Multi-level Comparative Evaluation on Multiple sources of User-generated Reviews

Joseph Cilli
School of Computing
Information Sciences
Florida International University
Miami, FL, USA
cillij@fiu.edu

Wei Xue
School of Computing
Information Sciences
Florida International University
Miami, FL, USA
wxue004@fiu.edu

Tao Li
School of Computing
Information Sciences
Florida International University
Miami, FL, USA
taoli@cs.fiu.edu

ABSTRACT

User-generated comments and reviews for hotels on the web are an important information source for hoteliers and for travel planners. In addition, understanding and reacting to these comments are important for quality control. To get a comprehensive and unbiased view of the user comments, one should be aware of the differences of the data sources. In this paper, we analyzed three review hosts for hotels: TripAdvisor.com, Yelp.com and Orbitz.com, and proposed a systematic framework to answer the research query of what are the variations in user comment data across different websites. The comparative study shows a three level difference from various statistics: basic statistic level, rating star level and semantic level. The results provide researchers insight into the consumer review data from multiple sources.

Categories and Subject Descriptors

H.4 [**Information Systems Applications**]: Miscellaneous; D.2.8 [**Software Engineering**]: Metrics—*complexity measures, performance measures*

General Terms

Applciation, Experimentation

Keywords

Data analysis, Opinion mining, Hotel reviews

1. INTRODUCTION

Online Travel Agents (OTA) have gained in popularity over the last decade and have become the primary source of consumer travel research and purchasing. This recent growth of can be attributed to both social media applications and mobile technologies. In addition, a 2010 eTRAK study examined the top 30 hotel brands and concluded that 57% of reservations were made online. This was a 19% increase from 2006 [?]. A recent ComScore study found that 37% of smartphone users had accessed travel related sites from their phones for research; whereas, one in five actually made a reservation with their mobile device[?]. Projections for 2012 indicated that there would be over 117 million online travel researchers and over 98 million reservations will be made online while traditional Global Distribution System reservations continued to decline to its lowest point. One of the most well-known user-generated comment sites is TripAdvisor.com [?], with nearly 30 million visitors monthly [?]. Most of this can be directly attributed to consumer online interaction[?, ?]. The growth and wide use of social media and mobile web applications fostered this evolution. Although many hoteliers began marketing campaigns across multiple sales channels in an effort to produce greater returns, the majority of consumers are less willing to believe advertising and more willing to trust peer reviews. 90% of consumers trust peer recommendations, while 70% trust online comments left by strangers; whereas, only 14% trust advertisements [?]. This would suggest that user-generated comments left by strangers and websites have significantly more value than company advertising campaigns; moreover, social networking services like Facebook may offer more significance because potential reviews are generated by peers or ĂIJfriendsĂİ. [?] determined that 97.7% of TripAdvisor.com users were influenced by other traveler comments. In addition, there exists a large financial impact for these websites to be viewed as the best. OTAĂŹs receive a financial commission for every reservation originating from their website. In short, if their website is viewed as providing credible, helpful reviews, then more people will frequent their website and ultimately increase the potential amount of reservations being generated from that website.

To understand this massive review data generated by users on the website, many research topics have been proposed, such as sentiment analysis, helpfulness analysis. However, most of these work only focus on one single data source, either hotel reviews crawled from TripAdvisor or product reviews from Amazon[?, ?, ?]. It is likely that the research results are biased by the data set. To the best of our knowledge, this is the first work on the comparison between multiple data sets. The purpose of this study is to explore multiple consumer review websites for hotels from several levels. We analyzed various statistics including star ratings, evaluated hotel reviews independently for helpfulness via classifiers, and measured the semantic conformity to the star ratings over several comment attributes. We provide information related to: review word length, the number of sentences and average sentence length, the distance between websites on star ratings, and sentiment word score. The data corpus was extracted from TripAdvisor.com, Yelp.com and Orbitz.com using a webcrawl. It consists of a minimum of 40 reviews from January 1, 2012 to December 31, 2013 for 25 hotels in the New York area. Related work is presented in Section 2. In Section 3 we describe the methods and experiments used to extract and analyze data and present a technique for a text classification approach. Conclusion and future work is presented in Section 4.

2. RELATED WORK

In this section we offer a concise review of related works. Mining user-generated comments for information has received increasing attention in recent years. Many research studies focus on consumer products. Movie reviews, Youtube.com reviews, Amazon.com product reviews and Twitter are also popular due to the large datasets that are publicly available. Various studies related to consumer reviews focus on sentiment analysis or the use of machine learning techniques for product helpfulness. Our work is related both of them. Identification of semantics of reviews provides more information than rating scores. [?, ?, ?] use topic models to extract ratable aspects and polarities on each aspect. The topic models are all based on Latent Dirichllet Allocation [?], where the aspects of hotels like *Service*, *Room* are generated as topics. The semantic for each aspect is modeled as additional latent variable. Using statistical inferences, aspects and sentiment are unsupervised learned. This topic is also generally related to classification. [?] describes a review recommender based on a supervised classification approach to identify helpful comment reviews from TripAdvisor. This study evaluated Weka classifiers [?]. The classification features focused on four categories: Reputation features, Content features, Social features and Sentiment features. Within these categories, twenty-five variables were analyzed. Similarly, [?] postulated a method to rank Amazon.com reviews based on helpfulness. This method used support vector machine regression on semantic and sentiment features. In addition, structural and syntactical features were analyzed. The results showed the importance of review length and its star ratings. Further understanding of helpfulness and how consumers view perceived helpfulness depends on more than strictly the content of the review. [?] identified that a helpfulness score has a corollary relationship to other scores, which is consistent of individual bias in the company of a differing opinion distribution. In essence, user comments have an affect on other user comments.

3. METHODS AND EXPERIMENTS

In this section, we provide comparative study on multiple hotel review websites in which three level analysis gain the insight into the difference.

3.1 Data

We crawled rating scores, review content, and attribute rating scores(if available) of all hotels in New York City from the three websites TripAdvisor, Yelp and Orbitz. There are about 300 hotels and 256,000 pieces of reviews. However the reviews were created in different time periods for different hotels and websites. Tripadvisor has more volume of reviews compared with other two websites. To fairly evaluate the three websites. We selected 25 hotels which have at least 40 reviews during 2012 to 2014, which includes 30,537 reviews. TripAdvisors has 27,462 revies, Yelp has 2,026 reviews, and Orbitz has 1,079 reviews respectively. Although one can have access huge amount of review in TripAdvisor, it is impossible to let user read through so many reviews before booking. Therefore, review summarization or review selection is critical for providing users summarized information about hotels.

Figure 1: Basic statistics: the number of reviews, the averaged length of reviews, the averaged length of sentences, the averaged number of words per sentences

Table 1: Basic statistics

	TripAdvisor	Yelp	Orbitz
# of reviews	1098.4(604.1)	81.0(36.0)	43.1(37.1)
# of words	149.3(12.9)	166.8(22.7)	94.3(10.9)
# of sentences	8.4(0.5)	10.0(1.3)	6.4(0.5)

The number of words and sentences are also different among three websites. We tokenized reviews by whitespace. For each hotel, we counted the number of words for all the reviews associated with that hotel and recorded the averaged number. As shown in Figure 1, the reviews on Orbitz have smallest length compared with those of TripAdvisor and Yelp. The reviews in Yelp are longest. The numbers of sentences for three websites have similar situations. We also computed the sample deviation of the number of words and sentences. For example, the averaged and the standard deviation of the number of words in the review set of TripAdvisor is 1098.4 and 604.1 respectively, as shown in Figure 1.

We investigated the distribution of the number of words. We plot the number of reviews against the length of the review in log scale. As shown in Figure 2, most of reviews have small length, there are also reviews but have very relative small amount. All three curves have perfect linear relation with the length of review. We denote variable x as the length of review, y as the number of reviews having that length, and fit linear regression models to the log-scaled number of reviews and the review length. Since the independent variables has only one variable, there are only two parameters k and b, which are the slope and the intercept respectively. Table 2 shows the three linear regression models have different slops. It confirms that the review data set of Yelp have longer reviews than TripAdvisor; whereas Orbitz has shorter reviews.

$$\log(y) = k\,x + b \tag{1}$$

Table 2: Linear regression on the review length and the review amount

	TripAdvisor	Yelp	Orbitz
k	-0.0037	-0.0062	-0.013
b	8.27	6.80	6.37
residual	11.03	0.49	4.69

Figure 2: The star rating distribution of the length of reviews

Figure 3: The EMD distance between over rating distribution over 25 hotels

3.2 Star Rating Analysis

We compared the star rating distribution between these three websites. In every website, the webpage displays how many reviews gives how many stars in total. An averaged rating score is also displayed along the hotel. However, the quality of hotel varies along the time. It is unknown the review websites take the time as an important factor into account when analyzing the user comments. Therefore we do not know the overall rating reflecting immediate opinions from reviews. In addition, the overall rating score ignores the actual rating distribution. It is entirely possible that one hotel receives same ratings from two websites, but have different rating distribution from user reviews.

We use a measure of the distance called Earth Moving Distance(EMD). It is widely used in computer vision to measure the difference between two histograms. Here we use it to evaluate the difference of the distributions of star ratings between the websites for the same hotel. Basically, EMD is interpreted as minimal mass within one distribution to be moved to form another distribution. For each hotel on each three websites, we count the number reviews giving one of 1-5 stars that are associated with that hotel. The EMD distance is then computed between these three distributions. Figure 3 shows EMD distance for 25 hotels between TripAdvisor, Yelp and Orbitz. It shows in the rating level, reviews from TripAdvisor does not so different with Yelp, compared with other two pairs.

3.3 Helpfuness

Helpfulness is another important factor for a review. Helpful reviews give customers and managers informative details about the hotels. TripAdvisor and Yelp provides buttons of reviews to let users vote if the review is helpful, and a counter to show how many votes for the review. Orbitz.com and other websites does not have such access. We computed the averaged percentage of helpful reviews in total reviews. In our data set TripAdvisor has 5.67% helpful reviews, while Yelp has 13.32% helpful reviews.

There are some research effort [?, ?, ?] on helpfulness using various text feature, user history feature and social feature [?]. The most related feature is the reputation of the user. Using all kinds of features, JRip (Repeated Incremental Pruning to Produce Error Reduction) reaches 0.92 AUC score. However, these feature are generally not public available in most review websites other than TripAdvisor. Helpfulness is a high-level semantic concept. To achieve better understanding of helpfulness, one could turn to recent development of deep learning [?].

3.4 Conformity

We also investigated the conformity of a review on attributes of reviews. We defined the conformity as the consistency between the text sentiment and the attribute rating score. A user might get confused when he finds the review text full of positive comments, but with negative rating stars. It is also likely that the rating scores is influenced by other reviews giving high ratings. The attribute rating scores can be found in the bottom of reviews. For example, *Room*, *Service*, *Location*, *Value*.

Figure 4: The word sentiment score with the attribute rating score over four attributes

Table 3: Linear correlation coefficient and p-value

	ρ		p-value	
	TripAdvisor	Orbitz	TripAdvisor	Orbitz
Room	0.896	0.658	0.000	0.000
Location	0.581	0.313	0.002	0.126
Cleanliness	0.683	0.340	0.000	0.096
Service	0.911	0.670	0.000	0.000

We first classified the sentences of reviews into aspects, then used the sentiment classifier implemented in NLTK to get the sentiment score for sentences. We adopted a simple boosting method from Latent Aspect Rating Analysis(LARA)[?] for aspect segmentation. It began with a few seed words for each aspect. For each aspect, the sentence is labeled with the aspect which has the most confidence. The confidence between a word and an aspect is measured by Chi-square statistics. Second, the most frequent not-labeled words in the labeled sentences were added as new seeds for the corresponding asepct. Reviews were segmented into aspects which consisted of sentences. We run a naive Bayes sentiment classifier from NLTK[?] to get the score of positive attitude. For each hotel and each aspect, we computed the averaged sentiment score over all reviews. We compared the averaged sentiment score against with the averaged attribute rating score.

Figure 4 shows the scatter plot of four attributes, *Room*, *Service*, *Location*, *Cleanliness*. The TripAdvisor data points are generally distributed below Orbitz points, because the reviews of TripAdvisor give higher scores than Orbitz. For attribute *Room* and *Service* almost form a linear relation between word sentiment scores and rating scores. We also computed the linear correlation coefficients *rho* in Table 3. In most cases, the word sentiment score is linear proportional to the star rating score, with strong statistical confidence. The coefficient *rho* of TripAdvisor is larger then Orbitz in all four attributes. The users on TripAdvisor give more consistent reviews and ratings.

4. CONCLUSION

The goal of this work is to evaluate the difference between multiple source of user-generated reviews. We presents a systematic framework for comparatie evaluation. Here we selected TripAdvisor, Yelp and Orbitz as the review sources to build the data set which incorporates about 30,000 reviews. The basic statistics were computed, the number of words and sentences and the number of words per sentence. To accurately compare the overall ratings of reviews from different websites, we employed EMD distance as the distance measure. We also surveyed the methods to predict helpfulness of reviews. Finally, we segmented the reviews into several predefined aspects, and used a sentiment classifier to label words in reviews with sentiment scores. Using the linear correlation coefficients, we evaluate the conformity between the review text and the attribute scores. We consider sentiment analysis and review summarization across multiple review sources in future work.