Clustered Layout Word Cloud for User Generated Online Reviews

Ji Wang

Thesis submitted to the Faculty of the

Virginia Polytechnic Institute and State University

in partial fulfillment of the requirements for the degree of

Master of Science

in

Computer Science and Applications

Christopher L. North, Chair

Yong Cao

Naren Ramakrishnan

October 25, 2012

Blacksburg, Virginia

Keywords: Word Cloud, Text Visualization, User Study

Clustered Layout Word Cloud for User Generated Online Reviews

Ji Wang

(ABSTRACT)

User generated reviews, like those found on Yelp and Amazon, have become important reference material in casual decision making, like dining, shopping and entertainment. However, very large amounts of reviews make the review reading process time consuming. A text visualization can speed up the review reading process.

In this thesis, we present the clustered layout word cloud – a text visualization that quickens decision making based on user generated reviews. We used a natural language processing approach, called *grammatical dependency parsing*, to analyze user generated review content and create a semantic graph. A force-directed graph layout was applied to the graph to create the clustered layout word cloud.

We conducted a two-task user study to compare the clustered layout word cloud to two alternative review reading techniques: random layout word cloud and normal block-text reviews. The results showed that the clustered layout word cloud offers faster *task completion time* and better *user satisfaction* than the other two alternative review reading techniques.

Contents

List of Figures

Chapter 1

Introduction

1.1 Motivation and Background

User generated reviews have become a popular resource for decision making in recent years. For example, Amazon, Yelp, and IMDB provide a large quantity of user generated reviews for user decision making or reference. However, in most cases, reading large quantities of reviews is a difficult and time-consuming task. Also, a person does not want to spend a large amount of time making the non-critical decision of where to dine or what movie to watch. In this situation, a visualization that summarizes the user generated reviews is needed for perusing reviews.

Current review visualization approaches [31] support the ability to visualize quantitative information of businesses. However, quantitative information does not always account for the descriptions given in the user review, which can often be detailed, providing more context about the business that is being reviewed. Text visualization has the potential to process user generated reviews and provide more context than only quantitative information.

Word clouds are a popular text visualization presenting word frequency, popularity or importance through font size. A majority of word clouds lay out the words in a random, though aesthetically appealing way. They can summarize large amounts of text in limited space and guide users' attention to discover more related context information. Word clouds have been widely used in both business and research, i.e. Opinion Cloud [7], Tirra [15], Review Spotlight [32] and Wordle [12] [26].

Although they are an informative tool, the random layout of word clouds does not provide an overview of the data. The random layout of word clouds can require significant mental demands when trying to understand the review content, because users need to scan the entire word cloud to gain an overview or to find specific keywords of interest. In word cloud, there is no multiple occurrences of one word. Random word clouds only provide word frequency information, but they do not encode any relationship among keywords. Thus, a word cloud's random layout strategy leads to a higher cognitive load.

My solution is to encode context of the review content into the word cloud. The context information here is the semantic relationships between the keywords, especially the grammatical relationships. Semantics are a very important dimension of review content [32] [14]. Grammatical dependency parsing is an effective NLP approach [5] which produces a *grammatical dependency graph (GDG)*. GDG has been used in applications [27] [4] and enhances the user's understanding of text sources. In this thesis, I will spatially cluster the word cloud based on GDG semantic relationships. Color encoding of the cluster memberships of word tags can help users to distinguish different clusters more quickly. My hypothesis is that the context information in clustered layout word cloud will improve user's review reading performance.

1.2 Research Questions

In this thesis, we present an approach to embed semantic information from grammatical dependency parsing into the layout of word clouds for review reading tasks. So, our two research questions are:

1. How to embed semantic information of GDG into word clouds?
2. Will embedded semantic information of GDG layout influence user performance? Specifically, will this word cloud layout, in which words that are close in GDG space are also spatially close in word cloud layout plus clustering color encoding, improve users' performance and experiences? If so, how?

1.3 Contributions

In this thesis, our primary contribution is the concept and prototype of embedding semantic information of GDG into a clustered layout word cloud (CLWC) to enhance user's performance in review reading. Our word cloud is designed to afford more insight about user generated reviews by creating clusters based on semantic information.

As a secondary contribution, we conducted a user study that compared our clustered layout word cloud with another two alternative review reading techniques: normal review reading and normal random word cloud. The results of user study show that CLWC improves users' performance: faster task completion time (while maintaining low error rate) and better user satisfaction.

Chapter 2

Related Work

2.1 Review Visualization

Review visualizations can be categorized as the following two types based on the characteristics of reviews.

2.1.1 Quantitative Feature Based Review Visualization

Quantitative feature based review visualizations are often used to visualize general rating, price level and other quantitative features of business. Wu et al presents a visualization to show hotel feedback based on quantitative features [31]. The drawback of quantitative feature based visualizations is that, for many products or services, it cannot use quantitative feature to describe whole things simply.

2.1.2 Content Based Review Visualization

Content based review visualizations visualize text content of reviews data, like review opinion. Liu and Street firstly use a natural language processing approach to analyze reviews content and then visualize the opinion extracted from these reviews using a bar chart [14]. Caternini and Rizoli presented a multimedia interface to visualize the fixed features extracted from review content that reflected user's opinions [2]. Koji et al presented Review Spotlight: a frequency based word cloud with "adjective plus noun" word pairing for visualizing user reviews [32], shown as Figure 2.1. In Figure 2.1, a) Adjective-noun word pairs based on term frequency. User can hover the cursor on the word pair to see more adjectives; b) User can select an adjective (e.g., long), then Review Spotlight shows the related raw review content. Huang et al present RevMiner: a smartphone interface that also used natural language processing to analyze and navigate reviews [10], shown as Figure 2.2. In Figure 2.2, the Tag Cloud interface (Left) shows the attribute-value pairs in different font size by term frequency. The Color Bar interface (Right) presents the attributes into different clusterings and shows the values in a stacked bar chart with color-encoded.

Both of the word clouds presented in Review Spotlight and RevMiner used a random layout word cloud approach. Review Spotlight use "adjective plus noun" word pairing extraction and RevMiner used text similarity for semantic information. Our clustered layout word cloud is also a *content based* review visualization. Different from Review Spotlight and RevMiner, we use grammatical dependency parsing NLP approach to provide semantic information.

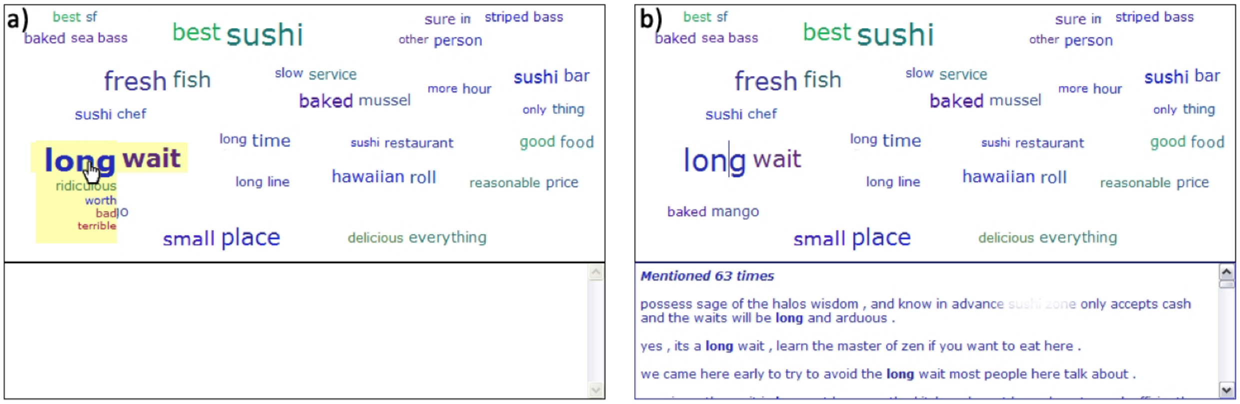

Figure 2.1: Review Spotlight [K. Yatani, M. Novati, A. Trusty, and K. N. Truong, Review spotlight: a user interface for summarizing user-generated reviews using adjective-noun word pairs, in Proceedings of the 2011 annual conference on Human factors in computing systems - CHI 11, 2011, p. 1541.]. Used with permission of K. Yatani, 2012.

2.2 Word Cloud Visualization

2.2.1 Basic Word Clouds

Word clouds (tag clouds) are a very popular text visualization tool.They provide the word frequency data of a variety of text sources and encode the frequency into the word's font size. They have become very popular on news website to summarizes news articles based on keywords [29] and on photo sharing website to show the content categories of images [28].

Kaser and Lemire presented an algorithm to draw word clouds in the limited space of HTML, based on table components [11]. Viegas et al. invented a more compact word cloud entitled Wordle, which uses the greedy space filling approach create the word cloud while maintaining

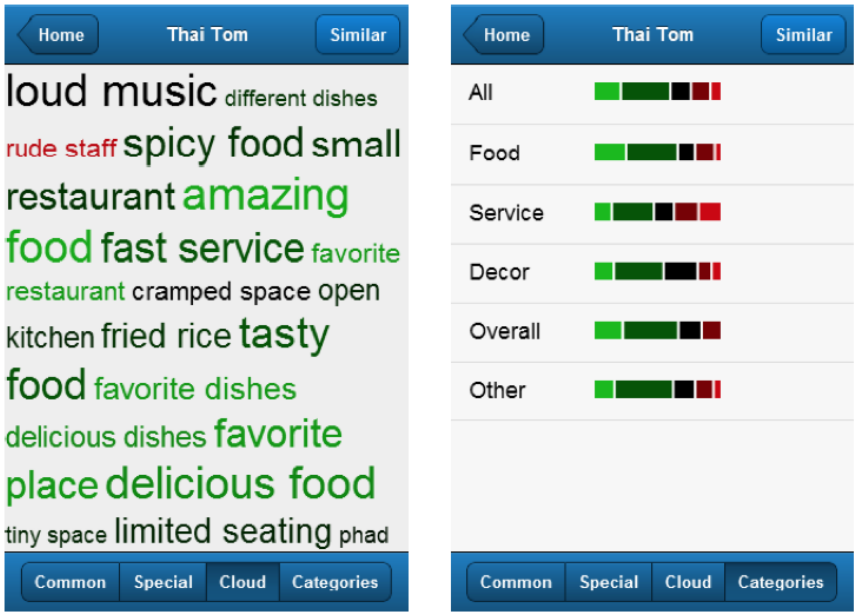

Figure 2.2: RevMiner [J. Huang, O. Etzioni, L. Zettlemoyer, K. Clark, and C. Lee, RevMiner, in Proceedings of the 25th annual ACM symposium on User interface software and technology - UIST 12, 2012, p. 3.]. Used with permission of J. Huang, 2012.

the word cloud's feature of using font size represent word frequency or importance level [26], shown as Figure 2.3.

The essential idea of Wordle generating algorithm is to place the word where it wants to be. In most of wordle, the place is random position. If it intersects any of the previously placed words, we need to move it one step along an ever-increasing spiral until it find the available position without intersections or overlapping. How to efficiently detect on the large images is the hardest work. In wordle project, they used boundary box to detect. But in our project, we use double image buffers to implement this part and the running time is on-the-fly.

Figure 2.3: Wordle [F. B. Viegas, M. Wattenberg, and J. Feinberg, Participatory visualization with Wordle., IEEE transactions on visualization and computer graphics, vol. 15, no. 6, pp. 113744, 2009.]. Used with permission of F. B. Viegas, 2012.

2.2.2 Word Cloud Embedded with Semantic / NLP Results

Recently, word clouds have begun to embed semantic information based on natural language processing results to enhance the detail provided in word clouds. Cui et al presented a context preserving tag cloud which used sophisticated similarity metrics to generate keywords and the distance between words [3], shown as Figure 2.4. Their similarity metrics included three parts: word importances (such as TF IDF), co-occurrence in time and co-occurrence in same sentences. But, in our project, the similarity metric is based on the weight of edges in GDG.

Paulovich et al presented ProjCloud, which was a clustering word cloud approach that provided semantic information for documents clustering [22], shown as Figure 2.5. It used the MDS approach to project documents into polygons which is defined by the documents clus-

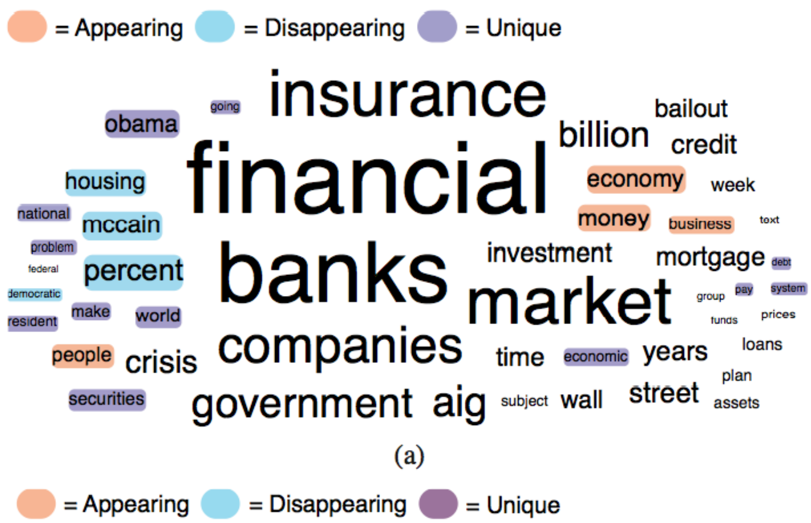

Figure 2.4: Context Preserving Tag Cloud [W. Cui, Y. Wu, S. Liu, F. Wei, M. Zhou, and H. Qu, Context-Preserving, Dynamic Word Cloud Visualization, IEEE Computer Graphics and Applications, vol. 30, no. 6, pp. 4253, 2010.]. Used with permission of W. Cui, 2012.

terings and present the relationships between documents. Then, the keywords are extracted from the clustered documents in same polygon and sorted based on document co-occurrence (Fiedler vector).

Both the Review Spotlight word cloud generation approach from Yatani et al [32] and the RevMiner word cloud generation approach from Huang et al [10] used natural language processing to extract semantic information from review content and embed it into their word clouds.

This thesis work is related to Review Spotlight [32] and ProjCloud [22]. Review Spotlight uses random layout word cloud to show review content. However, our clustered layout word

Figure 2.5: ProjCloud [F. Paulovich, F. Toledo, and G. Telles, Semantic Wordification of Document Collections, Computer Graphics, vol. 31, no. 3pt3, pp. 11451153, Jun. 2012.]. Used with permission of F. Paulovich, 2012.

cloud embeds semantic information into the layout to present review content. Different from ProjCloud, our word cloud focuses on term level of review content, not document level.

2.3 Word Cloud Evaluation

Word cloud evaluation is difficult because of measuring the insight of a visualization from users [20].

2.3.1 Evaluation of Word Cloud Layout

Lohmann et al evaluated how the layout of word clouds influenced user's performance in different types of task [16]. They found thematic layouts were good for finding words that belonged to a specific topic task. However, Lohmann's work was not based on real data, but on an experimental semantic layout word cloud that they created. Our design provided a clustered layout word cloud that was created based on algorithms to be described in following sections, so our user study is different in that the word clouds were dynamically created. At the same time, our study data is real review data from Yelp Academic Dataset [33].

2.3.2 Evaluation based on Task Completion

Task completion tests are one way to evaluate and compare word cloud visualizations. Bateman et al provided an evaluation for word clouds on the aspect of finding the most important word in a tag cloud based on font size, color and position [1]. They found font size and font weight were the aspects that caught the user's attention most when using word clouds. Rivadeneira et al conducted a user study to obtain performance metrics of four types of tasks using word clouds. The four tasks were search, browsing, impression formation, and recognition [23]. They used the impression formation metric for word cloud evaluation. Yatani et al also used impression formation to evaluate their Review Spotlight word cloud visualization [32]. Schrammel et al presented topical word cloud layouts that could assist users in finding specific words as well as words that were relevant to specific topics [24].

To the best of our knowledge, this thesis work is first attempt to use real review for clustered layout word cloud. At the same time, we also use real review data and real life task design for user study. So, this work can be easy to apply in our daily life.

Chapter 3

Clustered Layout Word Cloud

3.1 Design Principles

User generated reviews are a useful tool for casual decision making in daily life, when it comes to making selections in social settings and purchases. Currently though, there are cumbersome amounts of reviews and a user wants to assess and make their decision in a short amount of time, so current review reading makes it difficult for user to make casual decision based on the reviews.

In order to improve the user generated review reading process via the word cloud visualization approach, the word cloud needs to meet two design requirements, which emphasize adding more features of detail to the word cloud (based on previous work [32] [22] [12]):

- The word cloud needs to use its layout to present semantic information.

- The word cloud needs to support interaction for retrieval of review content.

We have implemented a prototype called the **clustered layout word cloud** to achieve these requirements above. The overview and techniques to accomplish these requirements are detailed in the remainder of this section.

3.2 System Overview

A clustered layout word cloud is a word cloud which embeds semantic information in layout, meaning it encodes detail of the context of the words based on visualization properties, such as spatially. The clustered layout word cloud we implemented also supports clickable interaction which allows the user to conduct keyword searching using words in the visualization. This provides the user the ability to further search reviews that contain keywords that are relevant to them on demand as opposed to conducting manual searches for their keyword.

In our paper, we use *grammatical dependency parsing* NLP approach to process review content for semantic information. The processing result is grammatical dependency graph (GDG). Then we embedded this semantic information into word cloud layout.

Figure 3.1 shows our processing pipeline of how we went from raw user generated review data to the clustered layout word cloud. The pipeline is divided into four parts:

> **Part 1:** A natural language processing technique was applied to the user generated review data in order to process the review content and generate the semantic graph

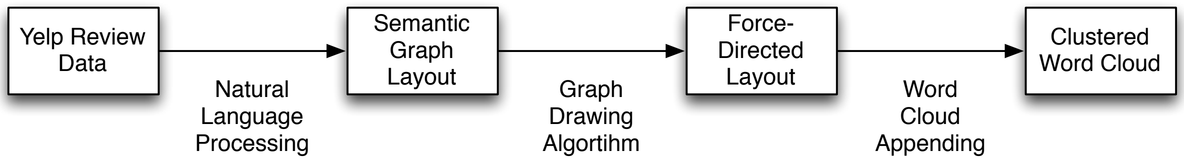

Figure 3.1: Processing pipeline of clustered layout word cloud

layout.

Part 2: From the semantic graph layout, the LinLogLayout energy model layout algorithm was used to create a force-directed graph layout to provide a basis for the clustered layout word cloud.

Part 3: A word cloud generation approach was then applied to the force-directed graph layout in order to display clustering information.

Part 4: Color encoding and clickable interaction were appended to the word cloud to provide faster recognition of clusters and achieve the ability to conduct quick keyword searching respectively.

Figure 3.2 (c) shows the user interface which we used to present our clustered layout word cloud. Section F of the interface shows the clustered layout word cloud. Section G shows the history of keywords in the clustered layout word cloud clicked by the user. In the bottom portion of the interface, Section H, shows the search results of the specific keyword clicked by the user. The keyword clicked by the user is highlighted in search results in section H.

More clustered layout word cloud results are shown in Figure 3.3 and Figure 3.4.

In the following subsections, we explain more of the details to this design.

3.3 Data

In our prototype, we used the Yelp Academic Dataset [33] as our user generated review dataset. The Yelp Academic Dataset provides the business profiles and user review content of 250 businesses, like shopping centers and restaurants, from 30 universities for academic and research exploration. The dataset is includes three objects: business profile objects, user profile objects and review content objects. Yelp provides these objects in JSON format. In our prototype, we loaded these objects to a MySQL database as three tables:

Business Profile Table Business Profile Table stores the basic profile information of local businesses, like restaurants, theater, supermarket. business_id is the primary key of this table, at the same time, it is also ID to query more information from Yelp API online. By this table, we can get all the information of a business, including name, address, telephone number, general rating. price level, geographic information and so on.

User Profile Table User profile data stores basic user information on Yelp. The fields include name, user ID, location and all the public information listed on yelp user page, like photo, url and so on. user_id is primary key in this table.

Review Table Review Table obtains review content and votes information of this specify review content from Yelp users. user_id and business_id are foreigner keys of this table. user_id is associated this review with others by the same user. business_id is associated this review with others of the same business in this dataset. For each review content at yelp, there is one record in this table correspondingly. This record includes the fields of the time of review, reviewer's user id, the reviewed business and the review's feedback. The feedback of review at Yelp is defined as useful, cool and fun. The number of feedback is also recorded in this record.

In this thesis work, I utilized the *Business Profile Table* to select businesses and restaurants and the *Review Table* to process the user generated review content and generate our clustered layout word cloud.

3.4 Nature Language Processing (NLP)

In order to add *grammatical dependency* to our clustered layout word cloud we used the records in a Review Table for each business to obtain semantic information. By pre-processing with NLP tools, review data was extracted and converted to textual graph of key phrases.

To construct the graph, the review content for a specific restaurant was first extracted from the raw dataset and chunked into sentences. Then, the sentences were parsed based on grammatical relations and eventually the relationship information was filtered to form a context graph from the user generated review content for a specific business/restaurant from the Yelp Academic Dataset. We used the Stanford Parser [5] [25] and the OpenNLP toolkits [21] to create the context graphs.

First, we break down each sentence into typed-dependency parses using the Stanford Parser. In Figure 3.5, (a) shows the typed dependency parse for a sampling sentence. (b) shows a filtered sentence level grammatical relations. We filtered all the edges with relations such as *aux*, *auxpass*, *punct*, *det*, *cop*, etc., and also only terms with specific part-of-speech tags are retained, such as *VB*, *VBD*, *VBG*, *VBN*, *NN*, *NNP*, *NNS*, *JJ*, *JJR*, *JJS*, etc. From the sentence-level grammatical relations now available to us, we performed a concatenation of the relations to form a graph for all the reviews of the restaurant the user is querying. In the first part of this process, we extracted the main grammatical relations within a sentence. If relations amongst different sentences share terms (vertices), we connected the terms using these shared vertices. The *GDG* for each restaurant were usually large due to the amount of reviews we had for each restaurant.

In our final step, we assigned weight values to both the vertices and the edges for the semantic graph layout. For assigning weigh to vertices, W_i of term T_i, we use the traditional *IDF* value as described below,

$$W_i(T_i) = \log(N/df_i) \cdot (\log df_i + 1) \tag{3.1}$$

N is the number of sentences, and df_i is the document frequency which denotes the number of sentences that have this term. For weighting the edges, the same strategy for vertex weighting was used, the only difference being that the variable df_j means the number of sentences that have this edge type.

$$W_j(E_j) = \log(N/df_j) \cdot (\log df_j + 1) \tag{3.2}$$

3.5 Force-directed Graph Drawing

Although the *GDG* created provides information about relationships amongst the user generated review set for a specific business, it is dense in its amount of connections, so it cannot be presented clearly. The density of the *GDG* led us to use a force-directed graph layout approach to create a two-dimensional representation of the semantic layout.

The force-directed graph layout is defined based on the force model in physics. The relationship between two entities is defined by force. In our design, the force between any two entities is the weight on the edge of two entities in *GDG*. The greater the weight of an edge, the stronger the relation with two entities, therefore the greater force between two entities, and vice versa; the smaller the weight of an edge, the weaker the relation between the two entities. In a force-directed graph layout, the strength of a relation is represented by the distance between entities; the closer an entity to another, the stronger the relationship between the entities, and, again, vice versa.

Force-directed graph layout has been widely used in visualization presentations. Traditionally, people use the Fruchterman-Reingold model for their force-directed graph layout generation. The process behind this model is based on Hooke's Law and the algorithm iterates until the whole system is stable, which means the termination condition is very slight changes in the whole graph.

Force-directed graph layout is a good model for performing the clustering process in our word cloud design. The process of changing a semantic graph layout to a force-directed graph layout is easy and widely used in [3] [22] and the resulting force-directed graph layout is easy to understand. However, the force-directed graph layout has a scalability problem; the layout computation process is time-consuming when there is a large number of iterations due to the large amount of edges and nodes.

In comparison to the Fruchterman-Reingold model, the energy model process of performing a force-directed graph layout is better. The energy model directly influences the graph layout germinating quality and speed [19]. The essential idea of the energy model is to map the layout to an energy value and this number is related to the optimal goal of whole layout. Then, the iteration algorithm is used to search all of the possible solutions for the lowest energy case of the entire layout. A good layout is considered to have minimal energy [18].

At the same time, the energy model also can provide clustering information based on force-directed graph [19]. In our design, we use the clustering information to encode different colors. Color encoding can help users to distinguish different clusters of word tags in our word cloud.

In our design, we use LinLogLayout open source code [13] as the energy model to create our force-directed graph layouts. The running time of generation process was to our satisfaction level. During our user study, all clustered layout word clouds are generated on-the-fly.

3.6 Word Cloud Visualization and Interaction

The force-directed graph layout is then turned into our clustered layout word cloud. Our design is mainly based on the traditional word cloud generation approach in [8].

However, clustered layout word cloud is based on NLP results and not random placement like in traditional word clouds. Thus, there is one difference with the process of traditional word cloud generation and our word cloud generation: the initial positions of word tags are not random. The initial position of the word tags in our design are provided by the force-directed graph layout, which is generated by LinLogLayout energy model.

In our project, the number of word tags are set manually ($N = 100$). The generation process of the clustered layout word cloud visualization can be described as following steps:

Step 1: Scan the vertex list of GDG and use the weight of the vertices to generate the word list for rendering. High frequency words will be rendered first.

Step 2: For each word in the word rendering list, the initial position is defined by force-directed layout algorithm. Similar to the traditional word cloud approach, the font size is also encoded by the weight of vertex (Equation (3.1)). There is also a check to see whether the word is in collision or overlap with other words. In this check, we use double buffer mask check for the overlap and collision detection. We draw previous results in one buffer of image and then draw word-to-be-placed on the other buffer, then we check the two buffers whether collision or overlap.

Step 3: If collision and overlap occurs, the word follows the Archimedean spiral in order to find the new position to draw, Until it finds a position without collision or overlap to draw the word.

Step 4: Step 2 and 3 are repeated until all the words are drawn.

We used the Java 2D graphics library to implement this clustered layout word cloud appending process. We used color to indicate different clusterings based on the force-directed graph layout. The clustering information is also provided by LinLogLayout [13]. The color is an added dimension of visual stimuli, allowing the user to be able to find distinct clusters faster.

In order to achieve the second design requirements of supporting interaction for review content based on keywords, we enabled all of the words in the word cloud to be clickable to retrieve the review content about the business that contain the clicked word. When the user clicks a word in the word cloud, a section of our interface shows reviews containing that keyword (Section E of Figure 3.2 (b) and Section H of Figure 3.2 (c)), with the keyword highlighted in each of the reviews. The open source search engine, Apache Lucene [17], was used to build an index for review content retrieval.

The white space in our clustered layout word cloud is a limitation. The reason is that our GDG is a very high density and large network. It is difficult to find a proper small 2D panel for all vertex projection. Wu et al presented a seam carving approach to the white space in a semantic word cloud [30]. But, the reason we did not use Wu et al's approach is that we thought the white space can offer the help to distinguish the clusters. In addition, the seam carving approach is an image processing based approach, not a semantic based approach.

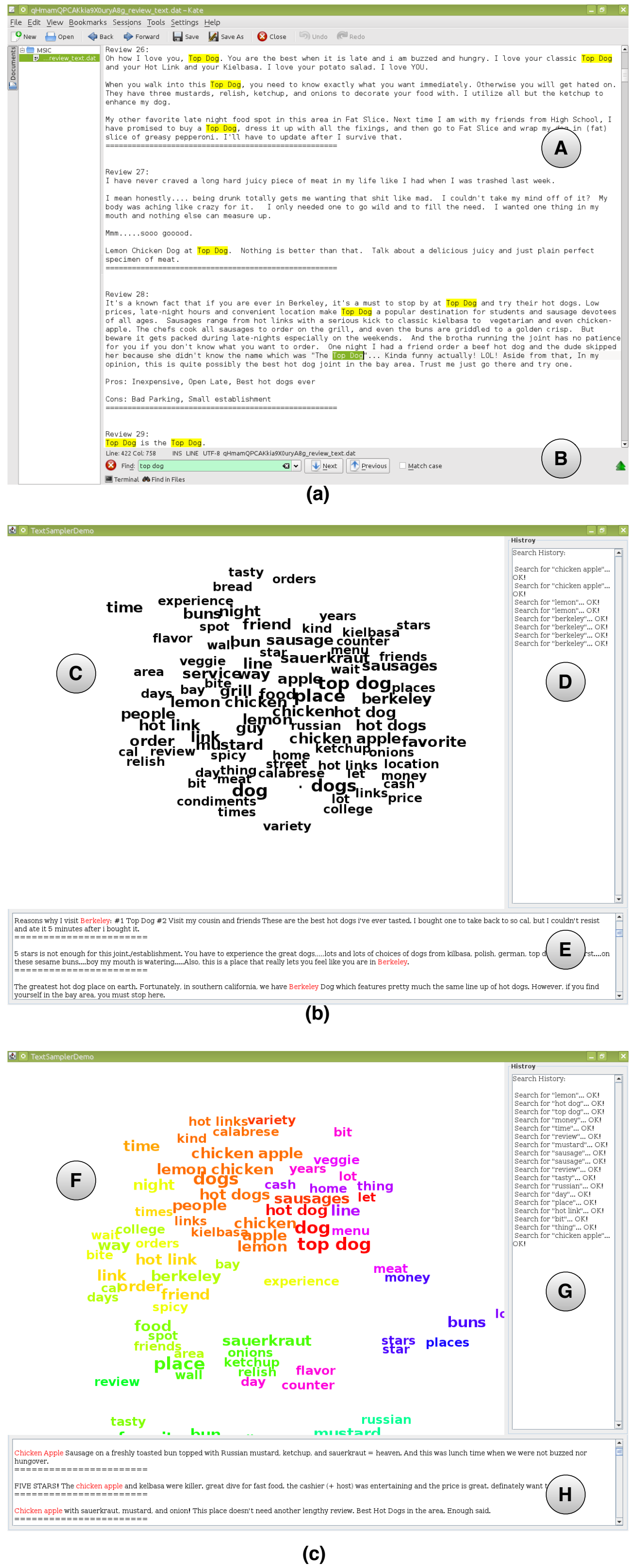

Figure 3.2: (a) Normal Review Reading UI, (b) Normal Random Word Cloud UI, (c) Clustered Layout Word Cloud UI.

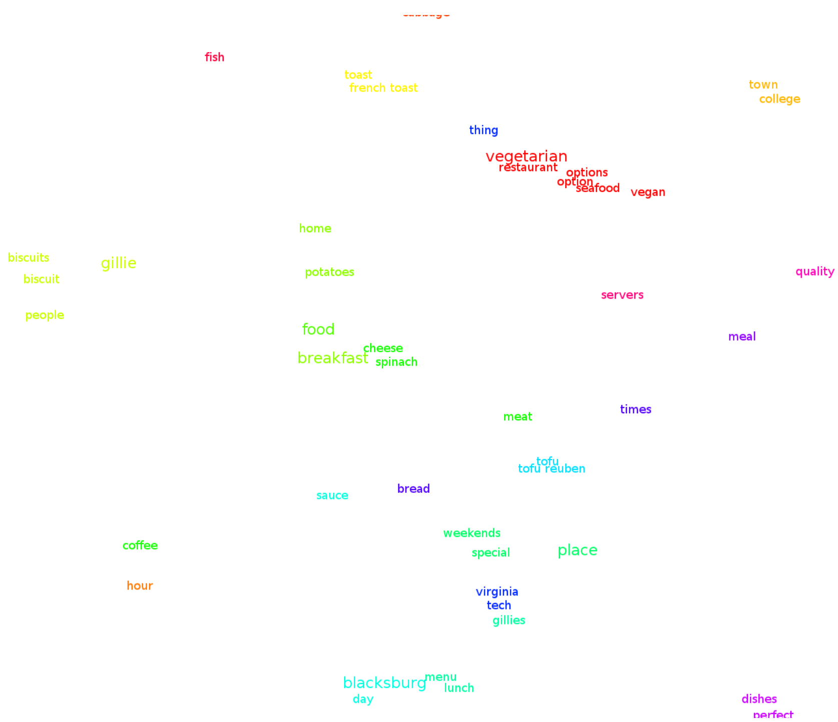

Figure 3.3: Clustered Layout Word Cloud of Gillie's at Blacksburg, VA

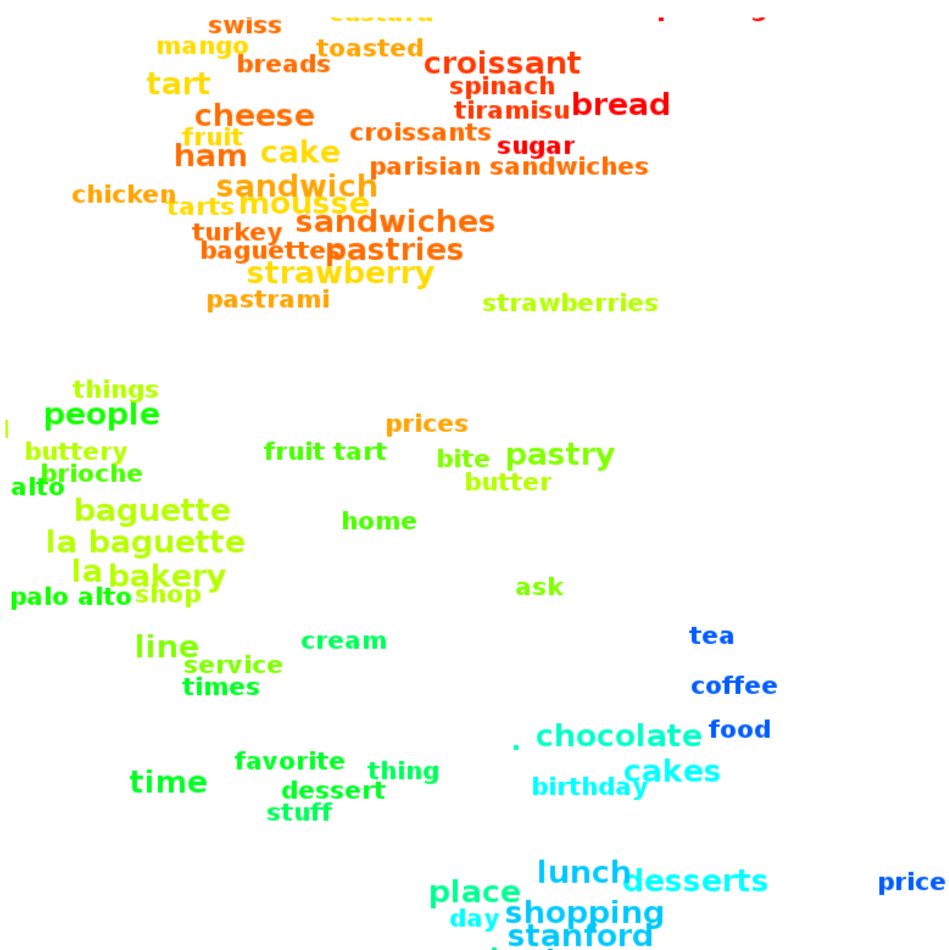

Figure 3.4: Clustered Layout Word Cloud of La Baguette Bakery at Stanford, CA

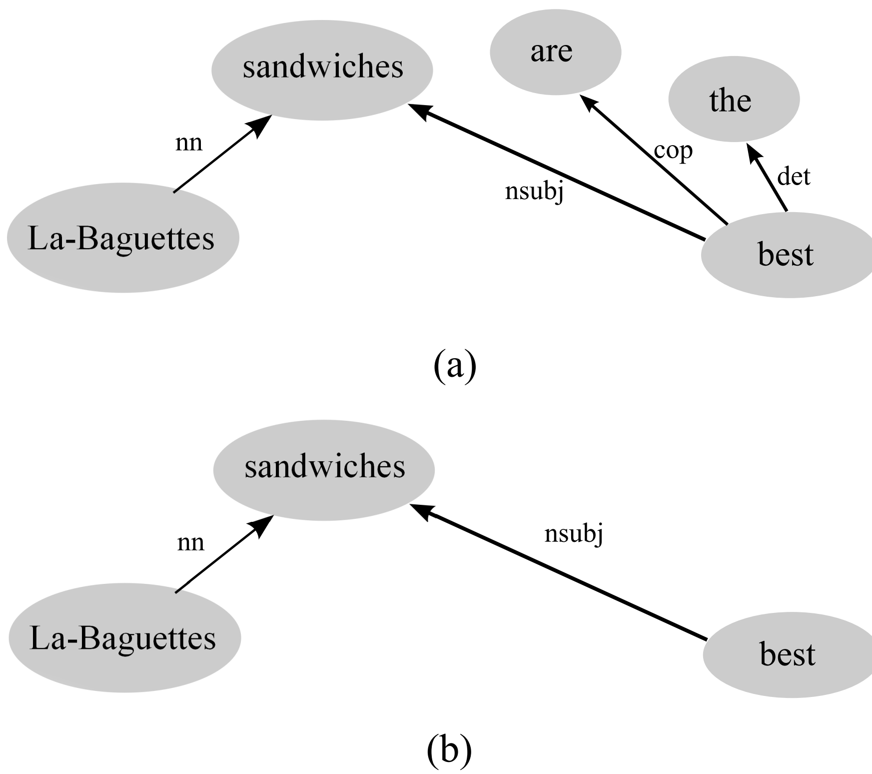

Figure 3.5: (a) A collapsed typed dependency parse for the sentence "La-Baguettes sandwiches are the best!". (b) The sentence level graph transferred from the parse in (a).

Chapter 4

User Study

The major goal of this user study was to assess the effectiveness of a clustered layout word cloud in users' performance. To the best of our knowledge, our work is the first to use real review data to conduct a user study on a non-random layout word cloud.

In this section, we will solve the second research question in Chapter 1.2:

- Will embedded semantic information of GDG layout influence user performance? Specifically, will this word cloud layout, in which words that are close in GDG space are also spatially close in word cloud layout plus clustering color encoding, improve users' performance and experiences? If so, how?

4.1 Participants

We recruited 15 participants (7 female) in our study. Their average age is 24.39, ranging from 20 to 32. All participants were familiar with normal word clouds. And all participants were no color blindness and have normal or corrected-to-normal vision. All participants were native English speakers. They were all undergraduate (4) or graduate students (11) from our university. Their academic majors are in the fields of computer science, mathematics, electrical engineering and veterinarian science. The study lasted about 50 minutes and each participants were compensated with $30 cash.

4.2 Apparatus

All the tasks are performed at a Lenovo Thinkpad T61 laptop with Intel Centrino 2.10GHz CPU and 4 GB memory with Kubuntu 12.04. The keyboard and mouse were external keyboard and mouse, like a normal desktop. The display used was one Dell 19-inch LCD monitor with 1280×1024 resolution. The users were able to switch between multiple desktop views. Task completion time was measured using a stop watch. After performing the visualization tasks, participants completed a TLX-based Likert-style questionnaire and survey on a Lenovo Thinkpad T400s laptop.

4.3 Review Reading Techniques

During this user study, we also compared the clustered layout word cloud review reading technique to two other alternative techniques: normal review reading and normal random word cloud. The normal review reading technique was used because this is commonly how users read reviews in their daily life and the normal random word cloud technique was used because it doesn't include clustered layout. Thus, both of them are good control groups in our study.

4.3.1 Normal Review Reading (RR)

In this technique, we listed all of the textual user generated review content in one pair of restaurants in a normal text editor where user could switch freely between two restaurants' review content if needed (Figure 3.2 (a)). The overall rating for each review was omitted. User can scroll up and down to check all review content and use the search box in the text editor to search the keywords they want to refine their search with and the keyword will be highlighted in the text.

4.3.2 Normal Random Word Cloud (RW)

In this technique, we used the random word cloud layout to generate normal word cloud (Figure 3.2 (b)). The user can click the word they feel is a keyword in their search and review content with this keyword in it will be displayed in Section E in Figure 3.2 (b).

4.3.3 Clustered Layout Word Cloud (CL)

In this technique , the user used the Clustered Layout Word Cloud to read reviews (Figure 3.2 (c)). The user can click the word they feel is a keyword in their search and review content with this keyword in it will be displayed in Section H in Figure 3.2 (c).

In Figure 3.2, we show an example of the three review reading techniques for the same Yelp review data. (a) is Normal Review Reading, (b) is Normal Random Word Cloud, and (c) is Clustered Layout Word Cloud.

4.4 Tasks

The two types of tasks we used to assess users' performance were a decision making task and a feature finding task. Both tasks mirror events that happen in daily life:

1. The decision making task occurs when people are comparing similar restaurants. This task is overview task where a restaurant can be distinguished.
2. The feature finding task occurs when people attempt to find non-quantitative detail., This task is a specific information retrieval task where more detail can be located depending on the specific restaurant of their choice.

For each of the review reading techniques we will let users to do two types of tasks described below:

4.4.1 Decision Making Task

In this task, we let users perform different review reading techniques to make a decision about which restaurant he or she will go to from the pair of restaurants provided in the test. The users will use same review reading technique for two pairs during the test, resulting in six decision making task tests.

We selected 6 pairs of restaurants for this type of task. Three of the pairs were good-good restaurant pairs, meaning the restaurants in these pairs had high ratings (4 or 5 stars on Yelp Business Profile Table). The other three restaurant pairs were good-bad pairs, meaning the restaurants in these pair had significant differences in their rating: ”good” was high(4 or 5 stars) and ”bad” was low(1 or 2 stars). Participants are unaware of how the good and bad restaurants were chosen. All the restaurants in the same pair had similar price level, similar location, and similar cuisine. In order to make counter-balanced ordering, intra-test procedures were changed between participants, meaning one participant would conduct the task on a good restaurant and then a bad restaurant during a good-bad test, and the next participant would conduct the task on the restaurants in opposite order (bad-good ordering).

4.4.2 Feature Finding Task

In this task, we want to evaluate how people find the features (food and non-food) of restaurants. The features we defined for this task are food and non-food features, non-food features being for example the restaurant’s environment or service. For the 3 review reading techniques, we use each technique to complete the two distinct feature finding tasks. One is to find the food-relevant feature, another one is to find non-food relevant feature. For the 6 tests conducted we provided 6 different restaurants to the participant. The 6 restaurants had high ratings (4 or 5 stars) similar to the "good" restaurants in the decision making task. The ordering of the review reading technique used for the task is randomized for each user, each technique had the participant locate the food and non-food feature though.

During the feature finding task we imposed a time limit on half of users, the other half of user will did not have time limit. The time limit for each of the 6 tests was 2 minutes. All tasks are shown as Figure 4.1.

4.5 Hypotheses

Based on our research questions we present in Introduction Chapter, we formulated the two hypotheses listed as below.

[**H1**] *CL leads to better user performance than a RW and RR for decision making and feature finding on user generated reviews.*

The assumption is that, based on the *CL* which is embedded with NLP results, we expect users to obtain more semantic information from the clustered layout compared to the other forms of presenting review content. We expected the semantic information and clustering will also influence the user's task completion time as well as accuracy of decision making and feature finding.

Task Type	Approach	Data	Task
Decision Making	Normal Review Reading	Good Good Pair 1	Which restaurant will you go?
	Normal Review Reading	Good Bad Pair 1	
	Normal Random Word Cloud	Good Good Pair 2	
	Normal Random Word Cloud	Good Bad Pair 2	
	Clustered Layout Word Cloud	Good Good Pair 3	
	Clustered Layout Word Cloud	Good Bad Pair 3	
Feature Finding	Normal Review Reading	Restaurant 1	Food Feature
	Normal Review Reading	Restaurant 2	Non-food Feature
	Normal Random Word Cloud	Restaurant 3	Food Feature
	Normal Random Word Cloud	Restaurant 4	Non-food Feature
	Clustered Layout Word Cloud	Restaurant 5	Food Feature
	Clustered Layout Word Cloud	Restaurant 6	Non-food Feature

Figure 4.1: Task Design and Study Procedure

[**H2**] *CL has a positive impact on user satisfaction in the decision making and feature finding tasks on user generated review.*

It is expected that user satisfaction feedback will show that *CL* will be more preferred amongst participants than RW and RR.

4.6 Procedure

The study was a within-subject design study, meaning that each user will use the three different review reading technique to complete the two tasks. Before the study started, we will let user to familiar with three different techniques to use same demo data set.

In order to get quantitative and qualitative feedback, we choose to evaluate how the clustered layout word cloud influence user's decision on the restaurant choice. The evaluation composed of task completion time, error rate, TLX-based Liker-style questionnaire [9], user's preference ranking and qualitative feedback.

For the decision making task, in each trial, we measured completion time and error rate of each task. We assume the performance of normal review reading as the baseline in our user study. For the feature finding task, only task completion time was recorded for each trial of participants who are in no time limit group.

After each review reading technique trial, we let users complete the TLX-based Likert-sytle questionnaire [9] to evaluate their feedback in mental demand, physical demand, and other metrics to measure task difficulty levels. After each of the two task sessions, the user provided a ranking of preference between the three review reading techniques. At the conclusion of the two task sessions, the user gave a general evaluation for each review reading technique.

Chapter 5

Results and Analysis

In our user study, we conducted two types of tasks, a decision making task and a feature finding task, to evaluate three different review reading techniques: *Normal Review Reading (RR)*, *Normal Random Word Cloud (RW)* and *Clustered Layout Word Cloud (CL)*.

5.1 Decision Making Task Results

In the decision making tasks, users evaluated 6 pairs of restaurants and choose which one they felt was the best.

5.1.1 Task Completion Time

The Shapiro-Wilk normal distribution test evaluated the task completion time (in seconds) on good-good pair restaurants group (p=0.09) and good-bad pair restaurants (p=0.43). p >0.05 means that data follows normal distribution.

We ran repeated measured ANOVAs on *task completion time* in for the decision making task. We found that *restaurant pairing* had a significant effect ($F_{1,14}$=52.465, p <0.001) on *task completion time* . We also found there is no significant effect of *review reading techniques* on *task completion time* .

We ran a post-hoc one way ANOVA for good-good pair restaurant groupings and good-bad restaurant groupings. We found that the *review reading technique* have a significant effect on *task completion time* ($F_{2,42}$=3.157, p=0.05) in good-good pairing of restaurants, but no significant effect in good-bad pairing of restaurants ($F_{2,42}$=0.253, p=0.78). From Figure 5.1, we can found that participants spent less time in the decision making task using *CL* compared to the other two techniques when participants are presented good good pairings. In Figure 5.2, it is shown that their is not a significant difference in *task completion time* amongst the three techniques when participants are presented good bad pairings.

5.1.2 Error Rate

In Decision Making Task, the error rate is calculated in good-bad pairings and we assume the correct answer is the good restaurant in good-bad pair of restaurants. Based on the

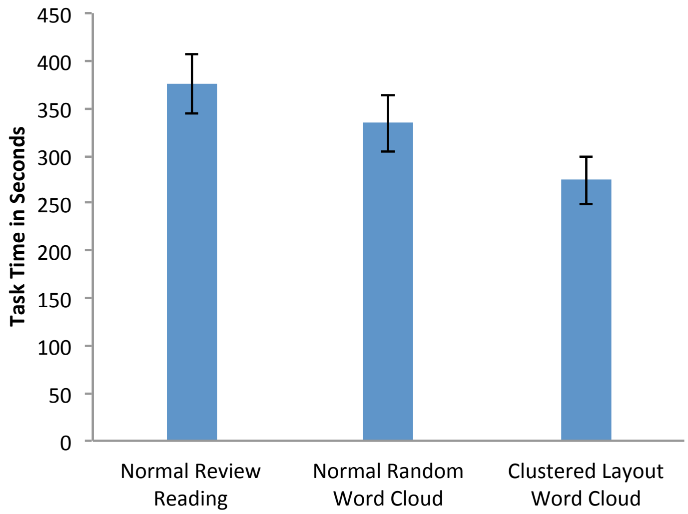

Figure 5.1: Decision Making Task Completion Time in Good-Good Pair

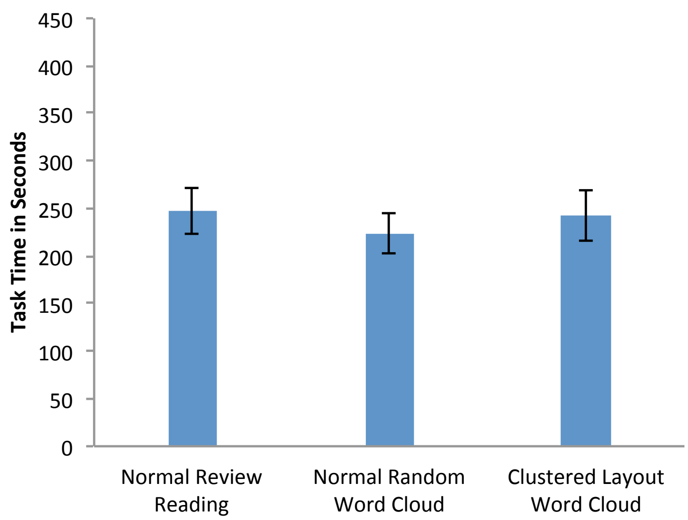

Figure 5.2: Decision Making Task Completion Time in Good-Bad Pair

decision making tasks, the error rates are as followed: *RR* (0%), *RW* (6.7%, 1 error in 15 trials) and *CL* (6.7%, 1 error in 15 trials).

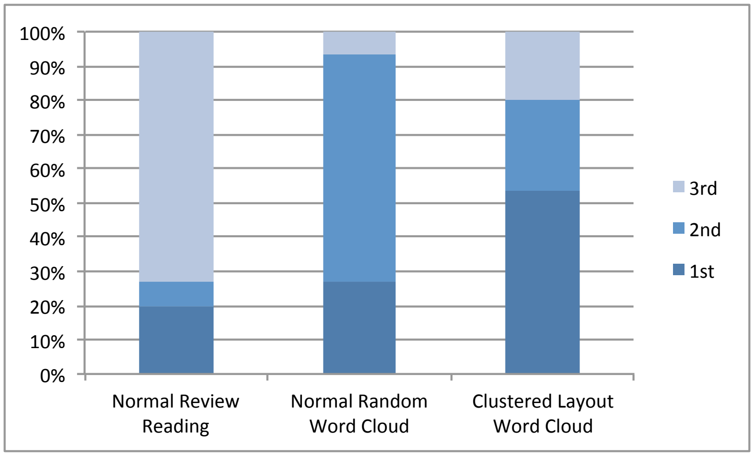

Figure 5.3: Preference ranking in *Decision Making Task*: The percentage of participants who ranked each review reading technique in *Decision Making Task* in overall preference. Note: 1st means most preferred and 3rd means least preferred

5.1.3 Preferential Ranking

We asked participants to provide an overall preference ranking of reading approaches, 1 being most preferred and 3 being least preferred. The preferential ranking was analyzed with a Friedman test to evaluate differences in median rank across three techniques. The test showed a significant difference between the three based on the preferential ranking (χ^2(2, N=15)=6.533, p=0.04). The follow-up pairwise Wilcoxon tests found that *RR* had a significantly lower preference ranking than *CL* (p=0.05) and *RW* (p=0.04). There is no significant difference in preference ordering between *CL* and *RW*. The description results are shown in Figure 5.3.

5.2 Feature Finding Task Results

In this task, we divided the participants into two groups. One group had a two minute time limit where the other group did not have a time limit. We used six restaurants for this task. Restaurant # 1,3 and 5 were used for food-feature finding. And Restaurant # 2,4 and 6 were used for non-food feature finding. The goal of this task was to replicate a situation similar with our daily life events and get users' feedback.

5.2.1 Task Completion Time

We ran a one way ANOVA for task completion time for those participants who did not have a time limit. The time data also passed the Shapiro-Wilk normal distribution test (p=0.43). We did not find any significant effect on task completion time between reading approaches ($F_{2,42}$=0.253, p=0.78), as shown in Figure 5.4.

5.2.2 Preferential Ranking

We asked participants to provide an overall preference ranking of reading approaches, 1 being most preferred and 3 being least preferred. The preferential ranking was analyzed with a Friedman test to evaluate differences in median rank across three techniques. The test, again, showed a significant difference between the three based on the preferential ranking (χ^2(2, N=15)=8.133, p=0.02). The follow-up pairwise Wilcoxon tests found that *CL* had a significantly higher preference ranking than *RR* (p=0.02) and *RW* (p=0.005). There is no

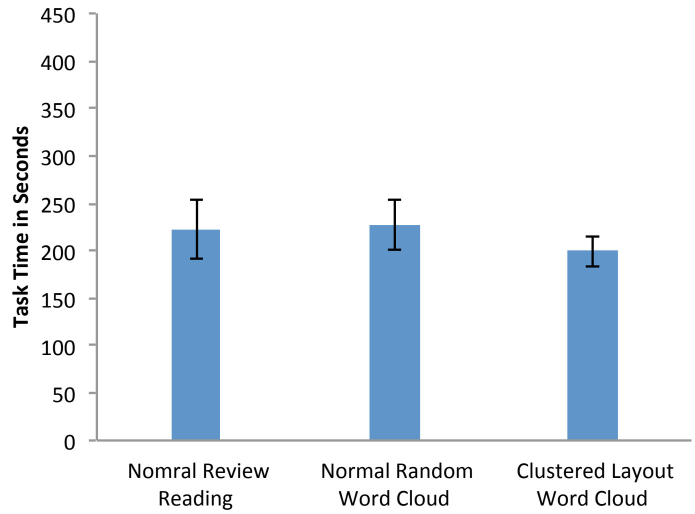

Figure 5.4: Feature Finding Task Completion Time

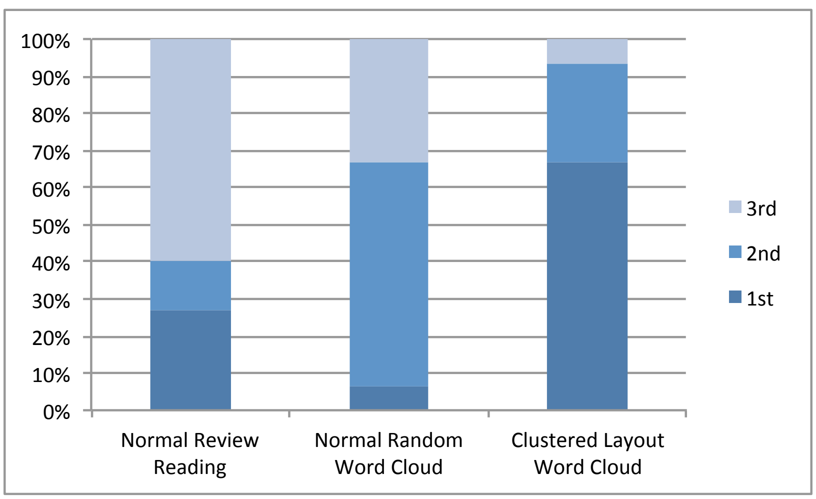

Figure 5.5: Preference ranking in *Feature Finding Task*: The percentage of participants who ranked each review reading technique in *Feature Finding Task* in overall preference. Note: 1st means most preferred and 3rd means least preferred

significant difference in preference between *RR* and *RW*, shown in Figure 5.5.

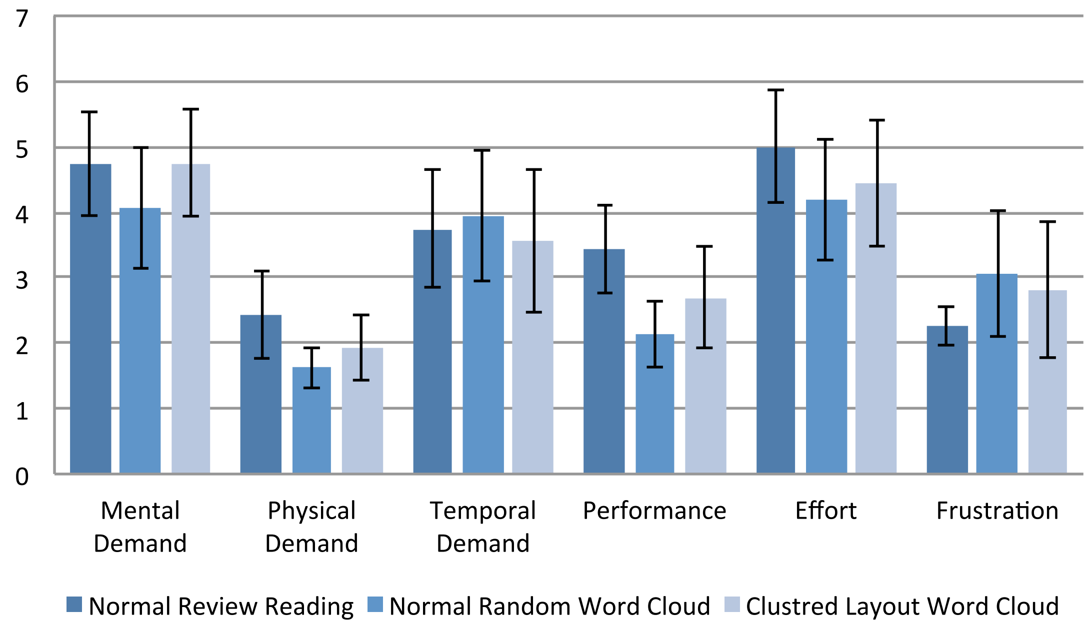

Figure 5.6: TLX-based Liker-style questionnaire results without time limit (where lower is better)

5.2.3 User Satisfaction Level without Time Limit

The TLX-based Likert-style questionnaire was used to acquire feedback of the three *review reading techniques* from participants whom did not have a time limit imposed on them during the feature finding task. A Friedman test was conducted to observe any differences in scores in the questionnaire and as shown in Figure 5.6, the test results showed there was no significant difference in any of the responses.

5.2.4 User Satisfaction Level with 2 Min Time Limit

The same TLX-based Likert-style questionnaire was used to acquire feedback of the three *review reading techniques* from participants who did have a time limit (2 min) imposed on them during the feature finding task. Friedman test was conducted to observe any differences in scores in the questionnaire. The test results showed a significant difference in *mental demand* (χ^2(2, N=8)=11.826, p=0.003), *physical demand* (χ^2(2, N=8)=6.5, p=0.04), *temporal demand* (χ^2(2, N=8)=10.129, p=0.006) and *effort* (χ^2(2, N=8)=7.00, p=0.03) (Figure 5.7).

The follow-up pairwise Wilcoson tests showed: *mental demand*: *CL* is significantly lower than *RR* (p=0.02) and *RW* (p=0.03); *physical demand*: *CL* is significantly lower than *RR* (p=0.04); *temporal demand*: *CL* is significantly lower than *RR* (p=0.01); *effort*: *RR* is significantly higher than *RW* (p=0.04) and *CL* (p=0.02).

5.3 Qualitative Analysis

5.3.1 Semantic Information Retrieval

Based on user feedback, we found the semantic information, provided by natural language processing in layout, improved user performance in decision making.

It is an easy way to navigate through several reviews that use similar terminology

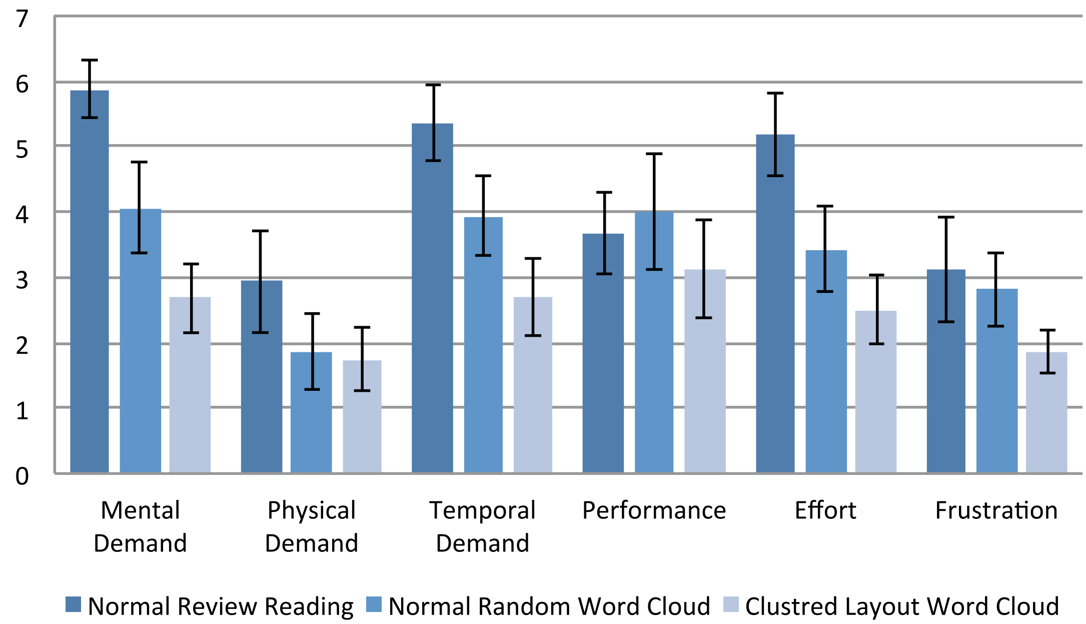

Figure 5.7: TLX-based Liker-style questionnaire results with 2 minutes time limit (where lower is better)

> *to pinpoint specific aspects of a restaurant. It is easier to find out what I'm looking for. (Subject 14)*

> *Finding the keywords for 'service' or 'sandwich' was made easiest with the color coded clouds. It was easy to pick out the keywords that I needed to look at to make my decisions.(Subject 9)*

The visual aspects of font size and color had positive impact on user's review reading process by using clustered layout word cloud as well.

> *The size of the words also made it easier to know what was important/more used*

> *in the reviews. (Subject 7)*

> *It makes it much easier to look for keywords that help when deciding on which option to pick. It's well organized and groups similar words and distinguishes them by color. The black and white words with no grouping make it difficult to find tags. (Subject 15)*

5.3.2 Keywords Query by Interaction

Through user feedback, the clickable interaction to query keywords in review content was found useful.

> *From these three context preserving could saves my time since I can click on the things I am interested in and quickly see them highlighted in the reviews (Subject 2)*

> *I still liked the context preserving tag cloud over all the other techniques because it helped me find better keyword to search so I could read more details in the actual reviews. (Subject 9)*

5.3.3 Natural Language Processing

There was mixed reviews on factors that were determined to be attributed to our natural language processing:

Users feel not all the necessary word tags were presented in word cloud generated by natural language processing.

> *I would have ranked the tag clouds higher, but I was unable to finish one of the tasks because there were no tags regarding the quality of service at a restaurant. Normally, I liked the context preserving tag cloud more, but I got the impression that fewer tags were included. I liked the clustering, but sometimes couldn't tell why terms were included in specific clusters. (Subject 1)*

In few test cases, users felt the clustering and font size generated by natural language processing was not clear enough for them to complete their tasks quickly.

> *I think that after I read one or two of the words in one cluster, and found them irrelevant, I would move ahead to another cluster to see if it was more relevant. I remember that it was difficult to find the words that I was actually looking for (e.g., "service" was really small one one of the word clouds, and it was partially off the screen in another). (Subject 8)*

People want to assess the personality of individual reviewers based on their complete review and the word cloud generated based on the natural language processing used did not provide specific information about individual reviewer's personalities.

> *I preferred reading full reviews because I felt like I could better understand the personality and interests of the reviewers, which factors a great deal into the way*

I interpret the quality and reliability of the review.(Subject 8)

Chapter 6

Discussion

Based on the results of our user study, we can verify that both of hypotheses are supported.

6.1 Better Users' Performance

In our decision making task, we found that the clustered layout word cloud had a significant lower task completion time than the normal random layout word cloud in good-good pairs of restaurants (p=0.05). In Figure 5.1, it shows that CL is significantly faster than RR (27% on average time) and RW (18% on average time).

Good-bad pairs are easy to distinguish in all three techniques. However, because of good-good pairs have similar high general rating, they need more context information to support users' decision. In order word, users need to spend more time to find evidences. In this situation, the clustered layout word cloud with semantic information can offer more context information to users and enhance users' performance.

The error rate of the two word cloud layouts was the same, 6.7%, and users preferred the clustered layout word cloud over the random layout word cloud. In our natural language processing techniques and word cloud visualization approach, there is word or content discrimination and bias. Through the experiment results, we can see that there are no significant influence the error rate.

Thus, the experiment results support **[H1]** in respect to our decision making task.

6.2 Better User Experience

In our decision making task, we saw that participants significantly preferred the Clustered Layout Word Cloud and Normal Random Word Cloud more than Normal Review Reading. In our feature finding task, the user preference ranking of Clustered Layout Word Cloud was significantly higher than that of Normal Review Reading (p=0.024) and Normal Random Word Cloud (p=0.005).

In our time-limited feature finding task test, we found that the user's satisfaction levels of the clustered layout word cloud test was significantly less than that of the normal random word cloud test in *mental demand, physical demand, temporal demand* and *effort*. However, there was no significant different in feature finding task without time limit. We think the reason is our task design. The idea behind the feature finding task is to perform the categorization and clustering process in our mind. It was assumed that the clustered layout word cloud would help users perform the clustering process better.

Based on user feedback from our two tasks, qualitative and quantitative, [**H2**] is fully supported.

6.3 User Strategies

Based on our observation and users' qualitative feedback, we found that users have different strategies to use CL, RR and RW for review reading tasks.

In RR, users followed their normal reading process for feature finding: reading reviews documents, finding a potential keyword, using Find function to highlight all of this keyword in all reviews, checking all the occurrences, finding another potential keyword, and repeating previous steps. In this strategy, users try to guess the keywords based on their own review reading experience and understand all reviews in short time by these potential keywords they guessed.

In RW, users firstly scanned the entire word cloud and found the potential keywords for retrieval. After clicking the keyword, they read the reviews which contain the keywords. Then, users clicked another relevant keyword for review reading.

In CL, users scanned all clusters of word cloud, chose a cluster to focus on, and clicked the keywords. The rest of steps are similar with RW. The only difference is that CL has the clustering information to guide users for review reading.

The advantages of CL are: First, it provides the clustering information to guide users for the keywords targeting. However, in RR and RW, users need to spend time searching the potential keywords for retrieval. Second, for CL and RW, the clickable query speeds up the switch process between one keyword review reading to another keyword review reading.

6.4 Usability Problems

Based on users' feedback, we found that some users have mixed reviews on NLP results. First, some users felt that not all the necessary word tags were presented in word cloud generated by NLP. Second, some users thought the keywords generated by NLP cannot provide clear clustering for review reading tasks. Third, one users said she preferred to understand the review content based on the reviewer's personality. But the keywords provided by NLP cannot aid her to complete this type of task.

Despite mixed reviews given on the natural language processing technique we use, statistical analysis results show that error rate was low. If our NLP algorithm was improved, we hope that content will be more vivid for users; all the necessary word tags will be shown, clustering and font size will be more accurate, and personality information will also be available in the word cloud.

Another issue is that the review writing styles of some reviews on Yelp are different from normal review articles on newspaper or books. There are web acronyms and internet symbols in internet review contents. Dent et al introduce a new NLP approach to process micro-text, like Twitter and Facebook updates, for the new internet writing style [6].

Chapter 7

Conclusion and Future Work

7.1 Conclusion

In this thesis, we present one new visualization technique: semantic based clustered layout word cloud. We used a *grammar dependent parsing* natural language processing technique to get a semantic graph of user generated review content. The energy based force directed graph layout algorithm was applied to the semantic graph to create a force directed graph layout of the content. Based on this layout, we used a word cloud visualization approach on the force directed graph layout to generate clustered layout word cloud. In order to have better user experience, we appended interaction functionality and clustering color encoding to the word cloud. The intention of the clustered layout word cloud was to reduce review reading time, gain more insight and improve user experience.

We also ran a user study to evaluate how the clustered layout word cloud would help users gain more information, task performance, error rate and user satisfaction level. We used the Yelp Academic Dataset and designed two types of tasks for our evaluation process: a decision making task and a feature finding task. The user study results supported our hypotheses that the clustered layout word cloud can provide better user performance in decision making task with good-good pairs of restaurants and let users have better satisfaction level.

In conclusion, clustered layout word clouds can provide users with an efficient and more satisfactory experience during their review reading process.

7.2 Future Work

In the future, we plan to append more information on the clustered layout word cloud, like time-series restaurant reviews and sentiment analysis of review information. We will also apply a more sophisticated NLP technique for processing the review content data as well as enable a search box functionality for finding words easier within the word cloud.

Then, by manipulating the NLP algorithm, we will also try to expose keywords that currently does not appear in the clustered layout word cloud and therefore provide a better more customizable and possibly interactive review reading experience for the users.

Based on the observation results of our user study, users preferred to bring relevant word tags together in order to get clusterings for understanding review contents. We will combine different types of similarities, like lexical distance and syntactic distance, for text visual analytics.

Furthermore, another direction of future work is to add more interaction throughout the whole pipeline, shown as Figure 3.1 in Chapter 3. Users can modify the NLP results, visualization results, and even raw data by interactions in order to process the customized visual analytics results they need. We also plan to use more than one type of data mining or NLP technology to process the data and get the different levels or context visualization results. Users' feedback can steer the algorithm to produce more customized word clouds.

Bibliography

[1] Scott Bateman, Carl Gutwin, and M. Nacenta. Seeing things in the clouds: the effect of visual features on tag cloud selections. In *Proceedings of the nineteenth ACM conference on Hypertext and hypermedia*, HT '08, pages 193–202, Pittsburgh, PA, USA, 2008. ACM.

[2] Giuseppe Carenini and Lucas Rizoli. A Multimedia Interface for Facilitating Comparisons of Opinions. In *IUI '09: Proceedings of the 14th international conference on Intelligent user interfaces*, pages 325–334, 2009.

[3] Weiwei Cui, Yingcai Wu, Shixia Liu, Furu Wei, Michelle Zhou, and Huamin Qu. Context-Preserving, Dynamic Word Cloud Visualization. *IEEE Computer Graphics and Applications*, 30(6):42–53, 2010.

[4] Aron Culotta and Jeffrey Sorensen. Dependency tree kernels for relation extraction. In *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics - ACL '04*, pages 423–es, Morristown, NJ, USA, 2004. Association for Computational Linguistics.

[5] Christopher Manning de Marneffe, Marie-Catherine, Bill MacCartney. Generating typed dependency parses from phrase structure parses. In *Proceedings of the 5th International Conference on Language Resources and Evaluation*, pages 449—-454, 2006.

[6] Kyle Dent and Sharoda Paul. Through the twitter glass: Detecting questions in micro-text. In *Workshop on Analyzing Microtext at the 25th AAAI . . .*, 2011.

[7] Economist Opinion Cloud. http://infomous.com/site/economist/.

[8] Jonathan Feinberg. Wordle. In *Beautiful Visualization*, chapter 3. 2009.

[9] SG Hart and LE Staveland. Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. In P. A. Hancock and N. Meshkati, editor, *Human mental workload*. North Holland Press, Amsterdam, 1988.

[10] Jeff Huang, Oren Etzioni, Luke Zettlemoyer, Kevin Clark, and Christian Lee. RevMiner. In *Proceedings of the 25th annual ACM symposium on User interface software and technology - UIST '12*, page 3, New York, New York, USA, 2012. ACM Press.

[11] Owen Kaser and Daniel Lemire. Tag-Cloud Drawing: Algorithms for Cloud Visualization, 2007.

[12] Kyle Koh, Bongshin Lee, Bohyoung Kim, and Jinwook Seo. ManiWordle: Providing Flexible Control over Wordle. *Visualization and Computer Graphics, IEEE Transactions on*, 16(6):1190–1197, November 2010.

[13] LinLogLayout. http://code.google.com/p/linloglayout/.

[14] Bing Liu and South Morgan Street. Opinion Observer : Analyzing and Comparing Opinions on the Web. In *WWW '05 Proceedings of the 14th international conference on World Wide Web*, pages 342–351, 2005.

[15] Shixia Liu, Michelle X Zhou, Shimei Pan, Yangqiu Song, Weihong Qian, Weijia Cai, and Xiaoxiao Lian. TIARA : Interactive , Topic-Based Visual Text Summarization. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 3(2), 2012.

[16] Steffen Lohmann, Jürgen Ziegler, and Lena Tetzlaff. Comparison of Tag Cloud Layouts: Task-Related Performance and Visual Exploration. volume 5726 of *Lecture Notes in Computer Science*, book part (with own title) 43, pages 392–404. Springer Berlin / Heidelberg, 2009.

[17] Lucene. http://lucene.apache.org/.

[18] Andreas Noack. Energy Models for Graph Clustering. *Journal of Graph Algorithms and Applications*, 11(2):453–480, 2007.

[19] Andreas Noack. Modularity clustering is force-directed layout. *Physical Review E*, 79(2), February 2009.

[20] Chris North. Toward measuring visualization insight. *IEEE Computer Graphics and Applications*, 26(3):6–9, May 2006.

[21] OpenNLP. http://opennlp.apache.org/.

[22] FV Paulovich, F Toledo, and GP Telles. Semantic Wordification of Document Collections. *Computer Graphics*, 31(3pt3):1145–1153, June 2012.

[23] A. W. Rivadeneira, Daniel M. Gruen, Michael J. Muller, and David R. Millen. Getting our head in the clouds: toward evaluation studies of tagclouds. CHI '07, pages 995–998, San Jose, California, USA, 2007. ACM.

[24] Johann Schrammel, Michael Leitner, and Manfred Tscheligi. Semantically structured tag clouds: an empirical evaluation of clustered presentation approaches. CHI '09, pages 2037–2040, Boston, MA, USA, 2009. ACM.

[25] Stanford Parser. http://nlp.stanford.edu/software.

[26] Fernanda B Viégas, Martin Wattenberg, and Jonathan Feinberg. Participatory visualization with Wordle. *IEEE transactions on visualization and computer graphics*, 15(6):1137–44, 2009.

[27] Mengqiu Wang, NA Smith, and Teruko Mitamura. What is the Jeopardy model? A quasi-synchronous grammar for QA. In *Proc. of EMNLP-CoNLL*, 2007.

[28] Word Cloud in Flickr.com. http://www.flickr.com/photos/tags/.

[29] Word Cloud in News. http://www.globalnews.ca/word+cloud/6442706583/story.html.

[30] Yingcai Wu, Thomas Provan, Furu Wei, Shixia Liu, and Kwan-Liu Ma. Semantic-Preserving Word Clouds by Seam Carving. *Computer Graphics Forum*, 30(3):741–750, June 2011.

[31] Yingcai Wu, Furu Wei, Shixia Liu, Norman Au, Weiwei Cui, Hong Zhou, and Huamin Qu. OpinionSeer: interactive visualization of hotel customer feedback. *IEEE transactions on visualization and computer graphics*, 16(6):1109–18, 2010.

[32] Koji Yatani, Michael Novati, Andrew Trusty, and Khai N. Truong. Review spotlight: a user interface for summarizing user-generated reviews using adjective-noun word pairs. In *Proceedings of the 2011 annual conference on Human factors in computing systems - CHI '11*, page 1541, New York, New York, USA, 2011. ACM Press.

[33] Yelp Academic Dataset. http://www.yelp.com/academic_dataset.