Assessing Language Discrepancies between Travelers and Online Travel Recommendation Systems: Application of the Jaccard Distance Score to Web Data Mining

Sangwon Park, Ph D

School of Hospitality and Tourism Management, University of Surrey, 15AP02, Guildford, Surrey, GU2 7XH United Kingdom

Dae-Young Kim, Ph D

Hospitality Management, University of Missouri-Columbia, Columbia, MO 65211, USA

Assessing Language Discrepancies between Travelers and Online Travel Recommendation Systems: Application of the Jaccard Distance Score to Web Data Mining

Abstract

By using a human-centric approach to online recommender systems, this research aims to estimate the language discrepancies of which travelers and destination marketers describe the travel experiences across 11 tourism destinations in USA. In order to address the research purpose, data has been collected from two different sources that reflect the views of travelers and service providers. Then, a set of text data mining methods (i.e., clustering analysis and Jaccard distance score) was applied to identify the language differences between travelers and CVB websites, according to the following categories: shopping, dining, nightlife/activities, and attractions. Some possible methodological extensions that can improve recommendation capabilities, and managerial implications of these findings are provided.

Key words: Smart tourism, online recommender system, web data mining, Jaccard distance score

Introduction

The notion of smart tourism has recently gained attention from academics and practitioners. The concept aims to accelerate service innovation and improve tourism experience as well as enhance destination competitiveness by developing IT infrastructure and capabilities (Gretzel, Koo, Sigala and Xiang, 2015). Particularly, the intelligent smart tourism system uses information aggregation and ubiquitous connectedness to facilitate travelers to obtain personalized information (Gretzel, Sigala, Xiang, and Koo 2015). In this sense, the fundamental role of destination marketing organizations (DMOs) – that is, to understand travelers' expectations on visiting a destination and to offer tailored information and services – has become more crucial than ever before (Werthner and Klein, 1999).

Convention and visitors bureaus (CVBs) are important information brokers and disseminators in the local tourism industry and act as a layer of destination management in the U.S. With financial support from the local community, one of the critical goals of CVBs is to promote their destinations to both leisure and business travelers. As a result, providing useful/helpful information to travelers is an essential part in the CVBs' marketing activities and tasks (Stepchenkova et al., 2010; Kim, Jang, and Morrison, 2011). With the emergence of the Internet, many CVBs have adopted online applications that facilitate providing a substantial amount of information to travelers and, as a result, help plan their trips. Nonetheless, increased accessibility of destination-related information via the CVBs' websites may bring about "information overload", which creates challenges for online travelers to find appropriate information and make choices (Choi, Lehto, and Oleary, 2007; Kim, 2009). Besides, this information is often presented in a way that does not match how travelers search for information (Pan and Fesenmaier, 2006).

At present, the continuous evolution of information technology allows CVB websites to adopt recommender systems that can simplify the decision-making process for travelers (Fesenmaier, Wöber, and Werthner, 2006). This system enables travelers to lessen search costs and cognitive efforts by identifying alternatives that meet the specific needs of online users and by offering information in a personalized way (Gretzel, Hwang, and Fesenmaier, 2012; Kabassi, 2010; Wind and Rangaswamy, 2001). Accordingly, the recommender systems should be human-centric in their design and functionality. This requires system-user interactions by understanding cognitive styles of online information seekers and adjusting the recommender system to address the needs/desires (Bauernfeind, 2003; Zins, Frew, Hitz, and O'Connor, 2003). Particularly, this research focuses on linguistic interactions between users and the system in the context of tourism (see Dann, 1996; Gretzel et al. 2012). Based upon the definition of linguistic interactions referring to the way in which language appears interactions in everyday to represent cognitive process (Couper-Kuhlen and Selting, 2001), providing destination information that online travelers actually required with proper language is a cornerstone of recommendation, which facilitates smart tourism.

Matching the language in tourism is important to fulfill the effective communication between travelers (or visitors) and destination marketers (hosts) (Xiang, Wober and Fesenmaier, 2008). Previous studies identified a number of cases of incongruent destination images projected by marketers and perceived by travelers (MacKay and Fesenmaier, 2000) as well as across different online travel resources (e.g., blogs, magazines, guides, and travel trade) (Choi et al., 2007). The previous studies indicate the perceptions and/or images that destination travelers bring to mind do not seem to coincide with those highlighted by suppliers. Thus, an important research questions can be induced to test whether the experiences that travelers are looking for at

the specific destinations (“perceived” or “expected”) correspond to those promoted by marketers (“projected”) as a foundation for implementing a human-centric approach to the recommender system.

Therefore, the main goal of this paper is to estimate the effectiveness of CVB websites by comparing the nature of the language travelers used to describe their expected trip experiences and the contents provided by destination marketers on CVB websites. This research used two types of text mining techniques to examine the language discrepancies: text clustering (Stepchenkova, Kirilenko, and Morrison, 2009) and Jaccard distance score (Maedche, Pekar, and Staab, 2003). The findings of this study provide an extensive understanding of travel experiences expected by potential travelers and promoted by destination marketers across 11 tourism destinations, and assessment of language discrepancies focusing on multi-facets of travel products including “shopping”, “dining”, “night life and activity”, and “attraction”. Furthermore, this paper suggests applying the advanced data mining method to capture traveler preferences through analyzing textual data. As a result, the findings of this research suggests implications to develop more effective online recommender systems within the context of the tourism-related industry.

Literature Review

Online recommendation system

Ricci (2002) defined recommendation systems as applications associated with online platforms to suggest products/services and offer travelers personalized information to help with their decision-making process. Considering the complex nature of travel planning that involves numerous decision tasks – not only a destination but also, e.g., accommodations, activities,

restaurants – travelers are subject to experience an excess of information over their capability in the process to make diverse decisions. In this context, online recommendation systems have great potential for the usability in not only reducing the search costs but also improving decision qualities (e.g., Häubl and Trifts, 2000; Häubl and Dellaert, 2004). The online system can assist this task by matching the consumers' needs and preferences through providing tailored services and available options.

Studies on the development of online recommender systems can be categorized into two classes of research focuses: (1) process of the systems by which the recommendation systems operate with certain algorithms according to different contexts and (2) outcome of the systems based upon different individual and situational features (Fesenmaier et al., 2006). The recommendation systems vary in sophistication, ranging from simple retrieval or filtering approaches to comprehensive computing systems (Spiekermann and Paraschiv, 2002). There are basically two classifications: content-based and collaborative filtering systems (Yeh and Cheng, 2015). The assumption of content-based filtering is that characteristics of an item determine the user's preferences of the item (Ricci, 2002). Specifically, the content-based filtering approach provides a user with suggested products/services that are similar to those s/he has purchased or searched in the past. The systems attempt to match the attributes of the products/services with the characteristics of the users stored in the data base. To the contrary, collaborative filtering (or social filtering) systems infer the behavior of users toward products/services from other users who show similar interests or preferences and mimic social processes (Breese, Heckerman, and Kadie, 1998). This application assumes that the evaluation or opinions of others are an important information source that travelers use in their decision-making process (Gavalas, Konstantopoulos, Mastakas, and Pantziou, 2014).

The later aspect of the recommendation research investigating outcome of the systems is directly related to understandings of information processing and evaluations as well as decision making behaviors (Kabassi, 2010). For example, the consumer styles inventory has been applied to comprehend travel decision making styles in a way to envisage different information sources and contents travelers searched as well as attributes of the destinations they preferred (Zin et al., 2003). Gretzel et al., (2012) proposed a theoretical framework of destination recommender systems, suggesting the design components should be responsive to travelers' needs in terms of personal characteristics of the travelers (e.g., demographics and personality), situational needs and constraints (e.g., travel party and lengths of stay) and aspects of the decision-making process (e.g., the specificity of the choice task and decision frames). The focus on the traveler as the user of the system is highlighted by anticipating user needs and offering recommended alternatives according to specific consumption contexts (Buhalis and Amaranggana, 2015).

In brief, the constant findings of those previous studies argue that a human-centric approach is extremely important to make the recommender systems helpful and successful as decision-making support tools (Chung, Lee, Lee, and Koo, 2015; Gretzel, Hwang, and Fesenmaier, 2006). Design and functionality of human-centric computing require an intensive understanding of the individual behaviors so that the system enhances the ability to fulfill the interactions between the recommender systems and users. Zin et al. (2003) stated that the adjustment of the recommender system interface to fit a user's cognitive style is vital for enhancing the quality of the interaction. With the development of information technology, travelers are able to assert their needs for information, which are formed within their individual contexts. DMOs that manage contents/design of the destination websites become a primary agent that establishes a basic lens to represent a destination and experiential aspects as well as a

process by which travelers gain information (Pan and Fesenmaier, 2006). Thus, accomplishing the axiom "speaking the right language" is an important aspect in the online recommender system, which addresses the slogan of user-centered design: "Recommender systems are about people, not machines" (Ricci, 2002; pp. 57). The statement emphasizes issue of the product description language. That is, even if the recommendation system is well-developed in the engineering aspect, users will have challenges when information presented (or destination descriptions) is too terse or does not fit their needs. In this vein, to address the research question of this study, this research estimates language discrepancies between expressed by travelers and marketers. Due to the textual format of data analyzed in this research, a set of text mining approaches has been used to examine the differences. The following sections discuss the methods of text mining in general and Jaccard distance score approach in particular.

Text mining techniques

Text mining, also known as text data mining, is the process of deriving useful information from a text dataset (Feldman and Sanger, 2007). Machine learning, data mining, and information retrieval techniques have enabled the text mining field to advance dramatically during the past decade. The development of the data mining field brought about a diverse set of text mining techniques, such as text categorization, text clustering, concept extraction, sentiment analysis, and entity relation modeling (Ikonomakis, Kotsiantis, and Tampakas, 2005; Feldman and Sanger, 2007). These text mining techniques have been largely applied to various fields. For example, in marketing, text mining was used in the context of customer relationship management in order to develop prediction models for customer attrition (Coussement and Poel, 2008). The fuzzy cluster technique was used to classify customers based on their historical loyalty analysis

(Simha and Iyengar, 2006). In customer preference prediction analysis, Bayesian-based cluster models were adopted to predict the active user's preference (Adomavicius and Tuzhilin, 2005; Breese, et al., 1998). It was also used in gaining tourism knowledge as to how people search information online by clustering similar phases according to their meanings (Xiang, et al., 2007). Xiang, Schwartz, Gerdes, and Uysal (2015) tried to comprehend the associations of hotel guest experience with satisfaction by analyzing traveler reviews with text mining analytics.

Jaccard distance score

Based on the principle that online recommender system should speak in the same language as travelers in order to have optimal persuasive effects, the purpose of our study is to uncover and quantify the difference between the languages of the two parties. Since clustering analysis normally generates complex language clustering outcomes, it is difficult to evaluate the difference between the clustering outcomes by simply viewing the observations. In order to follow the principle, it is important to address the extent of the differences between the language used by travelers and the online recommender systems. Important questions may require normalized statistics to effectively quantify the language cluster outcomes. Simply observing complex language clustering outcomes could uncover the gap. However, it provides little information to the size of the gap due to the deficiency of simply observing the data in identifying differences from complex clustering outcomes.

Instead, this research suggests a method of Jaccard distance score, which is a sort of supervised measure, to analyze the training data and produce an inferred function (Tan, 2006). This method has been largely applied in computer science and bioinformatics fields. It has been used in flow cytometry data and to cluster quality evaluation and comparison (Liu, et al., 2008), internal and external measures of the colon dataset (Varshavsky, Horn, and Linial, 2007), user-

to-user similarity in collaborative filtering systems (Millan, Trujillo, and Ortiz, 2007), and when ranking correlation with human judgments on the MarkovLink model (Hughes and Ramage, 2007). While these previous attempts prove the usefulness of the Jaccard score method, limited applications have been adopted within tourism research, particularly in language analysis and online recommender systems.

The Jaccard distance score measures the dissimilarity between the clustering results. It measures the unique pairs of data objects in both the evaluated solution S and true solution T, adjusted by the unique and common pairs of data objects in S and T. Therefore, the values of the Jaccard distance score range from 0 (no difference) to 1 (perfect difference). The formula to calculate Jaccard distance shows as following:

$$Jaccard = J(T,S) = \frac{n_{10} + n_{01}}{n_{11} + n_{10} + n_{01}}$$

Where J(T,S) is Jaccard distance score of a solution S against the true solution T. In our study, the true solution T is based on the traveler's language, and the solution S is based on the online recommender system's language. n_{11} is the number of pairs of data objects that are clustered together in the same cluster in both S and T, n_{01} is the number of pairs of data objects that are clustered together only in S, and n_{10} is the number of pairs of data objects that are clustered together only in T.

Another important attribute of the Jaccard distance score is its normalization (Strube and Ponzetto, 2006). Since the formula excludes n_{00}, which is the number of pairs of data objects that are not clustered together in S or T, the Jaccard distance score reflects a real value that does not depend on the size of the cluster. This feature is important in our problem setting because all the selected online recommender systems in this study have different sizes of language clusters,

and we need a normalized statistic to evaluate the language difference in order to make the result comparable. In this study, we do not use the unsupervised measures that do not require prior training in order to mine the data: instead, the method we use measures how far a word is inclined towards separate groups (Tan et al., 2005). This is because this current research only uses one language clustering algorithm, and the absolute error itself may not be meaningful without comparing clustering results obtained from different clustering algorithms. In addition, since there is a limited potential of biases towards the data dimension reduction (e.g., a curve fitting approach), an unsupervised measure is unnecessary (see Liu, et al., 2008).

Methodology

The data analyses in this study followed four steps. First, a pilot study was performed to select distinctive destinations in the U.S. A total of 10 tourism and travel experts including scholars, industry practitioners, and graduate students were employed to check a face validity of representative destinations in the U.S. Second, two types of language data were collected: (1) travelers' written expression of destinations and (2) descriptive text presented on CVB websites. Third, qualitative data pre-processing was employed to refine the data for the data mining approach. Fourth, the clustering technique and Jaccard score method were performed to measure the different nature of languages between clusters. A flow diagram of our approach is displayed in Figure 1.

[Figure 1 is about here]

For text data analysis, a computer-aided text mining software, CATPAC (Woelfel and Woelfel, 1997), was used to cluster phases (Stepchenkova, et al., 2009). CATPAC detects textual patterns and identifies phases' associations by measuring the distance between phase concepts using artificial neural networks. It can cluster phases based on concepts and generate clustering files, and these files can be viewed by the software ThoughtView (Woelfe and Woelfel, 1997). CATPAC provides efficient clustering and visualization of the phase clustering analysis (Lowe, 2002) in a form that is logically precise, human friendly, and computationally tractable (Montes-y-Gómez, Gelbukh, López, and Baeza-Yates, 2001). It excludes theoretical and personal biases (Geana, 2006) and has been proven extremely useful in identifying clusters based on the concept of phases (Dinauer and Fink 2005).

Selection of CVBs websites

One hundred college students in the Midwest portion of the U.S. were recruited to rate 34 major U.S. cities where they had previous trip experience or perceptions regarding second-hand information in terms of the two dimensions that are used for many studies about tourism experience (i.e., attractiveness and excitement) (Oh, Fiore, and Jeong, 2007; Pine and Gilmore, 1999). The students rate integral scores on each city ranging from 0 to 10 where 0 is least favorable and 10 is most favorable. To select distinctive cities, we classified the cities into four tiers according to a visual map, which is based on mean scores of attractiveness and excitement. The logic behind the selection is that examining differently attributed cities increases the generalizability of the study findings. We selected 11 distinctive cities out of 34, as shown in Figure 2. The 11 cities include Las Vegas, Honolulu, NYC, Chicago, Orlando, Los Angeles,

Tampa, Denver, Baltimore, Pittsburgh, and Detroit. The cities' CVB websites are listed in Table 1.

[Figure 2 is about here]

[Table 1 is about here]

Data collection

Two types of language data describing the cities were collected from travelers and marketers, respectively. In regard to traveler language, 85 samples were collected from college students. The students who have domestic trip and web information search experiences in a tourism marketing class were recruited voluntary based. Based upon the premise that travel is an experiential product, travelers are more likely to keep information about experiences in memory within a story format, referring to mental imagery of their experiences (Tussyadiah, Park, and Fesenuamer, 2011; Govers, Go and Kumaer, 2007). Hence, subjects of this study were asked to elaborate their expected experiences or perceptions about the selected 11 cities regarding major travel components, such as dining, shopping, night life, and activities (Park, Nicolau and Fesenmaier, 2013). The story format was suggested with more than 7 sentences; as a result, the total number of descriptive scripts was 935.

The questions were adopted by measurements Govers et al., (2007) used. Specifically, the instruction for writing was: "*First, imagine that next week you will visit the each following cities. Project yourself in the scenario, and tell me your story. What do you think your experience in the destination would be like (ex. activities, shopping, dining, night life, environment, and so on)? What images and thoughts immediately come to mind? What would you expect to see, or feel, hear, smell, taste there? Without any research or additional information, kindly be spontaneous and share with me whatever thoughts come to your mind right now, whether*

positive or negative. Make your response as detailed, and try to write in story format, using complete sentences, not just loose words. If you know little about a destination, your story will probably be short. If you already have clear ideas about it, your story might be very long. But remember, there is no right, wrong, or best model answer; simply express your own ideas about each destination, and NOT what you think we want to hear. Share your ideas about each destination with me right now."

Second, the textual contents from the CVB websites owned by the selected 11 cities were extracted. These data extracted from the websites were used as the language representation of the CVB website owners. Once we obtained the collection of textual data from both two sources, three steps of pre-processing of the textual data were employed: (1) remove all the "outlier" words (e.g., "a", "an", "the", "you", "and", "but") so that only meaningful phases could be analyzed; (2) replace plurals with singles, e.g., replacing "stores" with "store"; (3) substitute all verb tenses with present tenses in order to transfer all the verbs with the same linguistic roots, e.g., replace "ate" with "eat".

Then, the Jaccard Distance Score technique was conducted to measure the language differences between the travelers and the CVB websites. The Jaccard distance score is defined as follows. Let T be the "true" clustering and S the clustering we wish to evaluate. In this study, the "true" clustering is the clustering results of the language used by travelers, and the clustering result we want to evaluate is the one of the language presented on the CVB websites. Let n_{11} represent the number of pairs of elements in both S and T, n_{01} represent the number of pairs in the same clusters only in S, and n_{10} represent the number of pairs in the same cluster only in T. The Jaccard distance score is then defined as the following:

$$Jaccard = J(T,S) = \frac{n_{10} + n_{01}}{n_{11} + n_{10} + n_{01}}$$

Based on this definition, two identical clusters will generate a Jaccard distance score of 0, while two completely different clusters will generate the Jaccard distance score of 1. The larger score means that the difference between the two clusters is greater.

Results

Word frequency analysis

Word frequency analysis was first employed to identify the 10 most popular words on the CVB websites and the travelers' descriptions about destinations, respectively (see Table 2). By comparing the two parties, the results show the commonalities and differences in the frequent words. For example, travelers tend to show great interests in dining and shopping-related words as well as the nature of travel (e.g., 'fun'), whereas CVB websites highlighted general words related to attractions (e.g., 'museum' and 'culture/history'). Besides, CVB websites are more likely to use verbs (e.g., experience, get in the know about, come back to, etc.), while travelers use nouns more often (e.g., specific attractions, restaurant names, street name in each destination).

[Table 2 is about here]

Overall differences

In the next phase, the overall differences between the two languages were examined by the Jaccard distance score. The scores for the 11 selected cities are presented in Figure 3. Orlando, Tampa, Los Angeles, and Chicago have relatively smaller differences (around 0.87)

compared to other destinations, including Denver, Baltimore, and New York City that indicate the score of around 0.93. The remaining four cities are Detroit, Honolulu, Las Vegas, and Pittsburgh, which all have Jaccard distance scores near 0.98. It appears that the first (i.e., high attractiveness and excitement) and the fourth tier cities (i.e., low attractiveness and excitement), except Baltimore in the perception-positioning map (see Figure 3), show larger language differences. On the other hand, the second tier cities in the perception-positioning map, except New York City, show relatively low distance scores, which indicates smaller language differences. The specific estimations of language differences across four travel facets (i.e., shopping, dining, night life/activities, and attractions) are discussed in the following section.

[Figure 3 is about here]

Shopping

As illustrated in Figure 4, the shape of the perception positioning map figure presents the similar feature with one of the overall differences. However, the value of the Jaccard distance score on shopping information is slightly smaller than the score of the overall differences. This suggests that the ways to describe the shopping experiences at the tourism destinations between travelers and marketers are relatively coherent, particularly for those second tier destination involving Orlando, Tampa, Los Angeles, and Chicago. By contrast, Honolulu appears to be a destination where travelers' expectations and CVBs' promoted experiences are largely mismatched. Since customers' perceptions about shopping for all the 11 selected cities are consistent with their perceptions about the 11 selected cities as a whole, this indicates that the shopping difference greatly contributes in shaping the overall difference.

[Figure 4 is about here]

Dining

Figure 5 shows that the shape of the perception positioning map figure of dining also presents similar features with that of the overall difference. Destinations in the first tier including Las Vegas (0.97) and Honolulu (0.99) have the highest scores, while Orlando (0.84), Chicago (0.87), and Los Angeles (0.85) (i.e., second tier destinations) show relatively lower values. As a result, the different nature of language on dining presents very similar Jaccard distance scores with that of shopping difference, but they are both slightly smaller than the score of overall difference.

[Figure 5 is about here]

Night life and activities

As shown in Figure 6, the shape of night life and activities represents a different form compared to the shapes of overall differences, shopping, and dining. Specifically, Jaccard distance scores for most cities are above 0.92, which indicates that there is a larger gap between the language used by the CVBs websites and travelers when describing night life and activities. In particular, Jaccard distance scores of the three destination (e.g., Orlando, Tampa, and Los Angeles) were below 0.90 in shopping and dining experiences; however, the result for night life and activities show scores around 0.95. The reverse pattern was also identified in Detroit (fourth tier) and Denver (third tier) destinations. While the scores of the two destinations were 0.95 and 0.92 in shopping and dining attributes, respectively, in case of night life and activity, the estimated values appear 0.87. It is interestingly identified that travelers used trendier and fancier

phases in night life and activities than in shopping or in dining, whereas the CVB websites normally maintain a consistent style across all the functions.

[Figure 6 is about here]

Attractions

It is obvious that all of the 11 cities have higher distance scores in attractions than other facets of travel, which is around 0.95 for all destinations. This indicates that the different nature of language between the two parties is an influential factor that causes the overall difference. In other words, attractions are where travelers and marketers show the biggest difference in language descriptions when depicting the destination experiences. There is no destination depicting the distance score below 0.90.

[Figure 7 is about here]

The detailed scores of Jaccard distance analysis are presented at Table 3 across individual destinations and aspects of travel experiences.

[Table 3 is about here]

Discussions and Implications

One of the main functionalities of a personalization system is the presentation of results, which involves technologies for improving the interactivity of systems and achieving human-computer interaction (Diaz, Garcia, and Gervas, 2008). Based upon this premise, this study

attempts to evaluate the effectiveness of online recommender systems by comparing qualitative differences in travelers' perceptions and destination website contents. More specifically, this research applied a word frequency analysis to compare descriptions about tourism destinations in order to show the commonalities and differences in the language used by travelers and online recommender systems. Data mining tools including a neural network clustering analysis of qualitative data and the Jaccard distance score were then used to capture and measure the language differences in terms of four travel product categories (i.e., "shopping", "dining", "night life and activities", and "attractions") as well as overall travel experiences combining all the facets.

This study suggests that understanding the language used by travelers to describe destination information is crucial to develop effective online recommender systems, which is the foundation of smart tourism (Neuhofer, Buhalis, and Ladkin, 2015). The result of this study reveals that the nature of the language used by travelers and the language used by CVB websites in describing the destination information and services are different across four facets of travel experiences. Comparatively, the distinctive discrepancies in night life/activities and attractions have been identified. With regard to classified destinations consisting of attractiveness and excitement, first tier destinations (e.g., Las Vegas and Honolulu) show relatively higher distance scores. This finding suggests an idea that tailored descriptive information should be developed according to different categories of travel experience in order to meet the information search needs for online travelers. This argument is the reason for the creation of online recommender system to suit online users' preferences at the right time, referring to foundation of context marketing (Buhalis and Amaranggana, 2015).

There are several implications for online recommender systems and smart tourism

researchers. First, the majority of the existing online recommender systems are built by marketers who provide the information of the products or services. Marketers are always trying to develop products and services to meet customers' needs in order to avoid marketing myopia, which is the false implementation of product or service-oriented business. In practice, even though marketers have developed customer-need-oriented products or services, the failure to "speak the same language" (Dann, 1996) on online recommender systems is still apparent. It conveys distorted information to travelers that may lead the deviated perceptions. This drawback potentially brings about resource waste and time contributed in the process of developing customer-need-oriented products or services and thus may create more burden of information processing. This problem would result in a more overwhelming online environment, in which customers find it more difficult to find useful information and to choose a choice among rich services (Park and Nicolau, 2015).

Second, given the nature of tourism composing high-involvement and experiential products, the needs of information processing in trip planning appear to be multi-dimensional including not only functional but also novelty or sensation-seeking values (Cho and Jang, 2008; Vogt and Fesenmaier, 1998). Therefore, destination information provided on online recommender systems should be able to meet this heterogeneous need of travelers during their searching process. This calls for highly personalized designs that can effectively convey the experiential destination information so as to address travelers' special needs.

Third, the methodology used in this study utilized artificial neural network context clustering analysis and Jaccard distance score measures as practical data mining tools for capturing and quantifying context differences. Travelers use a different language in describing products or services other than the language used by online recommender systems that are

developed by marketers, but up to now no study employed the Jaccard distance score measures to quantify this context difference in the tourism field. Thus, limited understanding can be obtained from content clustering analysis, which could only display the complex clustering outcomes without knowing the extent of their differences. This limitation of previous studies largely restricts their utility for online recommendation system marketers. To overcome the limitation of current research, we recommend the use of Jaccard distance score measures for future research. This methodology should not be limited to online recommender system scopes, but it should be extended to various travelers' perceptions and image judgments, such as advertising research, persuasive message production, communications, campaigns, brandings, and other marketing related research.

Another contribution of our methodology is that we divided the destination experiences into several sub-categories: shopping, dining, night life and activities, and attractions. Through the Jaccard distance score measures, we were able to quantify the overall experiential and the perceptual differences for each of these four sub-groups. This method allowed the researchers to find which sub-facets of the experiences have obvious perception differences, which sub-facets' perception differences have the most influence in shaping people's perception toward the overall destination experiences. The results of this study show that the two sub-facets of destination, (i.e., "shopping" and "dining") have similar patterns in terms of perception difference, and they are the two most influential facets in shaping the overall perception difference depicted. In addition, "attractions" had larger distance scores, which indicates the most obvious perception differences between two parties. This reveals that traveler online searching is mainly devoted to searching for shopping and dining information. This finding is consistent with the study of Park and Fesenmaier (2014), who defined travel decision flexibility. Park and Fesenmaier identified

that travelers tend to develop different information search strategies for deciding shopping and restaurants compared to accommodations, additional destinations, and places/attractions to visit. In other words, people are likely to keep obtaining information for shopping facilities and restaurants from pre-trip to en-route stages and are subject to change their decisions according to the information obtained.

Meanwhile, travelers tend to seek more sophisticated contents when they search attraction-related online information; this resulted in larger Jaccard distance scores. That is, the current online travel recommender systems more seriously does not seem to properly deliver tailored information about attractions that the traveler wants. Therefore, a desirable online recommender system should take full consideration of travelers' different perceptions toward each facet of travel experiences when they promote their destination information. The result of this study also implies that marketers should focus more on "night life and activities" and "attractions" when describing destination information in order to effectively narrow down the perception difference between travelers and service providers.

Limitations and Future Research

Although this study provides penetrating insights into the differences between travelers' perceptions and online recommender systems, there are several aspects of our study that need to be explored further. First, this research collected customer language data on the selected cities from college students. Future research should overcome this sampling limitation by obtaining a more general representation of customers' language to improve the external validity. Our initial concern is that younger people represent the majority of internet users who are engaged actively in online behaviors and activities. However, this potential of bias in sampling may restrict

additional insights into travelers' perception differences without more representative sample populations. In this similar vein, controlling levels of prior knowledge about a destination is important so as to minimize any potential confounding effect.

Second, this paper suggested that CVB websites to focus on the identification of language discrepancies between descriptions perceived by travelers and promoted by destination marketers; however, optimal steps of this improvement were not explored in this research. Therefore, future work should investigate ways with separations between travelers' experiences and perceptions. In addition, more psychological studies needed to better understand the correlations between sub-facets of tourism and their influences on the overall tourism experiences.

Third, this research found that the language style of males and females are not the same, implying an important issue to understand heterogeneous ways of information representation across different age groups as well as genders. Therefore, it is suggested to conduct future research to measure the difference between the language used by the CVB websites and different genders in different age groups. Consequently, if the future research sheds light on the language biases of the CVBs websites against certain genders, it can give more specific recommendations to CVB website designers.

References

Adomavicius, Gediminas, and Alexander Tuzhilin. "Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions." *Knowledge and Data Engineering, IEEE Transactions on* 17, no. 6 (2005): 734-749.

Bauernfeind, Ulrike. "The evaluation of a recommendation system for tourist destination decision making." In *Proceedings of the XII International Symposium on Tourism and Leisure*. 2003.

Breese, John S., David Heckerman, and Carl Kadie. "Empirical analysis of predictive algorithms for collaborative filtering." In *Proceedings of the Fourteenth conference on Uncertainty in artificial intelligence*, pp. 43-52. Morgan Kaufmann Publishers Inc., 1998.

Buhalis, Dimitrios, and Aditya Amaranggana. "Smart Tourism Destinations Enhancing Tourism Experience Through Personalisation of Services." In*Information and Communication Technologies in Tourism 2015*, pp. 377-389. Springer International Publishing, 2015.

Cho, Mi-Hea, and SooCheong Shawn Jang. "Information value structure for vacation travel." *Journal of Travel Research* 47, no. 1 (2008): 72-83.

Choi, Soojin, Xinran Y. Lehto, and Joseph T. Oleary. "What does the consumer want from a DMO website? A study of US and Canadian tourists' perspectives." *International Journal of Tourism Research* 9, no. 2 (2007): 59-72.

Chung, Namho, Hyunae Lee, Seung Jae Lee, and Chulmo Koo. "The influence of tourism website on tourists' behavior to determine destination selection: A case study of creative economy in Korea." *Technological Forecasting and Social Change* 96, (2015): 130-143.

Cohen, Erik, and Robert L. Cooper. "Language and tourism." *Annals of Tourism Research* 13, no. 4 (1986): 533-563.

Couper-Kuhlen, Elizabeth,. and Margret Selting. "Introducing interactional linguistics.". In *Studies in interactional linguistics*, Edited by: Selting, M and Couper-Kuhlen, E. 2001, pp. 1–22. Amsterdam: John Benjamins.

Coussement, Kristof, and Dirk Van den Poel. "Integrating the voice of customers through call center emails into a decision support system for churn prediction." *Information and Management* 45, no. 3 (2008): 164-174.

Dann, Graham MS. "Noticing notices: tourism to order." *Annals of Tourism Research* 30, no. 2 (2003): 465-484.

Dann, Graham. "The language of tourism." A Sociolinguistic Perspective. Oxon: CAB International (1996).

Diaz, Alberto, Antonio Garcia, and Pablo Gervas. “User-centered versus system-centered evaluation of a personalization system.” *Information Processing and Management*, 44, (2008): 1293-1307.

Dinauer, Leslie D., and Edward L. Fink. "Interattitude structure and attitude dynamics." *Human Communication Research* 31, no. 1 (2005): 1-32.

Feldman, Ronen, and James Sanger. *The text mining handbook: advanced approaches in analyzing unstructured data*. Cambridge university press, 2007.

Fesenmaier, Daniel R., Karl W. Wöber, and Hannes Werthner, eds.*Destination recommendation systems: Behavioral foundations and applications*. Cabi, 2006.

Gavalas, Damianos, Charalampos Konstantopoulos, Konstantinos Mastakas, and Grammati Pantziou. "Mobile recommender systems in tourism." *Journal of Network and Computer Applications* 39 (2014): 319-333.

Geana, Mugur Valentin. "Penetration of innovation: taming the unexplored interactions between information, knowledge and persuasion in the innovation-decision model." PhD diss., University of Missouri--Columbia, 2006.

Govers, Robert, Frank M. Go, and Kuldeep Kumar. "Promoting tourism destination image." *Journal of Travel Research* 46, no. 1 (2007): 15-23.

Gretzel, Ulrike, Yeong-Hyeon Hwang, and Daniel R. Fesenmaier. "A behavioural framework for destination recommendation systems design." *Destination Recommendation Systems: Behavioural Foundations and Applications* 64 (2006): pp53-66.

Gretzel, Ulrike, Yeong-Hyeon Hwang, and Daniel R. Fesenmaier. "Informing destination recommender systems design and evaluation through quantitative research." *International Journal of Culture, Tourism and Hospitality Research* 6, no. 4 (2012): 297-315.

Gretzel, Ulrike, Chulmo Koo, Marianna Sigala, and Zheng Xiang. "Special Issue on smart tourism: convergence of information technologies, experiences, and theories." *Electronic Markets* 25, no. 3 (2015): 175-177.

Gretzel, Ulrike, Marianna Sigala, Zheng Xiang, and Chulmo Koo. "Smart tourism: foundations and developments." *Electronic Markets* 25, no. 3 (2015): 179-188.

Häubl, Gerald, and Benedict GC Dellaert. "Electronic travel recommendation agents and tourist choice." *The tourism and leisure industry: shaping the future* (2004): 317-24.

Häubl, Gerald, and Valerie Trifts. "Consumer decision making in online shopping environments: The effects of interactive decision aids." *Marketing science* 19, no. 1 (2000): 4-21.

Hughes, Thad, and Daniel Ramage. "Lexical Semantic Relatedness with Random Graph Walks." In *EMNLP-CoNLL*, pp. 581-589. 2007.

Ikonomakis, M., S. Kotsiantis, and V. Tampakas. "Text classification using machine learning techniques." *WSEAS Transactions on Computers* 4, no. 8 (2005): 966-974.

Kabassi, Katerina. "Personalizing recommendations for tourists." *Telematics and Informatics* 27, no. 1 (2010): 51-66.

Kim, Dae-Young. "The moderating effect of individual and organizational factors on information technology acceptance: The case of US CVBS'internet marketing." *Journal of Travel and Tourism Marketing* 26, no. 3 (2009): 329-343.

Kim, Dae-Young, Soocheong Jang, and Alastair M. Morrison. "Factors affecting organizational information technology acceptance: A comparison of convention and visitor bureaus and meeting planners in the United States." *Journal of Convention and Event Tourism*, vol. 12, no. 1, pp. 1-24. Taylor and Francis Group, 2011.

Liu, Lin, Li Xiong, James J. Lu, Kim M. Gernert, and Vicki Hertzberg. "Comparing and clustering flow cytometry data." In *Bioinformatics and Biomedicine, 2008. BIBM'08. IEEE International Conference on*, pp. 305-309. IEEE, 2008.

Lowe, Will. "Software for content analysis–A Review." *Cambridge: Weatherhead Center for International Affairs and the Harvard Identity Project* (2002). http://people.iq.harvard.edu/ ~wlowe/ Publications/rev.pdf.

MacKay, Kelly J., and Daniel R. Fesenmaier. "An exploration of cross-cultural destination image assessment." *Journal of Travel Research* 38, no. 4 (2000): 417-423.

Maedche, Alexander, Viktor Pekar, and Steffen Staab. "Ontology learning part one—on discovering taxonomic relations from the web." In *Web Intelligence*, pp. 301-319. Springer Berlin Heidelberg, 2003.

Millan, Marta, Maria Trujillo, and Edward Ortiz. "A collaborative recommender system based on asymmetric user similarity." In *Intelligent Data Engineering and Automated Learning-IDEAL 2007*, pp. 663-672. Springer Berlin Heidelberg, 2007.

Montes-Y-Gomez, M., A. Gelbukh, Alvaro Lopez-Lopez, and R. Baeza-Yates. "Text mining with conceptual graphs." In *IEEE INTERNATIONAL CONFERENCE ON SYSTEMS MAN AND CYBERNETICS*, vol. 2, pp. 898-903. 2001.

Neuhofer, Barbara, Dimitrios Buhalis, and Adele Ladkin. "Smart technologies for personalized experiences: a case study in the hospitality domain." *Electronic Markets* 25, no. 3 (2015): 243-254.

Oh, Haemoon, Ann Marie Fiore, and Miyoung Jeong. "Measuring experience economy concepts: Tourism applications." *Journal of travel research* 46, no. 2 (2007): 119-132.

Pan, Bing, and Daniel R. Fesenmaier. "Online information search: vacation planning process." *Annals of Tourism Research* 33, no. 3 (2006): 809-832.

Park, Sangwon, and Daniel R. Fesenmaier. "Travel decision flexibility." *Tourism Analysis* 19, no. 1 (2014): 35-49.

Park, Sangwon, and Juan L. Nicolau. "Asymmetric effects of online consumer reviews." *Annals of Tourism Research* 50 (2015): 67-83.

Park, Sangwon, Juan L. Nicolau, and Daniel R. Fesenmaier. "Assessing advertising in a hierarchical decision model." *Annals of Tourism Research* 40 (2013): 260-282.

Pine, B. Joseph, and James H. Gilmore. *The experience economy: work is theatre and every business a stage*. Harvard Business Press, 1999.

Ricci, Francesco. "Travel recommender systems." *IEEE Intelligent Systems*17, no. 6 (2002): 55-57.

Simha, Jay B., and S. S. Iyengar. "Fuzzy data mining for customer loyalty analysis." In *Information Technology, 2006. ICIT'06. 9th International Conference on*, pp. 245-246. IEEE, 2006.

Spiekermann, Sarah, and Corina Paraschiv. "Motivating human–agent interaction: Transferring insights from behavioral marketing to interface design." *Electronic Commerce Research* 2, no. 3 (2002): 255-285.

Stepchenkova, Svetlana, Andrei P. Kirilenko, and Alastair M. Morrison. "Facilitating content analysis in tourism research." *Journal of Travel Research* 47, no. 4 (2009): 454-469.

Stepchenkova, Svetlana, Liang Tang, SooCheong Shawn Jang, Andrei P. Kirilenko, and Alastair M. Morrison. "Benchmarking CVB website performance: Spatial and structural patterns." *Tourism Management* 31, no. 5 (2010): 611-620.

Strube, Michael, and Simone Paolo Ponzetto. "WikiRelate! Computing semantic relatedness using Wikipedia." *In AAAI,* vol. 6, pp. 1419-1424. 2006.

Tan, Pang-Ning. *Introduction to data mining*. Pearson Education India, 2006.

Tussyadiah, Iis P., Sangwon Park, and Daniel R. Fesenmaier. "Assessing the effectiveness of consumer narratives for destination marketing." *Journal of Hospitality and Tourism Research* 34, no. 1 (2011): 64-78.

Varshavsky, Roy, David Horn, and Michal Linial. "Clustering algorithms optimizer: a framework for large datasets." In *Bioinformatics Research and Applications*, pp. 85-96. Springer Berlin Heidelberg, 2007.

Vogt, Christine A., and Daniel R. Fesenmaier. "Expanding the functional information search model." *Annals of Tourism Research* 25, no. 3 (1998): 551-578.

Werthner, Hannes, and Stefan Klein. *Information technology and tourism: a challenging relationship*. Springer-Verlag Wien, 1999.

Wind, Jerry, and Arvind Rangaswamy. "Customerization: The next revolution in mass customization." *Journal of interactive marketing* 15, no. 1 (2001): 13-32.

Woelfel, J. K., and J. D. Woelfel. "CATPAC for Windows (Version 2.0)[Computer program]." *Amherst, New York: The Galileo Company* (1997).

Xiang, Zheng, Sang-Eun Kim, Clark Hu, and Daniel R. Fesenmaier. "Language representation of restaurants: Implications for developing online recommender systems." *International Journal of Hospitality Management* 26, no. 4 (2007): 1005-1018.

Xiang, Zheng, Zvi Schwartz, John H. Gerdes, and Muzaffer Uysal. "What can big data and text analytics tell us about hotel guest experience and satisfaction?." *International Journal of Hospitality Management* 44 (2015): 120-130.

Xiang, Zheng, Karl Woeber, and Daniel R. Fesenmaier. "Representation of the online tourism domain in search engines." *Journal of Travel Research* 47, no. 2 (2008): 137-150.

Yeh, Duen-Yian, and Ching-Hsue Cheng. "Recommendation system for popular tourist attractions in Taiwan using Delphi panel and repertory grid techniques." *Tourism Management* 46 (2015): 164-176.

Zins, Andreas H., A. J. Frew, M. Hitz, and P. O'Connor. "Adapting to cognitive styles to improve the usability of travel recommendation systems." In *Information and communication technologies in tourism 2003: Proceedings of the International Conference in Helsinki, Finland, 2003.*, pp. 289-297. Springer-Verlag Wien, 2003.

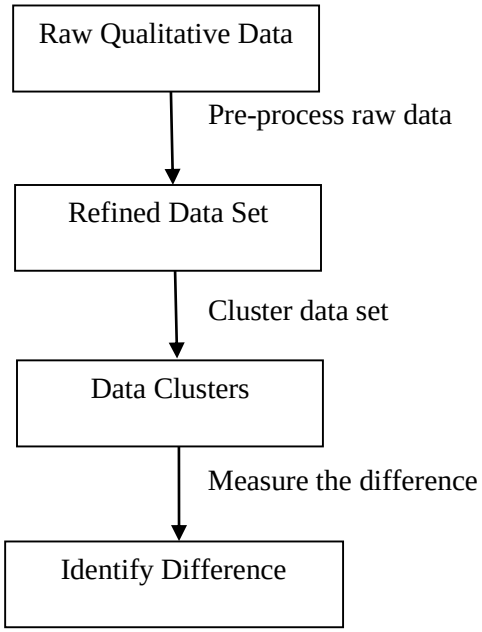

Fig. 1. Data Mining Process

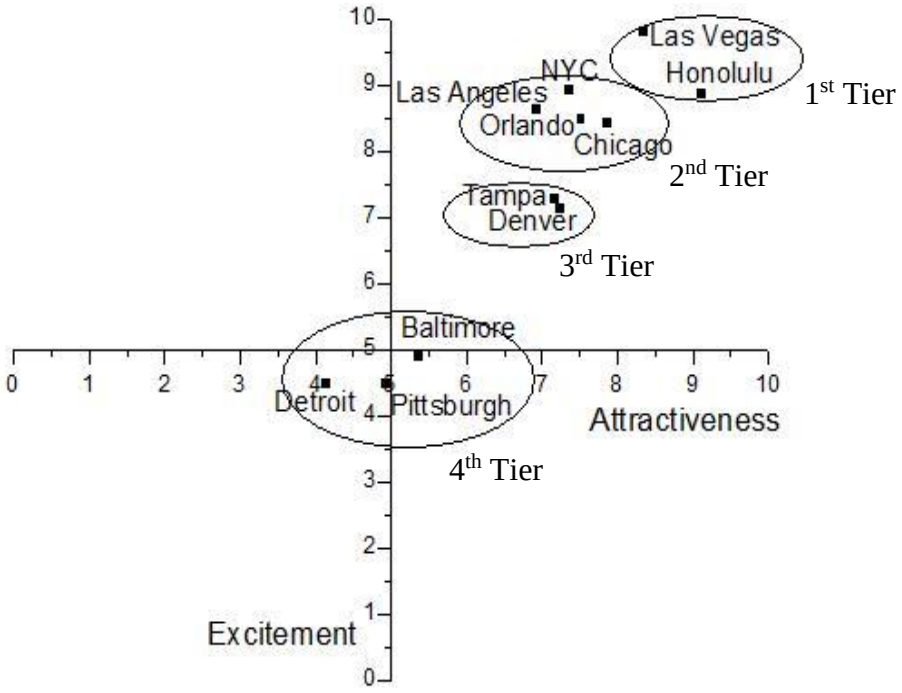

Fig. 2. Perception Positioning Map of Cities on Attractiveness and Excitement

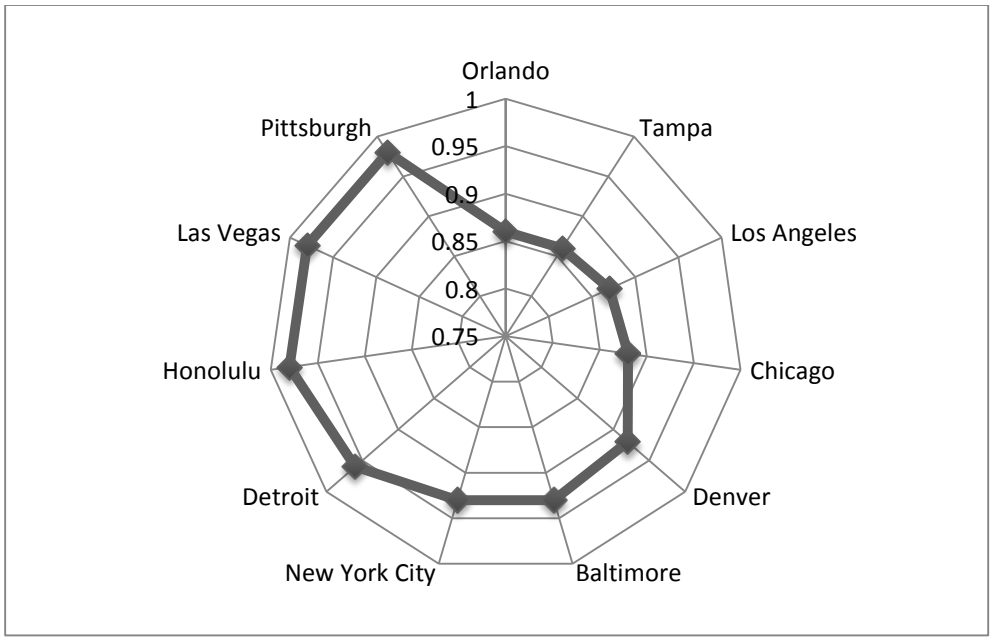

Fig. 3. Overall Jaccard Distance Score for the Selected Cities

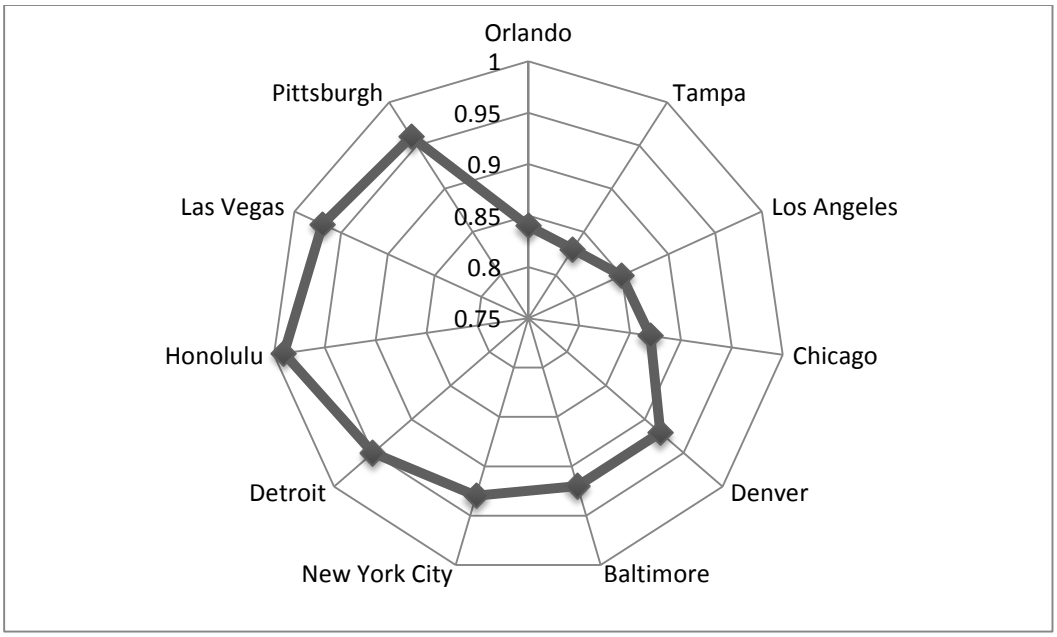

Fig. 4. Jaccard Distance Score for the Selected Cities of Shopping

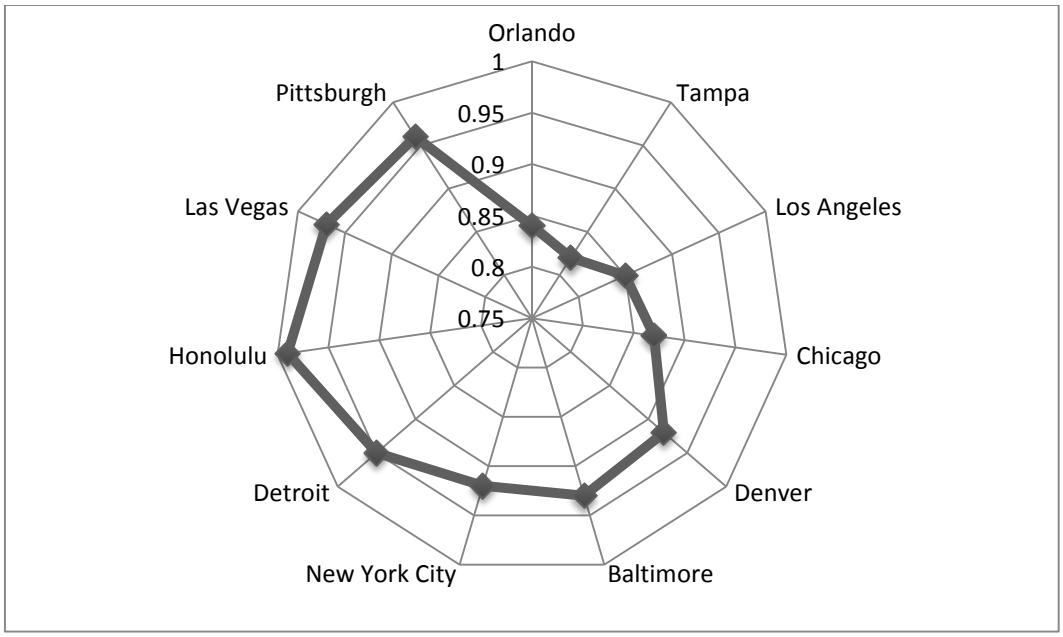

Fig. 5. Jaccard Distance Score for the Selected Cities of Dining

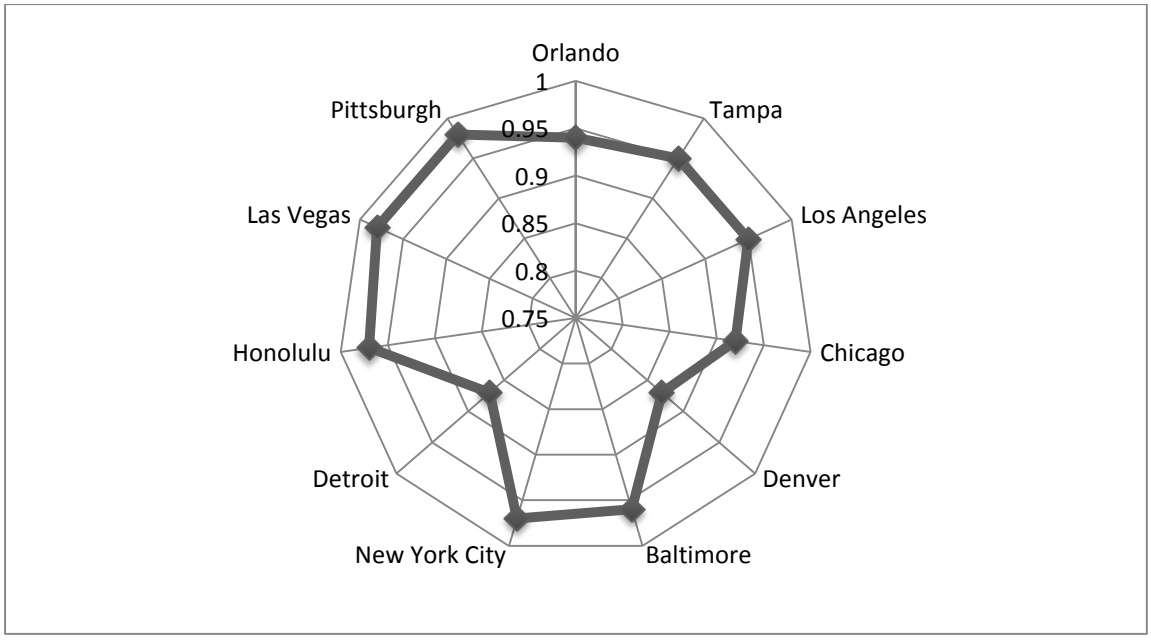

Fig. 6. Jaccard Distance Score for the Selected Cities of Night Life and Activities

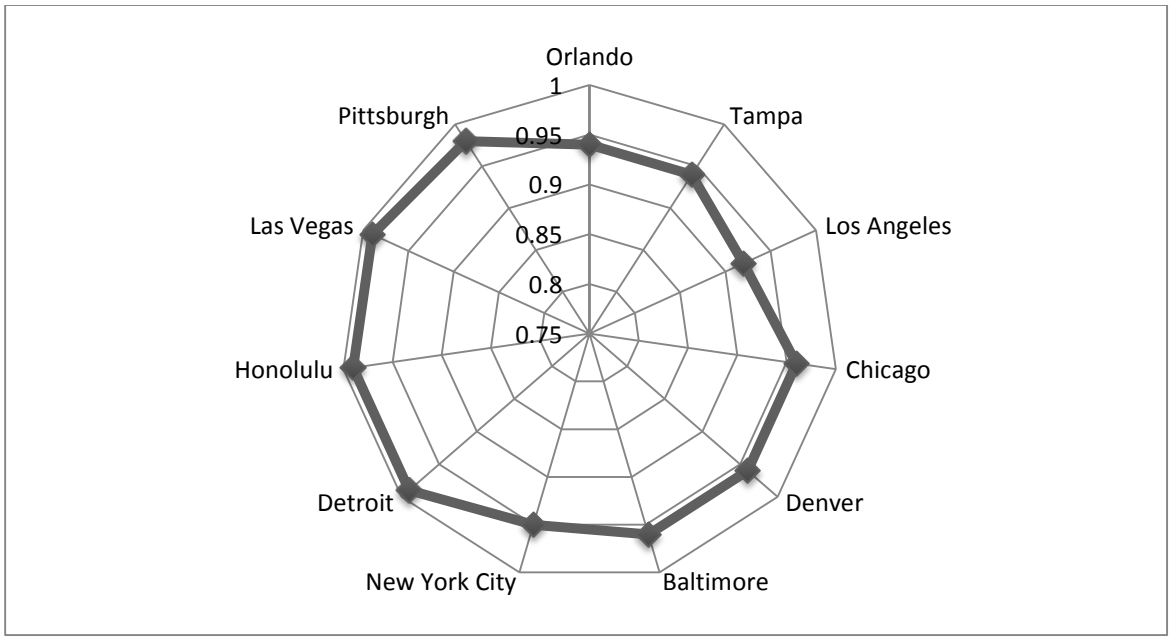

Fig. 7. Jaccard Distance Score for the Selected Cities of Attractions

Table 1

CVB Website URL of the selected cities

City	Website URL
Las Vegas	http://www.visitlasvegas.com/vegas/index.jsp
Honolulu	http://www.gohawaii.com/oahu/
Chicago	http://www.choosechicago.com/Pages/default.aspx
Los Angeles	http://www.discoverlosangeles.com/
New York City	http://nycvisit.com/
Orlando	http://www.orlandoinfo.com/
Tampa	http://www.visittampabay.com/
Denver	http://www.denver.org/
Detroit	http://www.visitdetroit.com/
Baltimore	http://baltimore.org/
Pittsburgh	http://www.visitpittsburgh.com/

Table 2

The result of word frequency for all the selected cities

City	The most popular 10 words from CVB website	The most popular 10 words from travelers
Orlando	Orlando, music, live, feature, dance, bar, downtown, restaurant, menu, entertainment	Orlando, Disney, fun, Florida, enjoy, entertainment, great, friend, dining, good
Tampa	Tampa, bay, art, culture, night, plaza, dining, Florida, fun, world	Tampa, beach, fun, sand, bay, art, shore, sunrise, food, sea
Los Angeles	Los Angeles, nightlife, event, sport, entertainment, bar, shopping, center, major, music	Los Angeles, fun, attraction, shopping, nightlife, dining, good, sport, bar, California
Chicago	Chicago, dining, music, food, hotel, choice, downtown, attraction, great, fun	Chicago, fun, lake, food, culture, dish, shopping, experience, downtown, bar
Denver	Denver, city, center, convenient, information, visitor, adventure, great, nightlife, art	Denver, nightlife, dining, fun, Colorado, attraction, good, city, music, mountain
Baltimore	Baltimore, city, restaurant, harbor, culture, history, new, Maryland, fun, easy	Baltimore, good, food, museum, bar, history, harbor, shop, nightlife, entertainment
New York City	New York city, experience, discover, information, personal, plan, event, find, shopping, dining	New York city, dining, art, finance, bar, attraction, fun, shopping, museum, restaurant
Detroit	Detroit, discount, attraction, deal, convention, hotel, offer, restaurant, fun, good	Detroit, culture, sports, auto, museum, food, hotel, attraction, history, home
Honolulu	Hawaii, Waikiki, pacific, beach, coast, harbor, history, Honolulu, surf, home	Hawaii, beach, seafood, ocean, dining, sun, fun, sea, island, attraction
Las Vegas	Las Vegas, plan, information, find, weather, explore, fun, perfect, shopping, casino	Las Vegas, casino, shopping, hotel, gamble, restaurant, strip, dining, nightlife, fun
Pittsburgh	Pittsburgh, park, city, find, perfect, enjoy, excitement, festival, night, fun	Pittsburgh, fun, history, dinning, sport, shopping, bar, entertainment, good, enjoy

Table 3

Jaccard Distance Score for all the selected cities

City	Overall	Shopping	Dining	Night life and Activity	Attractions
Orlando	0.86	0.84	0.84	0.94	0.94
Tampa	0.86	0.83	0.82	0.95	0.94
Los Angeles	0.87	0.85	0.85	0.95	0.92
Chicago	0.88	0.87	0.87	0.92	0.96
Denver	0.92	0.92	0.92	0.87	0.96
Baltimore	0.93	0.92	0.93	0.96	0.96
New York City	0.93	0.93	0.92	0.97	0.95
Detroit	0.96	0.95	0.95	0.87	0.99
Honolulu	0.98	0.99	0.99	0.97	0.99
Las Vegas	0.98	0.97	0.97	0.98	0.99
Pittsburgh	0.98	0.96	0.96	0.98	0.98

*Jaccard Distance: 0=identical, 1=completely different