest that providing the user with a way to obtain additional contextual information about the word pairs is a useful way to recover from the errors caused by the linguistic analysis.

Controlling Displayed Word Pairs

Some participants suggested a feature to allow the user to adjust which word pairs are displayed. For instance, Review Spotlight could have a scale widget to tune the percentage of positive/negative word pairs to appear in the interface. Based on participant comments, we also found two interesting parameters that could be adjusted by the user in order to enhance the usefulness of Review Spotlight.

Subjective-Objective Parameter

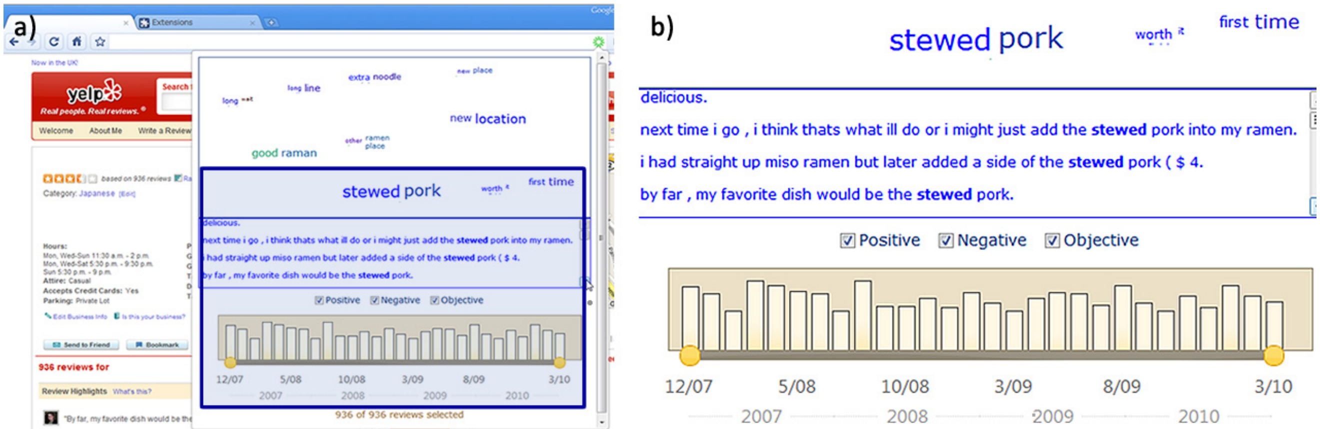
One interesting adjustment parameter is the degree of subjectivity or objectivity in the word pair. One participant told us that she wanted to have more objective (descriptive) words than subjective (evaluative) words to understand what the restaurant is like and what it offers. She explained the aspect that she did not like in the prototype Review Spotlight interface as follows:

> *These words [subjective word pairs shown in Review Spotlight], it's like somebody's opinions of the restaurants, not compared to "hot source" or "sweet potato", like description of the actual food or service. So, it's like you have mixed [subjective] comments as well as [a] description of type of food.*

However, subjective words were still useful for many participants to make decisions. Therefore, allowing the user to adjust the amount of subjective/objective words appeared in the interface could improve Review Spotlight.

Time Parameter

Another interesting improvement could be to display word pairs based on their post date, as the trend of the reviews often changes over time. For example, reviews might change dramatically after a food poisoning incident or the



delicious.

next time i go , i think thats what ill do or i might just add the **stewed** pork into my ramen.

i had straight up miso ramen but later added a side of the **stewed** pork ($ 4.

by far , my favorite dish would be the **stewed** pork.

Figure 5. The Google Chrome extension of Review Spotlight. The histogram indicates the number of the reviews over time, and the user can control which sentiment types to be displayed: a) the whole view of our extension; b) the enlarged view of the highlighted area in Figure 5a.

start of a new chef. Allowing the user to filter reviews from before and after an influential incident could enable them to better understand the entity as one participant commented:

> *One good thing about [the review page] is that you tend to read only the recent reviews... Maybe things change,… like if they… improve the service… the recent reviews indicate that as opposed to old reviews [in which] people complain about the service... So maybe if you have some sort of a time bar on the top you can sort of drag it and see how tags change from recent to oldest.*

Such a feature has been included in different web-based systems. The Zoetrope interface allows the user to examine how a portion of a particular webpage has changed over time [1]. Dubinko *et al.* developed an algorithm to show the most salient tags that were posted by users in Flickr for a given time interval [7]. Thus, this feature could also be beneficial in Review Spotlight.

REVIEW SPOTLIGHT EXTENSION

Based on the findings from the laboratory study, we revised our Review Spotlight interface and developed a Google Chrome extension which resembles what we illustrated in Figure 1. This Review Spotlight extension supports reviews available on Amazon.com as well as Yelp.com.

Figure 5 shows the revised Review Spotlight interface. We added several features suggested by participants in our laboratory study. First, the interface includes a histogram showing the number of reviews entered for a given entity along a timeline; this gives the user an overview of how the number of the reviews changes over time. The sliders below the histogram allow the user to specify the time period for which Review Spotlight should summarize the reviews. Finally, the new interface also provides checkboxes for selecting sentiment types. This feature allows the user to specify whether positive, negative, and subjective word pairs should be displayed by clicking the corresponding checkboxes.

We made our Review Spotlight extension available to anyone in our website and Google Chrome Extension Gallery for three months. To enable the extension, the user had to consent with our data gathering of their usage of the extension. We did not extensively recruit the users for our extension because it was still a research prototype. Eleven users volunteered to use our extension. Our log showed that on average the users accessed 29.0 different webpages (SD=38.0) on which Review Spotlight could have been used; and on average they used Review Spotlight 19.9 times on average (SD=25.7). Thus, 68.6% of the user's visits to Yelp or Amazon webpages (for which the Review Spotlight was available) resulted in its use. This indicates that our users did use Review Spotlight to view a summary of user-generated reviews provided by Review Spotlight. Because we implemented the extension to only log information when the user visits a Yelp or Amazon webpage for which it can summarize, it is not possible for us to know if a user still has the extension installed or has removed it. Thus, we can only report how many days after first installing the extension was a user's most recent use. On average, participants' most recent use of the extension came 35.7 days after they installed it (SD=29.8). This indicates that participants saw value in the tool and continued to use it.

CONCLUSIONS AND FUTURE WORK

The overwhelming number of user-generated reviews and their inconsistent writing style often require a significant amount of time and effort to read, and can result in important information being obscured from the user. To counter such challenges, we developed Review Spotlight, a user interface that helps users with impression formation by summarizing user-generated reviews using adjective-noun word pairs. Our interface provides a quick overview of the reviews and allows the user to explore the details of reviews in which word pairs are mentioned. Our laboratory user study with Review Spotlight showed that participants could form detailed impressions about restaurants from examining an overview of the reviews and the original context of the word pairs that interested them. The participants could also choose their preferred restaurant with Review Spotlight significantly faster than with traditional review webpages.

We do not claim that Review Spotlight would outperform other kinds of interfaces or visualizations. However, this study revealed that Review Spotlight has the potential to support the quick formation of an impression of the reviewed entity. We plan to perform a comparative evaluation between different kinds of interfaces which summarize user generated reviews.

One concern with using user-generated reviews is their reliability. The current Review Spotlight did not do any filtering of word pairs or review text. Thus, it could provide false impressions if the original reviews were posted under malicious motivations (*e.g.*, a reviewer posted false information or negative comments to decrease the venue's popularity). We can address this issue by adding more weight to reviews from highly-rated reviewers (*e.g.*, top 100 reviewers) or reviews other users found useful.

The current interface only supports reviews written in English. Adopting an international POS tagger like TreeTagger [20] would enable the system to extract adjective-noun pairs from the text and display the review written in a different language in the same way as that it currently does with English text. However, adjective-noun pairs might not be the best way to summarize review pages in other languages. Thus, an investigation on what users would find to be more appropriate in other languages is necessary for internationalization.

ACKNOWLEDGEMENTS

We would like to thank members of the DGP, Fanny Chevalier and Elaine M. Huang for their comments on this paper. We also thank all the participants in our user studies.

REFERENCES

1. Adar, E., Dontcheva, M., Fogarty, J., and Weld, D. S. Zoetrope: Interacting with the ephemeral web. In *Proc. of UIST 2008*, ACM Press, 239-248.
2. Bateman, S., Gutwin, C., and Nacenta, M. Seeing things in the clouds: the effect of visual features on tag cloud selections. In *Proc. of HT 2008*, ACM Press, 193-202.
3. Carenini, G., Ng, R. T., and Pauls, A. Interactive multimedia summaries of evaluative text. In *Proc. of IUI 2006*, ACM Press, 124-131.
4. Carenini, G. and Rizoli, L. A multimedia interface for facilitating comparisons of opinions. In *Proc. of IUI 2009*, ACM Press, 325-334.
5. Chevalier, F., Huot, S., and Fekete, J. D. WikipediaViz: conveying article quality for casual wikipedia readers. In *Proc. of PacificVis2010*, IEEE, 215 – 222.
6. Dave, K., Lawrence, S., and Pennock, D. M. Mining the peanut gallery: opinion extraction and semantic classification of product reviews. In *Proc. of WWW 2003*, ACM Press, 519-528.
7. Dubinko, M., Kumar, R., Magnani, J., Novak, J., Raghavan, P., and Tomkins, A. Visualizing tag over time. *ACM Transactions on Web 1*, 2 (2007), 7.
8. Esuli A., and Sebastiani F. SENTIWORDNET: a publicly available lexical resource for opinion mining. In *Proc. of LREC 2006*, ELRA, 417-422.
9. Ganesan, K. A., Sundaresan, N., and Deo, H. Mining tag clouds and emoticons behind community feedback. In *Proc. of WWW 2008*, ACM Press, 1181-1182.
10. Halvey, M. J. and Keane, M. T. An assessment of tag presentation techniques. In *Proc. of WWW 2007*, ACM Press, 1313-1314.
11. Hu, M., and Liu, B. Mining and summarizing customer reviews. In *Proc. of KDD 2004*, AAAI, 168-177.
12. Jo, Y., and Oh, A. Aspect and sentiment unification model for online review analysis. In *Proc. of WSDM 2011* (to appear).
13. Just, M. A., and Carpenter, P. A. *The psychology of reading and language comprehension.* Allyn & Bacon, Boston, 1987.
14. Kaser, O., and Lemire, D. Tag-cloud drawing: algorithms for cloud visualization. In *WWW 2007 Workshop on Taggings and Metadata for Social Information Organization 2007*, ACM Press.
15. Kuo, B. Y., Hentrich, T., Good, B. M., and Wilkinson, M. D. Tag clouds for summarizing web search results. In *Proc. of WWW 2007*, ACM Press, 1203-1204.
16. Lee, S. E., Son, D. K., and Han, S. S. Qtag: tagging as a means of rating, opinion-expressing, sharing and visualizing. In *Proc. of SIGDOC 2007*, ACM Press, 189-195.
17. Liu, B., Hu, M., and Cheng, J. Opinion observer: analyzing and comparing opinions on the Web. In *Proc. of WWW 2005*, ACM Press, 342-351.
18. Pong, B., Lee, L., and Vaithyanathan S. Thumbs up? sentiment classification using machine learning techniques. In *Proc. of EMNLP 2002*, ACL, 79-86.
19. Rivadeneira, A. W., Gruen, D. M., Muller, M. J., and Millen, D. R. Getting our head in the clouds: toward evaluation studies of tagclouds. In *Proc. of CHI 2007*, ACM Press, 995-998.
20. Schmid, H. Probabilistic part-of-speech tagging using decision trees. In *Proc. of NEMLP 1994*, 44-49.
21. Schrammel, J., Leitner, M., and Tscheligi, M. Semantically structured tag clouds: an empirical evaluation of clustered presentation approaches. In *Proc. of CHI 2009*, ACM Press, 2037-2040.
22. Sinclair, J., Cardew-Hall, M. The folksonomy tag cloud: when is it useful? *Journal of Information Science 34*, 1 (2008), 15-29.
23. Tsuruoka Y., and Tsujii, J. Bidirectional inference with the easiest-first strategy for tagging sequence data. In *Proc. of HLT/EMNLP 2005*, ACL, 467-474.
24. Turney P. Thumbs up or thumbs down? semantic orientation applied to unsupervised classification of reviews. In *Proc. of ACL 2002*, ACL, 417-424.
25. Viégas, F. B. and Wattenberg, M. Tag clouds and the case for vernacular visualization. *Interactions 15*, 4 (2008), 49-52.