IMPROVING TOURISM RECOMMENDER SYSTEM THROUGH QUANTIFYING REVIEWER CREDIBILITY

YUANYUAN WANG

Ph.D

The Hong Kong Polytechnic University

2015

The Hong Kong Polytechnic University

Department of Computing

Improving tourism recommender system through quantifying reviewer credibility

Yuanyuan Wang

A thesis submitted in partial fulfilment of the requirements

for the degree of Doctor of Philosophy

August 2014

Certificate of Originality

I hereby declare that this thesis is my own work and that, to the best of my knowledge and belief, it reproduces no material previously published or written, nor material that has been accepted for the award of any other degree or diploma, except where due acknowledgement has been made in the text.

_____(Signed)

WANG Yuanyuan (Name of student)

Dedicate to my parents.

Abstract

With the growing interconnectedness of the world and advances in transportation and communication, an increasing number of people are travelling as independent tourists, putting together their own itineraries and activities from information researched from social media. Moreover, a growing number of travellers post reviews and give ratings online to share experiences and opinions, which have become one important source of information. However, the explosive growth of travel information and the proliferation of uninformative, biased or false information make it very time-consuming and challenging for travellers to find helpful and credible information.

Recommender systems can assist travellers in managing the information available and facilitate their travel decisions. There have been some recommender systems developed in the tourism domain. However, these systems usually apply collaborative filtering-based, content-based or knowledge-based approaches, which require historical ratings, description of items, or extra knowledge about users' needs. It is difficult for them to generate reliable recommendation when the ratings, description as well as the knowledge are insufficient, and they cannot make recommendations when there is no such information. In the tourism domain, ratings, description and knowledge available are much fewer than the equivalent for books or movies. Therefore, they usually suffer from the sparseness and cold-start issues.

Addressing the sparseness issue, we apply rating inference method to augment ratings for rating-based recommender systems. We investigate several clustering

approaches to do sentiment analysis on travel reviews to generate numerical ratings. The clustering methods include K-means, co-clustering, hierarchical co-clustering, and six state-of-the-art traditional hierarchical clustering algorithms. Moreover, we compare different features extracted from reviews to choose more suitable features for each clustering method. Experimental results show that hierarchical algorithms (traditional hierarchical clustering and hierarchical co-clustering) with stepwise exploiting strategy lead to more accurate clustering results than non-hierarchical algorithms (K-means and co-clustering). Especially, hierarchical co-clustering method gets better clusters than all of the other clustering methods, no matter what features it uses. From the investigation, we also found that it is difficult to get very accurate multi-rating clusters by using these unsupervised approaches.

Rating inference on reviews can augment ratings for recommendation, but it is not helpful for solving the cold-start issue, since a new item or traveller has no review. Therefore, a demographic recommender system is applied to the recommendation of attractions to overcome the cold-start problem. Our system categorizes travellers using their demographic information and then makes recommendations based on demographic classes. Its advantage is that the history of ratings, description of attractions and extra knowledge are not necessary, so even a new traveller can obtain recommendations. Different machine learning methods are adopted to produce prediction of ratings, so as to determine whether these approaches and demographic information are suitable for providing recommendations. Our preliminary results show that these machine learning methods and demographic information can be used to predict travellers' ratings on attractions. But only limited accuracy can be achieved using demographic information alone.

Although recommender systems are able to provide travellers with recommendations, most of them make recommendations based on existing ratings or reviews, which may contain uninformative, biased or even false information.

Recommendations will be not so helpful or reliable if the recommender systems generate them based on these unreliable information. For instance, TripAdvisor, the world's largest travel community supplies a recommender system which can ranks reviews on an attraction for travellers based on posting dates or ratings. Travellers can then read some top ranked reviews on the attraction. However, there may be some incredible information involved in the top ranked reviews. Hence, it is critical to help travellers seek credible information from such amounts of travel information. Most current work applies mainly qualitative approaches to investigate the credibility of reviews or reviewers without quantitative evaluation.

This thesis presents a method that quantifies the credibility of reviewers, to help travellers find more credible information. We propose an Impact Index to quantify the credibility of reviewers by simultaneously evaluating the expertise and trustworthiness based on the number of reviews posted by reviewers and the number of helpful votes received by those reviews. Furthermore, Impact Index is enhanced into the Exposure-Impact Index by considering in addition the number of destinations on which the reviewer posted reviews. Our experimental results show that both methods perform better than the state-of-the-art method in discovering credible reviewers. To further examine the effectiveness and applicability of Impact Index and Exposure-Impact Index, we evaluate them on the data sets collected from two rather different online travel communities: TripAdvisor, the world's largest travel community, and Qunar, one of the most popular travel communities in China, by taking into consideration the differences between these two travel communities, such as different languages, scales and data distributions. Experimental results show that both Impact Index and Exposure-Impact Index lead to results more consistent with human judgments. They can not only discover more credible reviewers, but also provide better ranking of reviewers, which manifest their effectiveness and applicability across diverse travel communities.

Publications

The following papers, published or submitted, are the partial outputs of my PhD study in PolyU.

Journal paper:

- Yuanyuan Wang, Stephen Chi Fai Chan, Hong Va Leong, Grace Ngai and Norman Au. Multi-dimension reviewer credibility quantification across diverse travel communities, *Knowledge and Information Systems*, 2014. (submitted)

Conference papers:

- Yuanyuan Wang, Stephen Chi Fai Chan, Grace Ngai, and Hong Va Leong. Quantifying Reviewer Credibility in Online Tourism, In *The 24th International Conference on Database and Expert Systems Applications*, pp. 381-395. Springer, 2013.

- Yuanyuan Wang, Stephen Chi Fai Chan, Grace Ngai, and Hong Va Leong. Applicability of Demographic Recommender System to Tourist Attractions: A Case Study on Trip Advisor, In *Proceedings of the The 2012 IEEE/WIC/ACM International Joint Conferences on Web Intelligence and Intelligent Agent Technology*, pp. 97-101, 2012.

Acknowledgements

The endeavor of carrying out research is a fascinatingly non-isolated activity. I am grateful to the several individuals who have supported me in various ways during my PhD study and would like to hereby acknowledge their assistance.

First and foremost, I wish to express my deep thanks to my supervisor, Dr. Stephen Chi-Fai Chan, for your enlightening guidance, patient discussions and insightful ideas throughout the years. Thank for your continuous encouragement and support. I will always remember you showed me how to do presentation clearly and write papers accurately.

Furthermore, I would like to thank Dr. Grace Ngai, Dr. Hong-Va Leung and Dr. Alvin Chan, for all interesting discussions and directions they gave me on my studies. What I have benefited most from you is the rigorous and diligent attitude to research.

I also appreciate my colleagues, especially Wenhao Jiang, Weiping Zhu, Lin Zhang, Jiajia Li, Michael Huang, Hugo Sun, Will Tang, Kenneth Lo, Chi Kin LAU, and Andy Tam. Thank you for helping me a lot.

Thanks must be given to my friends, especially Cong Xie, Xin Sui, Xutao Li, Chunshan Li, Lei Qu, Xiaojuan Wang and Ting Fu. I am so lucky to be your friend.

Finally, I would like to express my special thanks to my parents, younger brother and my boyfriend. Thank you for your love, encouragement and support. To my dear grandma, thank you for your love, and raising me up.

Contents

List of Figures

List of Tables

List of Algorithms

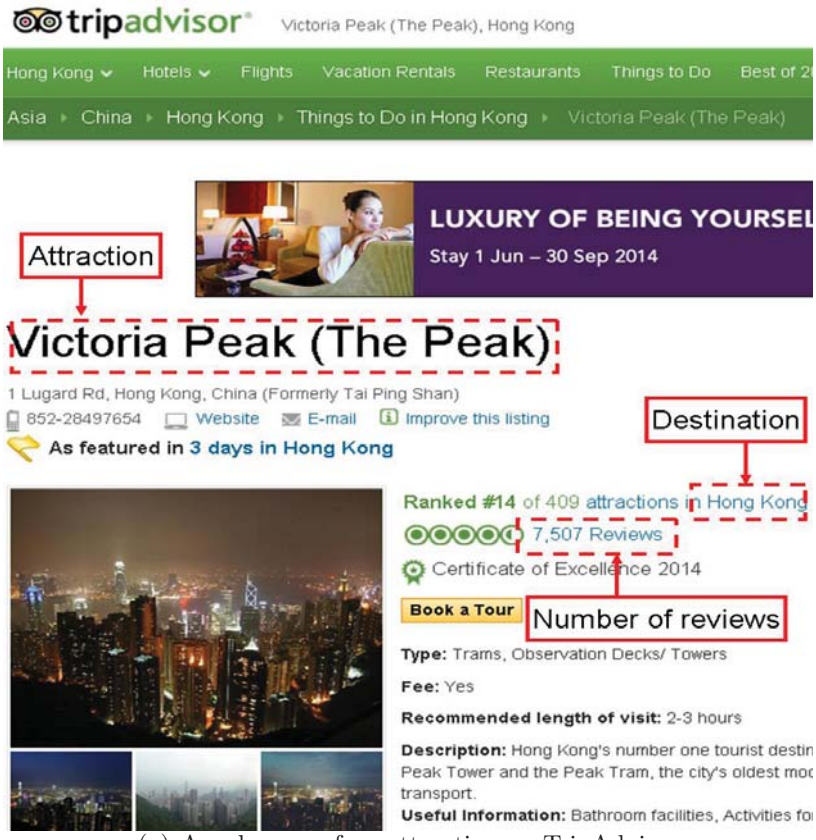
Chapter 1

Introduction

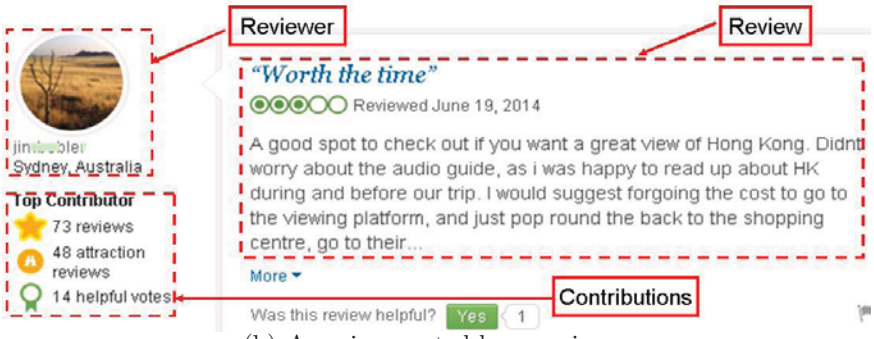
1.1 Background

Tourism has become one of the most popular e-commerce businesses online. Once upon a time, travellers rely on "official" travel information, such as travel magazines, travel agents, etc. However, with the development of social media technologies, anybody can post reviews and give ratings online to share their travel experiences and opinions, which have become one important online source of information. Meanwhile, the web helps any traveller access these travel information online to assist their travel decisions [63, 111, 121]. This has created a market for tourism websites, such as TripAdvisor, which is the world's largest travel community. For instance, as shown in Figure 1.1, TripAdvisor describes an attraction (Victoria Peak) in Hong Kong, and many reviewers post reviews on this attraction. Then new travellers can obtain information from these reviews.

However, tourism communities usually allow a reviewer to register at the websites using nickname and email address, without other identifying information, such as real name, photo and occupation. Even worse, it is not very easy to conduct a rigorous editorial process for factual verification [51, 71, 126]. For example, as shown in Figure 1.1(b), the reviewer uses a nickname to post a review without real name, personal photo or other identifying information. This leads to the explosive growth of



(a) A web page of an attraction on TripAdvisor



(b) A review posted by a reviewer

Figure 1.1: TripAdvisor website

information, and even the proliferation of uninformative, biased or false information, which make it very time-consuming and challenging for travellers to seek useful and credible information [126, 53, 70]. Take the attraction (Victoria Peak in Hong Kong) shown in Figure 1.1(a) as an example, it has received 7507 reviews as of July 19, 2014,

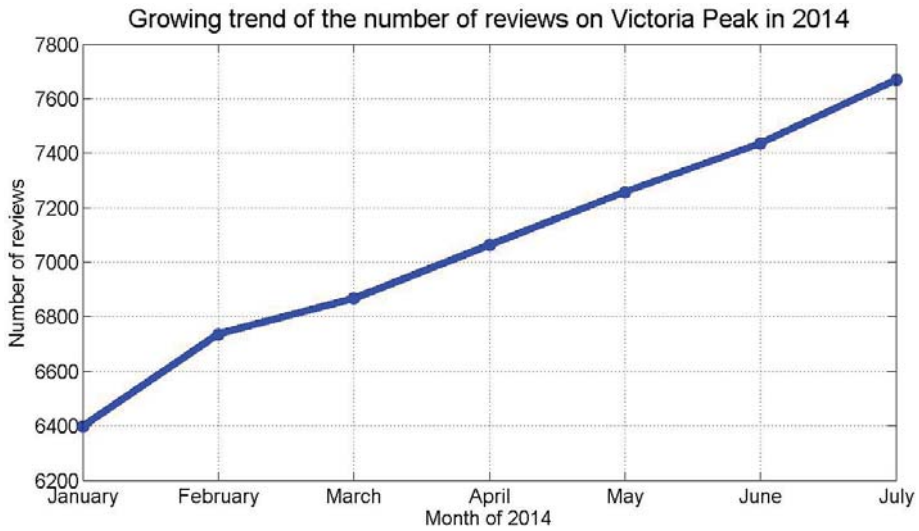


Figure 1.2: Growing trend of the number of reviews on an attraction

and the number is increasing quickly. For instance, monthly increase of the number of reviews on the attraction (Victoria Peak) is around 200 (as of July 31, 2014), as shown in Figure 1.2. Travellers will be very overwhelmed by the vast quantities of information provided from existing tourism communities.

1.2 Motivation

Different from books and movies, products and services in tourism area are mainly not physical and cannot be observed or experienced prior to purchase. They are intangible goods characterized by being inseparable, perishable and heterogeneous [42]. Due to such nature of tourism products and services, it is much more important for customers to search for information beforehand to help themselves make travel decisions. In the tourism domain, useful and credible travel information is crucial for travellers to reduce both financial and emotional risk [63].

With such huge amount of travel information, travellers usually need advice about which is useful and what to do. Recommender systems can assist travellers in managing the information available and facilitate travel decisions [94]. For instance,

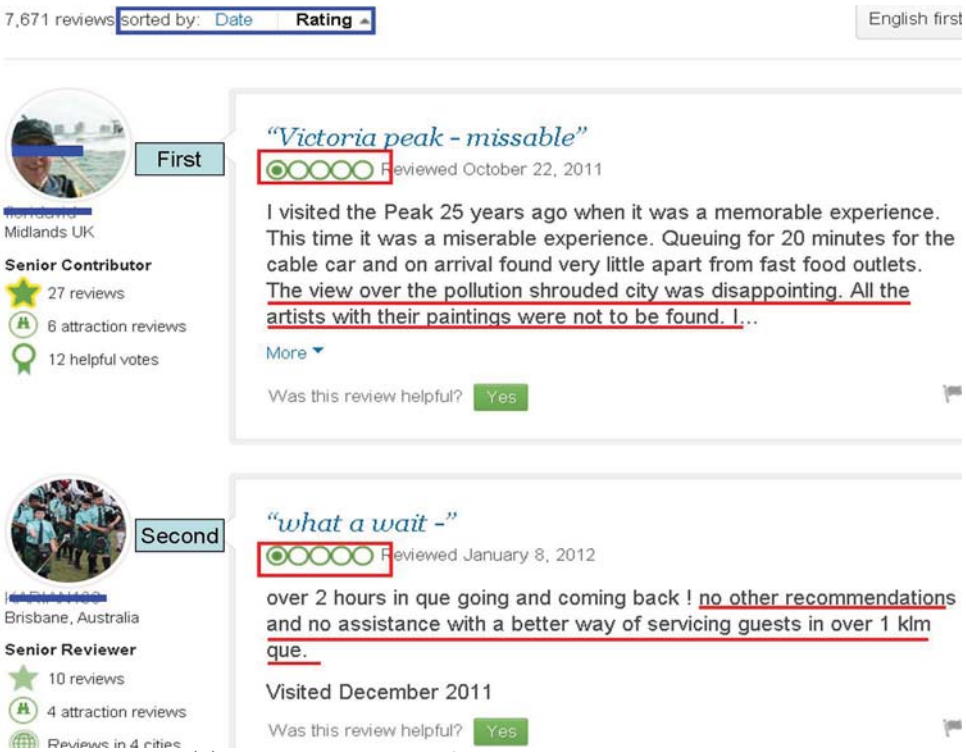
these systems can recommend places to go, things to do, hotels to stay in, etc. Existing recommender systems for classical application domains, such as books and movies, cannot be easily applied to the tourism domain. For instance, collaborative filtering method used by Amazon.com for book recommendations, relies on sufficient ratings given by similar users in a large community. However, individual travelling is not as frequent as books purchase or movie watching, and ratings available for them are much fewer than the equivalent for books or movies. Consequently, it is hard to model reasonable user preference profiles based on a user community in which users have similar taste and purchasing behavior .

Many tourism recommender systems have been developed as an integral part of some travel communities [89], such as TripAdvisor, TripleHop's SkiMatcher and Expedia's Inspiration Tool. However, most of them are collaborative filtering-based, content-based or knowledge-based approaches [94, 89, 41, 64], which need sufficient historical ratings, descriptions (textual document) of items or extra knowledge extracted from a conversational dialog. They usually suffer from the data sparseness issue that there is insufficient historical ratings for recommender system to learn user preference, and the cold-start issue that new users or new items have no historical rating. Moreover, it is time-consuming to learn the knowledge about travellers' requirements. For instance, Husain, W. et al. [41] presented a traveller recommender system integrating the content-based and collaborative filtering-based methods, which requires the descriptions of restaurants and historical ratings. The system can not provide good recommendations when there are sparseness and cold-start issues. Expedia's Inspiration Tool can help travellers to seek inspiring and exiting destinations [89]. It offers a choice of destination that meets travellers' requirements, based on their answers to simple questions. This tool needs extra knowledge about how an item satisfies a user's needs [12], which is extracted from a dialog.

Another issue of most existing recommender systems is that they are unable to handle the uninformative, biased or even false information. The recommendations will be not so helpful or reliable if the recommender systems generate them based on unreliable information. For instance, TripAdvisor, the world's largest travel community integrating recommender system as a part, assists travellers in their search for travel information. It provides travellers a QuickCheck-Tool to check hotel pricing, availability and ranks based on a popularity index that is evaluated by the quantity and quality of content written about the hotel on the web [89]. It ranks reviews on one attraction for travellers based on posted dates or ratings, so that travellers can just read some top ranked reviews. However, there may be some unreliable information involved in these reviews which will lead travellers to make bad decisions.

For example, Figure 1.3 shows three reviews on the attraction, Victoria Peak. They are ranked by ratings in ascending order, and their orders are first, second and fourth, respectively. When travellers read these reviews rated as "1", they need some negative opinions to know about the disadvantages of this attraction for facilitating their decisions. However, it is difficult to accept the information of the first two reviews (Figure 1.3(a)) as truth, since the description is not compressive or convincing, and the opinion expressed in these reviews is unfair and biased, especially the second one, in which the opinion is expressed in a very emotional and extreme way. Moreover, the review, as shown in Figure 1.3(b), expresses a very positive opinion, which is completely contrary to the rating. Therefore, these three reviews are not so credible. Travellers still need to read more other reviews to find helpful information. Hence, it is critical to find credible information from such amount of travel information.

Recommender systems in tourism are still in the developing stage, without achieving the level of success as in the domains of books or movies and insufficient to



(a) Reviews ranked in first and second, respectively

(b) The review ranked in fourth

Figure 1.3: Example of reviews on an attraction in TripAdvisor, which are ranked top by rating in ascending order

help travellers make travel decisions. Hence, this thesis focuses on improving tourism recommender systems by overcoming sparseness and cold-start issues, and especially, aiding travellers in finding credible information.

1.3 Objective

Our long term objective is to develop better recommender systems for the tourism domain. In this thesis, we aim at improving existing tourism recommender systems by addressing the issues of sparseness and cold-start recommendations, and the credibility of information. Particularly, we investigated the credibility of reviewers in detail.

To reduce the influence of the sparseness issue on rating-based recommender systems, we apply rating inference on reviews to augment ratings. Sentiment analysis is used to identify the overall sentiment of reviews and represent it as numerical rating. Existing studies on sentiment analysis mainly focus on semantic orientation and machine learning approaches. However, semantic orientation approaches need to compute the orientation and strength of opinions words, and machine learning approaches require enough labelled training data. In this thesis, we explore unsupervised methods to do sentiment analysis for rating inference. We investigate several popular clustering methods integrated in different feature extractions, hoping to generate additional information to improve the performance of ratings-based recommender systems.

Rating inference by sentiment analysis on reviews can augment ratings for recommender systems, but it is not helpful for overcoming the cold-start issue that a new item (attraction, restaurant, or hotel) or a new traveller has no historical rating for recommendation and no review for rating inference. Addressing cold-start issue, previous researchers have developed hybrid recommender systems based on collaborative filtering, content-based or knowledge-based methods, but did not try to make use of travellers' demographic information on their profile web page. Therefore, we employ demographic recommender system for the recommendation of attractions to alleviate the cold-start issue. Different machine learning approaches are examined

on the demographic information of reviewers to determine whether these approaches and demographic information alone are useful and effective to make prediction of the ratings.

Tourism recommender systems make recommendations for travellers on destinations, attractions, hotels or restaurants, mainly based on historical ratings, reviews, or extra knowledge extracted from dialog. If these information is uninformative, biased or even false, recommendations generated by existing recommender systems will be not reliable or helpful. Moreover, sentiment analysis on unreliable reviews will infer inaccurate ratings, which also lead to bad recommendations. Hence, it is more crucial to distinguish credible information from such amount of travel information. The main focus of this thesis is to handle uninformative, biased or false information for improving existing tourism recommender systems. To help travellers search for credible travel information, most of previous work just applied qualitative approaches to investigate the factors impacting the credibility of reviews or reviewers without quantitative evaluation. Lee et al [53] did a different work that they used Average RHR to quantify reviewer credibility. But this method tends to favor one dimension of reviewer credibility, which is the trustworthiness. Therefore, this thesis develops a method to quantify reviewer credibility, for the purpose of discovering credible reviewers automatically. Then travellers can get credible information to support their decisions. Furthermore, to examine the effectiveness and applicability of the proposed method across different travel communities, we evaluates them on the data sets collected from diverse travel communities.

1.4 Contributions

In this thesis, we work on improving existing tourism recommender system by addressing these issues: the sparseness and cold-start issue, and especially, the

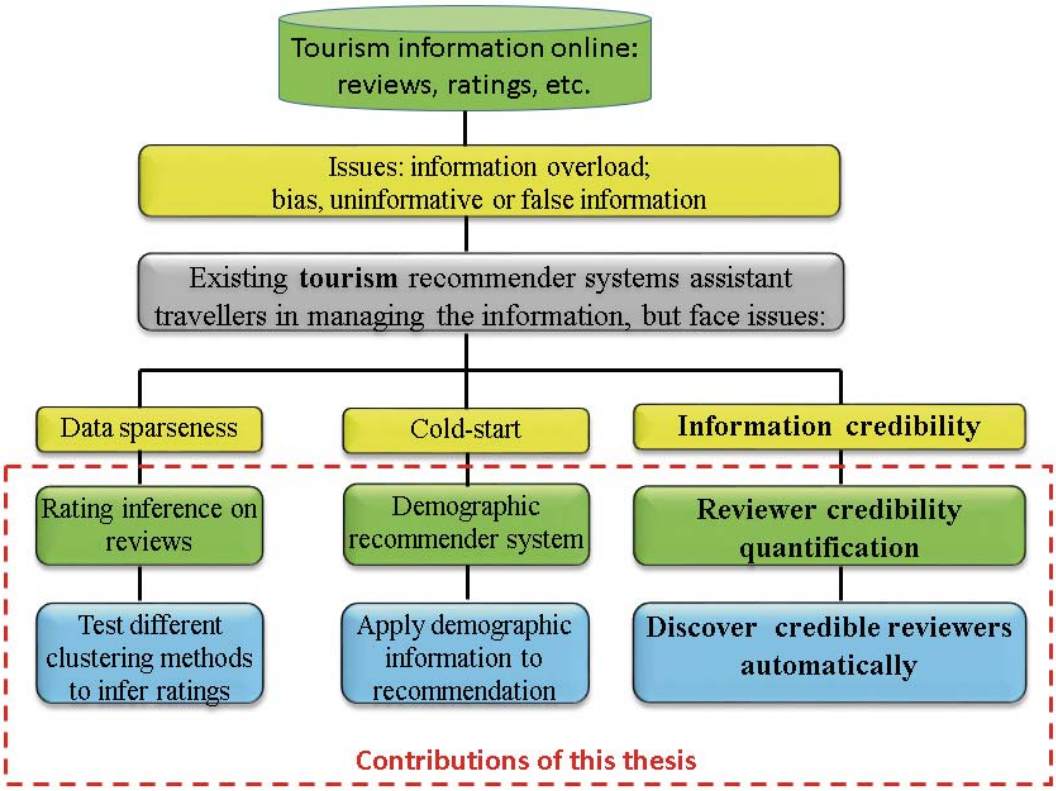


Figure 1.4: Framework

credibility issue. As shown in Figure 1.4. The original contributions of this thesis are as follows:

Rating inference by sentiment analysis on reviews: Addressing the sparseness issue, Chap. 3 presents a comparative study on several state-of-the-art clustering methods for sentiment analysis, in order to infer ratings for recommender systems. These clustering methods include K-means, co-clustering, traditional hierarchical clustering, and hierarchical co-clustering methods. Reviews with similar sentiment are clustered into one cluster with a numerical rating. It is important to transform the unstructured review text into numerical structured data which contains more useful key words that can manifest the sentiment orientation of the review. We compare different features used in the clustering process to find suitable

features for clustering methods. From preliminary experimental results, we found that hierarchical algorithms (traditional hierarchical clustering and hierarchical co-clustering) with stepwise exploiting strategy obtain more accurate clustering results than non-hierarchical algorithms (K-means and co-clustering). In particular, hierarchical co-clustering method gets the most accurate clusters, benefiting from its hierarchical strategy, feature clusters and feature reduction at each level, which may reduce the dimensionality. The results also suggested that only using Part-of-speech or opinion words obtained similar or even better results than those using all the words.

Demographic recommender system for tourist attractions: Addressing the cold-start issue, Chap. 4 investigates the applicability of demographic recommender algorithms for the prediction of ratings of tourist attractions. Based on the features extracted from travellers' demographic information, different machine learning methods are investigated to determine whether these approaches and demographic information are suitable for providing recommendations. we examine three machine learning approaches, including Naive Bayes, Bayesian Network and Support Vector Machine. Experimental results show that three machine learning methods based on demographic information performed better than the baseline method, especially SVM method. Our preliminary investigation result manifested that our demographic recommender system is indeed useful to make prediction of ratings, but demographic information alone is not sufficient to do accurate prediction of ratings, and more detailed experiment is needed to confirm our results.

Quantifying the credibility of reviewers in online tourism: Addressing the information credibility issue, Chap. 5 presents a method and its variant that quantitatively measure the credibility of reviewers in tourism. Based on previous literature on source credibility and message credibility, we build a positive linkage between reviewer credibility and review credibility. Then, our method is inspired by

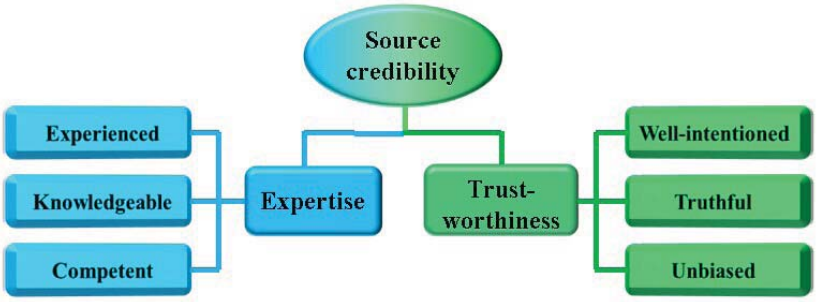


Figure 1.5: Key dimensions of source credibility

previous research on source credibility which was defined as two key dimensions: expertise and trustworthiness. The two dimensions are usually annotated with several words as shown in Figure 1.5. In the tourism domain, a reviewer is actually a source of travel information. Hence, reviewer credibility can be evaluated in terms of source credibility. As shown in Figure 1.6, we deduce that credible reviewer with high level of both expertise and trustworthiness should have posted many reviews, manifesting their expertise, and received many helpful votes (feedback from other travellers), implying their trustworthiness. Inspired by the idea of H-Index, we propose an Impact Index to compute reviewer credibility by evaluating the expertise and trustworthiness jointly, based on the number of reviews posted by the reviewer and the number of helpful votes received by the reviews. Compared to the previous Average RHR (average helpful vote) method, the Impact Index considers expertise and trustworthiness simultaneously, and does not emphasize one dimension only.

To better represent the multi-faceted nature of credibility, we further consider in addition the number of destinations on which a reviewer has posted reviews to improve the the model of credible reviewers based on the number of reviews. As shown in Figure 1.7, this new model evaluates the expertise of reviewer credibility directly referring to reviews' exposure, indicating rich experiences and broad knowledge at diverse destinations, which enable the reviewer to provide

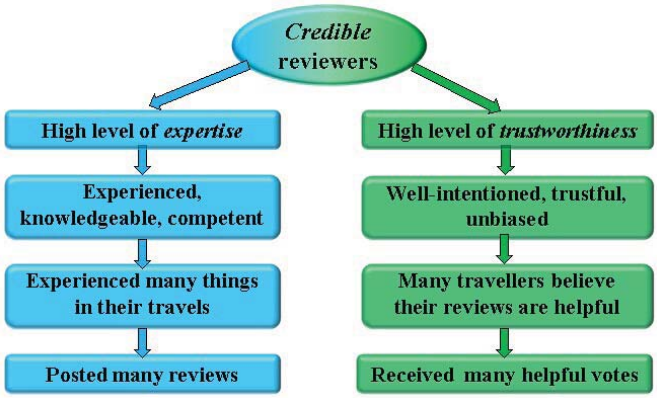


Figure 1.6: Model of credible reviewers in tourism

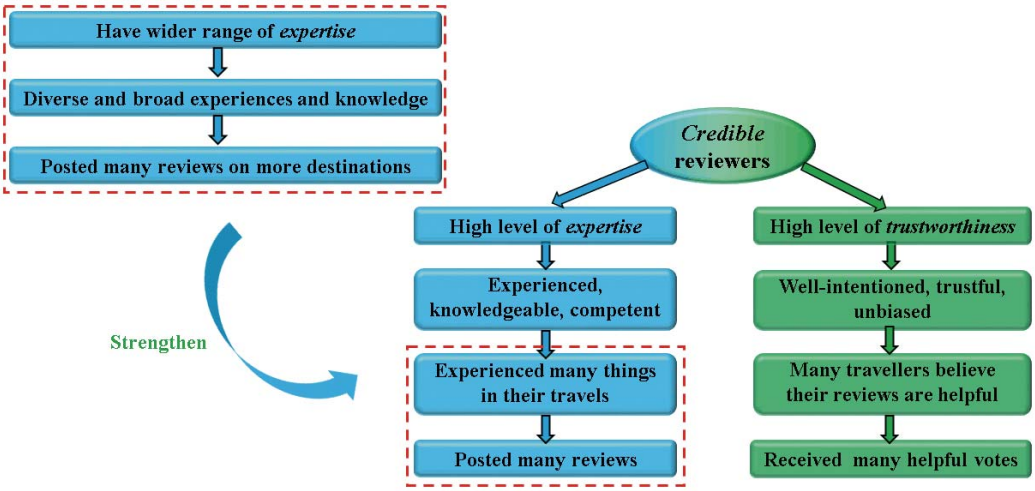


Figure 1.7: Improved model of credible reviewers in tourism

relatively more comprehensive and reliable information. According to the new model, the Impact Index is further improved into the Exposure-Impact Index by considering in addition the number of destinations on which a reviewer has posted reviews. Experimental results showed that both Impact Index and Exposure-Impact Index can discover more credible reviews than existing method.

In Chap. 6, we further validate the effectiveness and applicability of Impact Index and Exposure-Impact Index across diverse travel communities. We evaluate

them on several data sets collected from two travel communities: TripAdvisor, the worlds largest travel community, and Qunar, one of the most popular travel communities in China. We provided detailed comparisons between TripAdvisor and Qunar. Based on our analysis on the data, we discovered some possible manipulation behaviors. The manipulation issue is much worse in Qunar. Hence, we developed a helpful-vote cleansing method and a timestamp-based method to detect manipulation behavior and remove possible manipulators out. From the experimental result of the cleaned data, we found that both Impact Index and Exposure-Impact Index obtain results closer to expectation of human raters. They can not only discover more credible reviewers, but also provide better ranking of reviewers, which manifest their effectiveness and applicability for diverse travel communities.

1.5 Organization of the Thesis

The thesis is structured as follows.

- Chap. 2 is a literature review that provides background information for our discussions in the following chapters. It also reviews previous and related studies on sentiment analysis, recommendation systems and reviewer credibility.

- Chap. 3 investigates different clustering methods for sentiment analysis to infer ratings for recommender systems. Different feature selection methods are also examined to determine the suitable one for each method.

- Chap. 4 presents the demographic recommender systems to investigate the applicability of them for the prediction of ratings of tourism attractions.

- Chap. 5 proposes the Impact Index and Exposure-Impact Index methods to quantify reviewer credibility in the tourism domain. Both methods consider

two key dimensions of credibility: expertise and trustworthiness, based on the contribution factors of reviewers, such as the number of reviews, and the feedback of other travellers-helpful votes.

- Chap. 6 looks further into the effectiveness and applicability of Impact Index and Exposure-Impact Index on the data sets collected from diverse travel communities. It presents an deeper experimental evaluation of both methods, by taking into account the differences between those travel communities. Moreover, manipulation detection method is developed to clean data sets.

- Chap. 7 concludes the whole thesis and plans for the future research.

Chapter 2

Literature Review

2.1 Recommender System Overview

Recommender systems generate individualized recommendations and help the customers to search for interesting or useful objects. It was originally defined by Resnich and Varian in [92]. Recently, it has attracted much more attention in different domains [52, 105], such as financial services, electronic goods, restaurants or movies.

A recommender system includes three parts [99], including background data, input data, an algorithm. Background data is the information that the system should possess before execute the recommendation process; Input data should be given to the system so that a recommendation can be stimulated; An algorithm make suggestions based on the background data and the input data. According to these parts, recommender system can be categorized as three major types: collaborative filtering-based (CF-based), content-based and knowledge-based algorithms, and hybrid recommender systems.

2.1.1 Collaborative Filtering-based Recommender Systems

CF-based recommendation is probably the most widespread technology. It aggregates user ratings of items and then discovers additional items that the user may

be interested in. Tapestry [32] is a CF-based recommender systems which was implemented very early. In this system, users can record their annotations on the electronic document they have read. Such annotations may be interesting or uninteresting. Then these annotations can be accessed by other users in order to help them to decide which document to read. After that, CF-based recommender system achieved great success in both research and application [35, 50, 66, 93, 106, 104]. For instance, GroupLens [50, 93] is another popular system, which computes the similarity between users of Usenet newsgroups based on the comparison of their ratings of news articles they have read. It then predicts the user's interest in a new article based on the ratings given by similar users (nearest neighbors) who have read that before.

Traditional CF-based recommendation techniques can be classified into two types: user-based and item-based. User-based collaborative filtering recommendation predicts if current user would like a target item based on this user's rating data and the preferences of her/his neighbors. In contrast, item-based collaborative filtering recommendation [102] makes the prediction for current user based on if those items that (s)he has previously rated are similar to the target item.

Collaborative recommender systems have several advantages. Firstly, they are independent of any computer-parsable representation of the items because they do not take into account the content information of items or domain knowledge. Hence, they can make recommendations for complex items, such as music and movies. Secondly, they make recommendations based on users preference indicated by the ratings rather than the properties of items. Therefore, they can recommend items very different from those current user has previously shown a preference for. Thirdly, they are easier to be understood and implemented than other types of systems. They can be applied to numerous domains where the user preferences (ratings) are available.

Collaborative recommender systems also suffer from several issues [4, 55, 74] despite their success and popularity. Some main problems are as follows:

Sparseness issue: in practice, the number of ratings obtained by the items is usually very small. For example, on Amazon.com, the number of books that an active user has purchased is far less than the number of the total books. Also, there may be books that have been rated by only few users. CF-based systems make recommendations based on similarities between users or items. Such similarities are computed according to the co-rated items. In a system with large user and items (like Amozon.com), the number of co-rated items will be very limited. This may lead to less reliable computation of similarity, which result in poor recommendation. Some strategies have been used to overcome this issue. One strategy is to use the profile information of the user to calculate user similarity. For example, Pazzani [87] used the gender, education, age, place, and employment information of users to recommend restaurants for them. This method is called as demographic-based recommendation. Another approach is to augment rating making use of other information. Leung et al. proposed a rating inference approach [56] that used the textual review to do sentiment analysis in order to augment the rating. That approach represents the overall polarity of opinions in the reviews as numerical ratings. Besides, dimensionality reduction methods [100, 101], for example Singular Value decomposition (SVD), was also applied to reduce the dimensionality of the sparse rating matrices. Moreover, hybrid content-based and CF-based recommendation, which incorporates content-related features about items into collaborative filters, is also an effective and popular strategy. We will discuss this method in another subsection next.

Cold-start issue: traditional CF-based recommender should learn user's preferences from previous ratings of the user in order to make more accurate recommendation. Therefore, it cannot recommend new items that have not been rated by any user. It also cannot generate recommendation to new users who did

not rate any item. Particularly, this problem may happen when no similar neighbor can be found for items or users because of the lack of overlapping preferences. Those methods for addressing data sparseness problem can also be used to solve this problem, because this problem is actually an extreme of data sparseness. Most of existing methods to handle this problem use the hybrid recommendation approach, which combines the contented-based and CF-based techniques.

Scalability issue: user-based method uses the nearest neighbor algorithms which make statistical computation over the entire database of user preference. Those algorithms are memory-based and require the computation to grow with both the number of users and the number of items. Traditional user-based method will suffer from serious scalability problems because of a huge number of users and items. Addressing this problem, several model-based CF-based algorithms [102] have been proposed to construct compact models about users or items using various data mining methods, and then generate recommendation based on the compact models rather than the entire database.

2.1.2 Content-based Recommender Systems

Content-based recommender systems recommend an item to a user using the description information of items rated by the user and the description of the item to be recommended. That is, the system will make recommendations on items which are similar to those the user likes. User profiles are learned by making use of features extracted from these items. Lots of current content-based systems are mainly recommending items described by textual information, such as Usenet news messages and documents. Hence, many algorithms have been proposed to analyze the content of textual documents, using the vector model of the words in the text as features [1, 4, 86]. For example, Fab system [4] makes recommendation on web pages for users, by representing the web page content as the 100 most important words with

the highest TF-DF value. Syskill and Webert system [86] uses 128 most informative words to represent documents. These words are more associated with one class of documents than another. There are several other methods to extract important words as the features, such as Mutual information, and IDF. Other weighting schemes can also be used to give high weights to discriminating words.

Once the document content has been represented, some algorithms can be employed to learn the user profile model and predicts whether the user would be interested in a new item. Constructing user profile (or preference) can be viewed as a form of classification process [85]. The classification algorithms generate a function that can make an evaluation of the probability that the user will like the new item. Several traditional classification approaches have been used to do content-based recommendation, such as Bayesian classifier used in Syskill and Webert system [86], Rocchio's algorithm used in Fab [4], decision tree, and nearest neighbor methods.

The main advantage of content-based recommender systems is that they can make recommendations for all domain items as long as the content information of them is available. However, a pure content-based recommendation has several drawbacks: firstly, content-based approaches depend on the information about items which is usually extracted from various sources or manual defined. However, the representation of the information can only captures part of the content. Some other attributes that influence the user's choice may be ignored. Especially, there are some items have no textual information. Hence, the representation, without enough information to distinguish items that a user interests in from those the user does not like, will reduce the effectiveness of the system. Secondly, content-based techniques suffer from over-specialization problem. They can only recommend items similar to those a user has previously seen [4], because they make recommendations based on the user profile. Thirdly, these technologies have the cold-start issue. Sufficient number of items should be rated by the user before the systems construct the

user's profile. Therefore, a new user without enough ratings can not get accurate recommendation.

2.1.3 Knowledge-based Recommender Systems

Knowledge-based recommender systems use knowledge about users and items to generate recommendation, finding what items meet a user's requirements. They are commonly used in the domains where the decision making process of users is more complicated and constrained. One example of these systems is the restaurant recommender Entree [11] that recommends a user restaurants in a new city which are similar to those the user likes. For example, this system can give a restaurant guide for Chicago. Users choose some options from a set of menu to describe what they want in a restaurant, or type in a restaurant in some other city they like. Then the system retrieves restaurants in Chicago that are similar to user's choice. Various types of knowledge can be used by the systems to know the needs of users.

Knowledge-based systems do not suffer from sparseness or cold-start issues, because they are independent from rating information or other historical data. They can make wide recommendations if knowledge allows. The main limitation is that they need domain engineering process to build knowledge bases. Three types of knowledge are involved in these systems: one is catalog knowledge which is about the items being recommended and their features; Another one is functional knowledge which can match user's needs to the corresponding items that may satisfy those needs; The last one is user knowledge which is the information about user. Another shortcoming is that the recommendations they make are rather static [99].

2.1.4 Hybrid Recommender Systems

Hybrid recommender systems integrate two or more recommendation methods in order to avoid certain limitations of any individual one. The most common

combination is combing the CF-based and content-based methods [4, 8, 12, 18, 87, 109]. On one hand, the CF-based method can overcome some shortcomings of content-based method. For example, CF-based method can help content-based method to solve the over-specific problem by recommending the items similar to those current user's neighbors like. Therefore, some items can be different from those current user has seen before. On the other hand, content-based method can help CF-based method to avoid some shortcomings. For instance, content-based method can settle the early-rater problem of CF-based method, because the content-based system can recommend new items on the basis of their content, without the need for ratings.

To combine these recommender systems, different combination methods can be employed [99]:

Firstly, different recommender systems are implemented separately, and all results of these recommender systems are combined into a score of a recommended item. For example, the P-Tango system [18] makes a recommendation based on a weighted average of the CF-based and content-based methods, which can benefit from individual advantage of both methods.

Secondly, the system makes use of some criterions to switch between different recommendation methods according to current situation. The DailyLearner system [8] applies the content-based method first and switches the method between content-based and collaborative filtering methods. The collaborative method will be attempted when the content-based method can not generate a recommendation with sufficient confidence. This combination need a switching criteria to be determined, which make recommendation more complex.

Thirdly, recommendations from different recommender techniques are presented together. This method can be applied to make large number of recommendations simultaneously. For instance, the PTV system [109] uses the content-based and

CF-based techniques to recommend programs, and combines their recommendation results. This mixed hybrid can overcome the new-item cold-start problem because of the content-based method, but fails to solve the new-user problem since both methods need some historical data about user preference.

Fourthly, the system combines CF-based and content based methods by treating collaborative information as additional feature of content-based techniques, which is called feature combination hybrid. For instance, Basu et al. exploited both the ratings and content information to do recommendation [6]. This method reduces the sensitivity of the system because it just considers the rating data without relying on it.

Fifthly, the recommendation of one system is used as the input to next system. This method can be called feature augmentation. For instance, the Libra system[75] integrates the information generated by the internal collaborative systems of Amazon to make content-based recommendation of books.

Sixthly, a rough ranking of candidates is generated by one recommendation method firstly, and next method will refine the ranking. For instance, the restaurant recommender EntreeC [99] uses the knowledge about restaurants to make recommendations according to the interests of users, and then assign those recommendation candidates into different buckets with equal preference. After that, the collaborative filtering method is applied to further rank the candidates in each bucket.

Seventhly, entire model generated by the first method is used as the input of next one. For instance, Pazzani [87] built a content-based model to describe the features about user preference. This model is then used to a collaborative method to make prediction. Hence, the model can be viewed as compressed representation of user's interest, which make the collaborative technique work more easily on it than the raw rating data.

2.1.5 Recommendation Technologies in Tourism

Travel and tourism industry is the leading application field in the B2C e-commerce. Different from other online transactions, the product and service in the tourism domain are mainly not physical and exist mostly as information. Travellers can only make travel decisions according to the description about the tourism destination or attraction. In addition, there are various options for travellers, including different cities and places, different kinds of attractions. Moreover, with the rapid development of social media technologies, an increasing number of travellers like to post reviews and give ratings online to share travel experiences and opinions, which have become one important source of information. With such huge options and information, travellers usually need advice to help make decisions. Recommender systems can provide advice about cities to visit, places to go, attractions to see, events to participate, options for hotel, etc. Therefore, recommender system for the tourism domain has become an attractive research area that deserve to be exploited.

The complexity of the concepts and the decision process involved in travel recommendation challenges for the design of usable and effective recommender systems. Moreover, it is hard to establish reasonable user profiles. A lot of famous travel communities integrated recommendation systems as an part of them. For instance, Triplehop's TripMatcher and VacationCoach's Me-Print use a content-based method to make recommendation [94]. However, these systems only apply one of traditional filtering approaches, such as CF-based, content-based and knowledge-based methods [94, 89, 41, 64], which usually require enough historical rating information, descriptions of items or extra knowledge and suffer from data sparseness and cold-start problems. Moreover, these approaches are insufficient when trying to make recommendations for the complex travel products.

There have been several studies on developing tourism recommender systems.

Many systems try to capture users' needs by extracting the preferences and requirements in a conversational dialog. For instance, Stanley et al. [64] presented a recommender system that help travel agents discover suitable options for travellers according to collaboration and text analysis. This system uses text mining method to identify interesting areas from the messages of web chat between a travel agent and a traveller. Then, the system searches for a database for tourist options classified in these interesting areas. Ricci et al. [96] proposed an approach to generate mobile travel recommendation based on the interactive elicitation of users' needs and wants by critiques which is obtained through a dialog. The system involves users in a dialog about a candidate product, and makes users express critiques on the product. After that, the critiques are interpreted in users' initialized preferences model. According to this initialized preference of users, the system will provide a list of recommended products. Then, there will be three situations. First, if one (or some) of the recommended products satisfies users' needs, the recommendation process will be terminated successfully. Second, users may be somewhat interested in one of the products, but not like one or some features of the product. Then, users can criticize the product and better specify their preferences. Such critiques are used to adapt current query, and generate a new recommendation. Third, none of the products recommended are liked by users, then users terminate the system with failure. These recommender systems can provide travellers recommendations, but require extra knowledge obtained from dialogs.

Previous research has concluded that hybrid approaches to recommendation are more effective, because this approach is not strictly limited to one type of information or data [99]. Hence, several hybrid approaches have been developed in tourism area [95, 129, 120]. For instance, Ricci et al. [95] presented a personality recommendation system that integrates content-based methods, collaborative filtering techniques and case-based reasoning. This system is divided into three stages: the first stage is and

acquisition of travel preferences. The system asks users to provide some information about personal and travel characteristics, such as budget, travel period and type of accommodation, which are called collaborative features; the second stage is searching for travel products. Users start the process by seeking a destination or a product that meets their requirements, such as a budget hotel close to a lake, which are defined as content features; In the third stage, the system rank the results obtained from a successful query and present them to the user.

Zanker et al. [129] proposed a hybrid framework for computing item similarity based on information retrieval and case-based recommendation systems and then enrich it with additional knowledge-based concepts. Another example is that Wietsma et al. [120] applied a hybrid recommendation method with the textual reviews as a decision making aid to recommender tourist attractions. They aimed at designing a methodology for mobile recommender systems that incorporates different knowledge source and offer better recommendations. The knowledge source could be structured or not, including the description information of the product and the reviews. In their approach, if the product to be recommended has a rich structured description, then the system tends to use the content-based method. On the contrary, if the product is poorly described, then the system relies more on collaborative method.

2.2 Sentiment Analysis Approaches

With the rapid growth of the personal opinions on the internet, such as ratings, reviews and other forms of online information, the detection and analysis of opinions, feelings, or attitudes expressed in a text has attracted more and more attentions. One of the most useful techniques is sentiment analysis which classifies a text based on the semantic or sentimental orientation of the opinions it carries [62, 76, 84, 117]. It

can be applied to several areas, such as business and government intelligence, and review-related websites [23, 82, 110]. Hence, there have been a number of techniques adapted from different disciplines to do sentiment analysis. Among these techniques, the sentiment orientation methods (also called lexicon-based methods) and machine learning methods have received considerable research and were popularly applied in various tasks [9, 20, 56, 82, 46, 79, 48, 57, 84, 117].

2.2.1 Semantic Orientation Approach

Semantic orientation approach is a very common technique to do sentiment analysis. Turney [117] proposed a algorithm to determine the overall sentiment of reviews by the average semantic orientation of phrases in reviews that contains adjective or verbs. Firstly, a Part-of-speech tagger extracts phrases. After that, the algorithm estimates the semantic orientation of each phrase. The mutual information between the phrase and the reference word "excellent" and "poor" is used to calculate the sentimental orientation of a phrase. Finally, the review is assigned a class in terms of the average semantic orientation of the phrases identified from the review. If the average semantic orientation is positive, the review is recommended, otherwise, not recommended. Kamps et al. [46] determined the orientation of a word by WordNet, since words in Wordnet are connected by synonymy relations. Kim et al. [48] also made use of the synonymy structure in WordNet to compute the strength of the sentimental orientation by probability. Moreover, Ohana et al. [79] applied SentiWordNet lexical resource to automatically classify the sentiment of film reviews. Each term of SentiWordnet is associated with a numerical score to represent positive and negative sentiment information. Their method integrates positive and negative term scores to determine sentiment orientation.

Dave et al. [20] presented a method to automatically distinguish positive and negative reviews by distinguishing suitable features and scoring methods from

information retrieval. N-grams is used to do feature extractions, and then a threshold is used to filter the feature. After that, a score is computed for every feature to indicate its semantic orientation. To predict the sentimental orientation of a document, a value (a positive value represent positive and vice versa) is added to its features that also occurs in the feature dictionary generated before. In addition, Leung et al. [56, 57] proposed a method that constructs a feature dictionary using a training set to help the binary sentimental orientation. This method applies the Part-of-speech, Negation Tagging and feature generalization to extract features for feature dictionary construction. Then two values reflecting the opinion strength of positive and negative sentimental orientation are assigned to every adjective. Subsequently, aggregations of all occurring adjectives for the negative and positive sentimental orientation are computed respectively, and the high aggregation indicates the sentiment of the overall document.

2.2.2 Machine Learning Approach

Machine learning technique has been usually used in topic-based text classification. Recently, it is applied to the sentiment classification problem. Pang et al. [84] is the first team that employed machine learning approaches, including Maximum Entropy, Naive Bayes and Support Vector Machine (SVM) to do sentiment analysis. The method represents a review as a feature vector based on Unigrams and Bigrams and trains machine learning models from a set of movie reviews labeled as positive and negative. Their experimental results showed that SVM trained on Unigram features outperforms other methods. After this work, a large number of researchers [9, 90, 2, 14, 47, 123, 125, 115, 83, 80, 107] investigated a wide range of machine learning approaches to do sentiment analysis. Among them, Naive Byes [84, 9, 2, 14, 47, 123, 125], SVM [84, 83], and Maximum Entropy [84, 9] have archived great success in sentiment analysis. In addition, other popular machine learning

approaches were also utilized by existing work, such as Decision Tree (C4.5) [2, 14], K-Nearest Neighbour (KNN) [115].

Feature extraction is an issue needs to be handled by machine learning-based sentiment classification. There have been several studies that investigated different methods to resolve the issue [84]. The most commonly used features include Unigram, N-gram, Part-of-speech information, negations and opinion words (phrases)[84, 82, 9, 14, 47, 107].

2.3 Credibility Overview

2.3.1 Credibility

According to uncertainty reduction theory, individuals who face uncertainty will attempt to execute uncertainty reduction strategies [7]. They will try to reduce uncertainty by searching for credible information to facilitate decision making [43]. Credibility can be defined as believability [29], and has been investigated in diverse areas, such as communication, information science, psychology, etc. It was systematically investigated beginning with the communication area. Hovland et al. [37, 38] was a landmark that examined dimensions of source credibility. After that, credibility attracted many studies attention, and has been discussed in three aspects: source credibility, message credibility and media credibility [71, 49, 27, 113].

Previous studies on source credibility has focused on the expertise or trustworthiness of the source as two key dimension to determine if the source is able to provide credible information [3, 27, 28, 29], since the "Yale Group", led by Carl Hovland defined it as expertise and trustworthiness [37]. Message credibility is the assessed credibility of the information itself or communicated message, such as information accuracy, quality, or currency [71, 88]. For instance, well-organized and comprehensive message or information is more persuasive to individuals [88, 54].

Medium credibility is the perception of credibility on a specific medium, which disseminates message , such as newspapers, television, the Internet, or blogs [113].

Previous literatures [98, 31] have pointed out that credibility assessment of sources and messages are fundamentally and positively interlinked and influence each other. That means, credible sources tends to generate credible messages and credible message tend to originate from credible sources [31]. Moreover, the assessment of source credibility and media credibility are also interlinked. However, this is less clear than the relationship between source credibility and message credibility [98].

2.3.2 Source Credibility

Jensen et al. [45] pointed out that the identity of information source assists individuals in finding people who like themselves. Lee et al. [127] investigated the role of source characteristics in e-word-of-mouth. Their experimental results showed information/cues regarding source characteristics were important to the information seekers. Although many studies on source credibility have explored several different dimensions, the focus is still the initial two dimensions: expertise and trustworthiness [29, 37, 27, 88, 54, 30].

Fogg et al. [29] defined expertise as knowledgeable, experienced, and component. They said that this dimension can capture the perceived skill and knowledge of the source. The trustworthiness was defined by well-intentioned, trustful, unbiased. They also pointed out that this dimension captures the perceived morality or goodness of the source. Fogg et al. [54] investigated the factors affect users' assessment of the credibility of source. They found that comprehensive information, shared value and diverse communication affect the perception of credibility. This team also conducted a study to investigated these two dimension. They found that expertise is indicated through the accuracy and comprehensiveness of information, professionalism, and credentials [30].

Flanagin et al. [27] pointed out that expertise is not only a subjective perception, but also includes some relatively objective characteristics of the source as well as message, such as information quality and source credentials. Trustworthiness is judged based primarily on subjective factors.

Cho et al. [16] defined expertise that an information source is perceived to be capable of generating correct information. They also described trustworthiness as that it implies the degree to which an information source is perceived as providing information that reflects the source's actual feelings or opinions.

2.3.3 Credibility Issues in Tourism

In tourism, the products are the intangible and experiential service purchases, which can not be evaluated before their consumption. Therefore, travelers tend to search for information before travelling to reduce the degree of uncertainty and risk, and facilitate decision making [111, 60]. The development of social media technologies and travelers' willingness to post reviews online sharing experience and opinion have created created a market for communities, such as TripAdvisor, the most popular tourism community in the world, enabling millions of travellers to post reviews and search for travel information. However, as Kusumasondjaja et al. [51] pointed out, tourism communities usually lack the mechanism to rigorously verify the identity of reviewers and the content of reviews, which allow users to register without providing real identify information to post reviews without verification for correction. This leads to the explosive growth of reviews and the presence of uninformative, biased or even false information, which makes it very time consuming and challenging for the travellers to find credible reviews [53, 70].

2.3.4 Credibility Assessments in Tourism

To address the credibility problem, some researchers have investigated the factors that affect the perception of the credibility of the review [51, 124, 108, 34]. For instance, Kusumasondjaja et al. [51] investigated the impact of the review valence and the reviewer's identity on the perception of credibility, and found that a negative review with the identity of the reviewer disclosed can enhance the perceived credibility of the review. The work of Xie et al. [124] indicated that the hotel reviews with the presence of personal profile information were perceived more credible by the travelers. Sidali et al. [108] found that a review to be trusted must be assessed as an expert The study conducted by Gretzel et al. [34] indicated that the type of website on which the review is posted, the detailed description, and the date the review was posted are very important for evaluation of a travel review. However, most of the previous work developed qualitative guidelines based on surveys to help the travellers distinguish credible review, without developing a method to search credible reviews automatically.

In terms of these two dimensions of source credibility, there has been much work on investigating the credibility of reviewers. Gretzel et al. [34] discovered that reviewer's credibility is most frequently judged based on the reviewer's travel experience. The result of the survey conducted by Sidali et al. [108] showed that the number of posted review and travelling a lot are important to judge the expertness. Vermeulen et al. [118] have applied experience as proxy of expertise. However, these studies fail to make a quantitative evaluation of the reviewer credibility. Additionally, Lee et al. [53] used the average helpful vote (Average RHR), which is a feedback received by the review from other travelers, to represent the credibility of reviewers. This approach can evaluate the reviewer credibility quantitatively, but it tends to emphasize the trustworthiness only.

Chapter 3

Rating Inference by Sentiment Analysis with Clustering Methods

3.1 Introduction

As the explosive growth of reviews in online tourism, travellers are always overwhelmed by such huge amount of information. Recommender systems are able to assist travellers in making travel decisions. However, rating-based recommender systems, such as collaborative filtering method, usually suffer from sparseness issue that there is insufficient ratings to learn user preference [4, 55].

Sentiment analysis is used to identify reviewers' opinions from their reviews [62, 76, 84]. Opinions are usually classified into binary classes, including positive and negative. Recently, some new studies extended the binary sentiment classification into multi-rating scales, such as "1" to "5" stars or points [56, 58, 83, 80]. The ratings inferred from reviews can be used in recommender systems to augment historical ratings [56]. Hence, this chapter focuses on estimating the overall sentiments expressed in reviews and representing those sentiments in multi-rating scales, in order to reduce the sparseness problem of recommender systems.

There have been many sentiment analysis approaches from different disciplines to determine the overall sentiment of reviews [9, 20, 46, 48, 79, 82, 117]. Among

these approaches, semantic orientation (lexicon-based) method and machine learning method were popularly applied in various tasks. Semantic orientation method aggregates the semantic orientation of the opinion words in a review based on an opinion dictionary or a lexicon to get the overall sentiment of the review [46, 48, 56, 79, 117]. However, It is difficult for semantic orientation method to find opinion words with domain specific orientations. The opinion words is quite common [61] and does not adapt well to different domains [68]. Different machine learning methods have been applied to classify reviews into different sentimental classes based on the labeled reviews, including Maximum Entropy, Naive Bayes, Support Vector Machine, K-Nearest Neighbour, etc [2, 9, 14, 47, 83, 90, 115, 123, 125]. Although the machine learning method is effective to the sentiment analysis, it requires a large number of labelled training data, which is usually costly and time consuming to acquire. Therefore, it is necessary to develop unsupervised or very weakly supervised methods for sentiment analysis.

This work is a comparative study of several clustering methods for doing sentiment analysis. The clustering methods include K-means [65], co-clustering [22], traditional hierarchical clustering [130], and hierarchical co-clustering method [119]. We attempt to cluster the reviews with similar sentiment into one cluster by these clustering methods without any labeled data and other extra knowledge. It is especially important to transform the unstructured review text into numerical structured data which contains more useful key words that can manifest the sentimental orientation of the review. We compare different features for the clustering process in order to find which method can extract more useful opinion words to facilitate these clustering methods. The features we tried include word Unigram, words in Part of Speech, and opinion words from the dictionary. In this work, we focus on analyzing the reviews on the attractions (things to do) of tourism area, while previous work are mainly on products or movie reviews [81, 84].

The rest of this chapter is organized as follows. Chapter 3.2 presents different clustering approaches. In Chapter 3.3, different types of features are presented for clustering. In Chapter 3.4, comparison experiments are presented to investigate the effectiveness of different clustering methods on different features for sentiment analysis. Chapter 3.5 includes our conclusions.

3.2 Clustering Approaches

We investigate several clustering methods, most of which are very famous in topic-based text clustering, including K-means, co-clustering, hierarchical clustering, and hierarchical co-clustering.

3.2.1 K-Means

K-means algorithm [65] is an iterative method commonly used to cluster a data set into a predefined k clusters. This method is initialized by selecting k data points as initial cluster centers. Methods for choosing the initial seeds include randomly picking from the data set, setting them by clustering a small subset of the data or perturbing the global mean of the data k times. Then the method iteratively refines the centers as follows:

Step 1: Each data point is assigned to its closest cluster center.

Step 2: Each cluster center is recalculated to be the mean of those data points assigned to it.

The algorithm converges when the assignment of data points to clusters no further changes. It is so simple and easily implemented that it is the most widely used partitional clustering method. However, it is quite sensitive to the initialization and the outliers.

3.2.2 Hierarchical Clustering

Hierarchical clustering is usually based on agglomerative and partitional algorithms, which arrange data set in the form of a tree to provide a view of the data at different levels of abstraction. Previous researchers have studied and evaluated different hierarchical methods, and found that partitional algorithms usually perform better than agglomerative algorithms for high dimensional data [131]. Therefore, this work apply six state-of-the-art partitional algorithms evaluated by Zhao et al. [130] to cluster reviews. These methods compute a hierarchical clustering solution using a repeated cluster bisection approach which optimizes a global criterion function given in Table 3.1. These algorithms have been implemented in the toolkit CLUTO[1]. For detailed explanations about these criterion functions, please refer to [130].

Although these methods can obtain clustering results at different levels of granularity, they are still based on one-side partitional algorithms like the k-means methods, which attempt to identify clusters in the whole feature space and can not find the subspace clusters.

Table 3.1: The criterion functions of the six state-of-the-art hierarchical clustering algorithms

I_1	Max $\sum_{r=1}^{k} n_r(\frac{1}{n_r^2}\sum_{d_i,d_j\in S_r}\cos(d_i,d_j)) = \sum_{r=1}^{k}\frac{\|\|D_r\|\|^2}{n_r}$
I_2	Max $\sum_{r=1}^{k}\sum_{d_i\in S_r}\cos(d_i,C_r) = \sum_{r=1}^{k}\|\|D_r\|\|^2$
ε_1	Min $\sum_{r=1}^{k} n_r\cos(C_r,C) \Leftrightarrow \sum_{r=1}^{k} n_r\frac{D_r^tD}{\|\|D_r\|\|}$
H_1	Max $\frac{I_1}{\varepsilon_1} \Leftrightarrow$ Min $\frac{\sum_{r=1}^{k}\|\|D_r\|\|^2/n_r}{\sum_{r=1}^{k} n_rD_r^tD/\|\|D_r\|\|}$
H_1	Max $\frac{I_2}{\varepsilon_1} \Leftrightarrow$ Min $\frac{\sum_{r=1}^{k}\|\|D_r\|\|}{\sum_{r=1}^{k} n_rD_r^tD/\|\|D_r\|\|}$
G_1	Min $\sum_{r=1}^{k}\frac{cut(S_1,S-S_1)}{\sum_{d_i,d_j\in S_r}\cos(d_i,d_j)} = \sum_{r=1}^{k}\frac{D_r^t(D-D_r^t)}{\|\|D_r\|\|^2}$

[1] `http://glaros.dtc.umn.edu/gkhome/views/cluto`

3.2.3 Information-Theoretic Co-clustering

We employ the famous Information-Theoretic Co-clustering (ITCC) [22] to cluster reviews, which views review-feature matrix as a joint probability distribution of two random variables and treats the co-clustering problem as an optimization problem in information theory. This method minimizes the loss of mutual information between the original joint distribution and the new joint distribution which is constructed from review clusters and feature clusters. The loss of mutual information is measured by Kullback-Leibler divergence as follows:

$$I(X;Y) - I(\hat{X};\hat{Y}) = KL(p(X,Y)||q(X,Y)) \tag{3.1}$$

Where X represents the review random variable and Y represents the feature random variable. $\hat{X}$ and $\hat{Y}$ are the review-cluster random variable and the feature-cluster random variable, respectively. The joint probability distribution between X and Y are denoted by $p(X,Y)$. $I(X,Y)$ is the mutual information between X and Y, KL is the Kullback-Leibler divergence, and $q(X,Y)$ is a new distribution constructed by the $\hat{X}$ and $\hat{Y}$ as follow:

$$q(X,Y) = p(\hat{X},\hat{Y})p(X|\hat{X})p(Y|\hat{Y}) \tag{3.2}$$

To obtain a co-clustering with K review clusters and L feature, the whole algorithm proceeds as follows [41]:

Algorithm 3.1 ITCC

Input: k, L

k review clusters and L feature clusters are randomly initialized.
repeat
1. k review centers are computed incorporating the information of feature clusters,
2. each review is reassigned to its closed center in the measure of KL divergence,
3. L feature centers are computed incorporating the information of review clusters,
4. each feature is reassigned to its closed center in the measure of KL divergence,

until the procedure converges

The ITCC method can find subspace clusters because of iteratively clustering the reviews and features into clusters, which capture the relationship of reviews and features. However, this method attempts to identify all the clusters at a stroke, which may be difficult for some data set.

3.2.4 Hierarchical Information Theoretic Co-clustering

Hierarchical Information Theoretic Co-clustering (HITCC) [119] extends the flat Information Theoretic Co-clustering (ITCC) to enable hierarchical clustering. HITCC makes use of the subspace mining capability of ITCC to cluster data level by level, so that subspace clusters at different-level abstractions can be identified. It builds the hierarchical tree structure via a series of nested ITCC method. Moreover, a feature reduction strategy and the stopping criterion assist overall clustering procedure.

HITCC performs in a binary partition scheme which partitions a co-occurrence matrix of objects and features into two sub-co-occurrence matrices only based on object clusters obtained from $(2, L)$-ITCC. For each sub-co-occurrence matrix, the $(2, L)$-ITCC is employed to get 2 new sub-co-occurrence matrices. This procedure loops until the stopping criterion is satisfied, producing a 2-ary hierarchy. The binary partition scheme is used because arbitrary number of clusters can be generated by repeating binary partitions. According to previous work on ITCC [22], the best number of feature clusters L varies from one data set to another. Therefore, L is treated as an input parameter and determined empirically. At each level, less discriminative features of each matrix are reduced by a feature reduction method proposed by Dhillon et al. [21]. Features are ranked by their quality, which is measured as follows:

$$q(y) = \sum_{i=1}^{m} f_i^2 - \frac{1}{m}[\sum_{i=1}^{m} f_i]^2 \tag{3.3}$$

where m denotes the number of total objects in the sub-co-occurrence matrix, and

f_i is the co-occurrence frequency of feature y and object x_i. Higher value of $q(y)$ means better quality of feature y.

HITCC generates a cluster tree with 2-ary hierarchy based on binary partition scheme and feature reduction. It also needs a stop criterion to control the growth of the cluster tree. For a cluster node, if the statistical dependence between objects and features is high, that means these objects and features are highly correlated to each other. Then, this cluster node is judged as a leaf and stops to partition. The normalized mutual information is used to measure the statistical dependence.

The algorithm is as follows:

Algorithm 3.2 HITCC

Input: the number of feature clusters L, the threshold for stopping criterion, the review-feature matrix.

1. the original review-feature matrix is viewed as the root node of the hierarchical tree, and it is set as current node.

repeat

2. perform the ITCC on the current node five times with the review cluster number being 2 and the given feature cluster number L, and then obtain five different co-clustering results. Only partition the reviews according to the best result and generate two sub-co-occurrence matrices as children nodes.

3. some features (say 20%) of these two matrices will be reduced by a feature reduction process proposed in [46], so two new sub-co-occurrence matrices are obtained.

4. compute the normalized mutual information for each child node, judge if the node is leaf or not according to the threshold.

until each child node is leaf

Output: a hierarchy with clustering results

HITCC has been proven to perform better than traditional hierarchical clustering and ITCC on topic-based text clustering [119]. Compared to traditional hierarchical clustering methods, the advantage of HITCC is that feature clusters generated by co-clustering at each level can reduce feature dimensionality, convey some semantic concepts or topics, and help uncover subspace clusters, which leads to good partitions of objects. Compared to ITCC, HITCC benefits from its stepwise (hierarchical) exploiting strategy, which let it exploit the clusters level by level.

3.3 Features for Clustering

As we mentioned earlier, it is very important for sentiment analysis to transform the unstructured review text into numerical structured data with more useful key words that can manifest the sentiment orientation of reviews. To represent reviews with key features, we introduce four types of features usually used by previous work:

Unigram: this method is one of the N-grams which are a type of documents representation. It refers to every single word that makes up the vocabulary of the document.

Adjectives in Part of speech: To extract words of a specified part of speech is an effective way to represent a document. After the Part-of-Speech tagging, all the words will be tagged with the POS tag. Every part of speech can be selected as the feature of the document. We only select the adjectives as the features because adjectives are important indicators of opinions [35, 47].

Adjectives and their matched nouns in Part of speech: The adjectives are usually used to describe some features of the object, which are also useful information for sentiment analysis. For example, only extracting the adjective "long" from the sentence "the battery life is long" makes it difficult to distinguish the sentiment orientation. If we extract its matched noun "life", it is understandable that "long-life" is positive. Thus, we extract the adjective and the noun they modify to represent the review. And the extraction strategy we used is that finds the nearest noun for every adjective. If no noun is found, the adjective is still reserved.

Opinion words: opinion words are applied to express whether the sentiment is positive or negative. For example, positive opinion words means they are good, beautiful, great. Negative opinion words means they are bad, terrible, disappoint. These opinion words are the dominating indicator for sentiment analysis.

3.4 Experimental Results

In this section, experiments are conducted on the reviews on attractions of tourism to compare the performance of different clustering methods with different features, aiming at examining which clustering method is suitable to do sentiment analysis on tourism reviews for inferring ratings.

3.4.1 Data Sets

We collected a set of reviews on attractions of ten famous cities in Asia. These data sets were downloaded from the TripAdvisor which is one of the most famous providers of user-generated travel reviews and ratings. The reviews on attraction are very fuzzy, free format, not well-structured and more challenging to deal with. For example:

"9|0|An awesome experience|5|Kerrymore|$Jul1, 2010$|Johannesburg, South Africa|What an amazing experience! I did the 2 day trip with the overnight stay at the Great Wall, which I would highly recommend. The section of the Wall that we visited is deserted, but spectacular and you get to enoy a delicious picnic with wine, dessert and coffee! The guest house is amazing, set in a beautiful location with friendly hosts. The next day in Beijing was fun and lunch at Mr Shi was an awesome expereince. Thanks to Douwe for providing 2 funfilled days|"

The example shows that a review includes the review ID, attraction ID, title, rating, user Name, date, and the review body. The rating is in a 5-point integer scale that can be viewed as different classes of the review indicating different degrees of satisfaction user expressed.

We developed a simple web crawler for downloading information from TripAdvisor. As shown in Figure 3.1, the crawler starts with the URL of the "things to do" page, which is called seed. In the first step, it crawls all the required

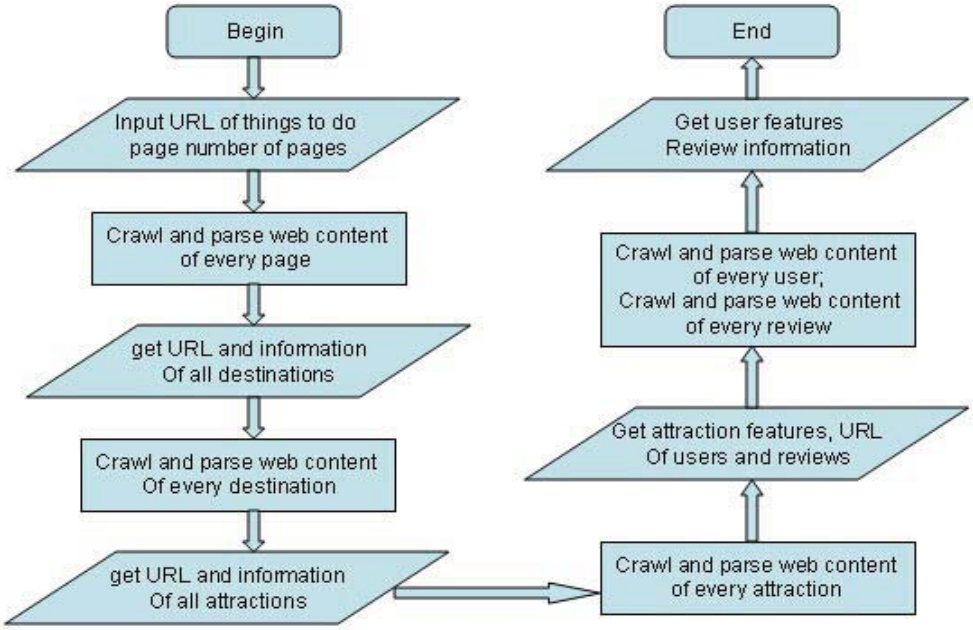


Figure 3.1: Flows of a crawler

pages. From these pages, the algorithm extracts all of the destination names, IDs and URLs with corresponding regular expression. After these destination URLs are obtained, the crawler visits every URL, to get the URL and information of attractions at the destination. Then the crawler visits the URL of attractions and downloads the web contents, from which the algorithm fetches all of the features of the attraction, including ranking, type, activities, average rating, rating distribution, etc. Simultaneously, the URLs of the reviews on the attraction and the corresponding users are extracted. In the next step, the crawler goes into every URL of reviews, to get the information of the reviews, such as title, body, rating, etc.

To prepare data sets for the experiment, review texts were changed into review-feature matrix, in which the element is the frequency of the feature in a specific review. At the same time, the stop words of the data sets were removed by Porter stemming. Then, data sets represented with the Unigrams (all the single words) can be changed into review-Unigram feature matrices.

Table 3.2: Number of reviews of each data set

Data	**Beijing**	**Shanghai**	**HongKong**	**Singapore**	**Tokyo**
No.	610	426	837	823	319
Data	**Bangkok**	**KualaLumpu**	**Dubai**	**Bali**	**Chiangmai**
No.	1348	406	619	1223	1922

In order to extract adjectives or adjectives and their matched nouns, we used the Stanford Log-linear Part-Of-Speech Tagger[2] to tag the data. Then extracted the adjectives alone or the adjectives and their matched nouns that are located near the adjectives with 3 words distance. After stemming and removing stop words and illegal words that is not existing or non-adjective, we got the review-adjective matrices and review-adjective and noun matrices. To obtain opinion words, we extracted positive and negative adjectives by making use of the dictionary of Inquirer[3]. Using these positive and negative adjectives, the data sets can be changed into review-opinion word matrices. Table 3.2 and Table 3.3 show the number of reviews and the number of features, respectively. For discarding some illegal words, we only selected 70% of top ranked features according to their mutual information. Cleaned data sets are shown in Table 3.4.

3.4.2 Implementation

We compared different clustering methods which include K-means, ITCC, hierarchical clustering and HITCC combining with different feature selection methods. All experiments were conducted on the review data sets as shown in Table 3.4, which is represented as document-word co-occurrence matrix. The results were evaluated with Fscore measure. K-means was implemented in Matlab. And the state-of-the-art hierarchical clustering algorithms were implemented in the toolkit CLUTO.

[2] `http://nlp.stanford.edu/software/tagger.shtml`

[3] `http://www.wjh.harvard.edu/~inquirer/`

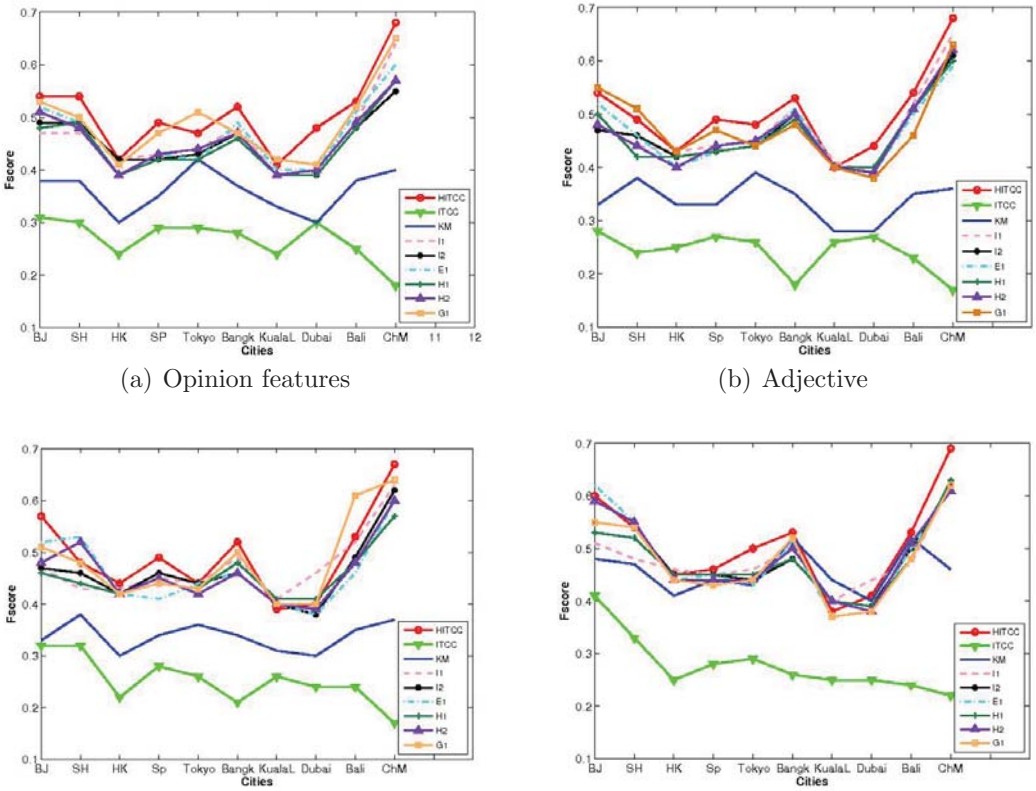
Table 3.3: Number of features after preprocessing and feature selection

Data Set	Unigram	Adjective	Adjective +Noun	Opinion words
Beijing	4501	1621	2760	274
Shanghai	3114	1100	1844	210
HongKong	4719	1757	2879	299
Singapore	4749	1632	2739	294
Tokyo	3008	992	1759	189
Bangkok	6192	2353	3811	386
KualaLumpu	3507	1140	1913	214
Dubai	4025	1404	2326	256
Bali	6026	2128	3441	340
ChiangMai	6143	2501	3924	380

Table 3.4: Number of features after removing illegal words

Data Set	Unigram	Adjective	Adjective +Noun	Opinion words
Beijing	3150	1134	1931	274
Shanghai	2179	770	1290	210
HongKong	3303	1229	2015	299
Singapore	3324	1142	1917	294
Tokyo	2105	694	1231	189
Bangkok	4334	1647	2667	386
KualaLumpu	2454	798	1339	214
Dubai	2817	982	1628	256
Bali	4218	1489	2408	340
ChiangMai	4300	1750	2746	380

For K-means and ITCC, the number of review clusters (K) is needed to input. Since the rating of reviews is in a 5-point integer scale, we input 5 as the number of the review clusters. The number of word clusters (L) should be determined empirically. To set the number of word clusters (L) for both ITCC and HITCC, we tried L with 2, 4, 8, 16, 32, and then choose the one that generated the best clustering result as the final input value. Both ITCC and HITCC algorithms initialized review clusters by a random perturbation of the "mean" review and initialized the word clusters by



(a) Opinion features

(b) Adjective

(c) Adjective and noun

(d) Unigram

Figure 3.2: Fscore values of different methods on different features.

the bootstrapping method [22]. According to previous research [119], the stopping criterion threshold of HITCC was set as 0.008. Parameters for the hierarchical algorithms were set as follows: -clmethod=RB, -clusters=10, -crun=I1, I2, E1, PH1, PH2, and G1 respectively (where I1, I2, E1, PH1, PH2, and G1 represent the six different criterion functions respectively). Considering the random components in the initializations, we evaluate all the algorithms by the average Fscore of three different trials.

3.4.3 Results and Discussions

Figure 3.2 shows the comparison results obtained by different clustering methods on different data sets with different features. Results of four types of features are

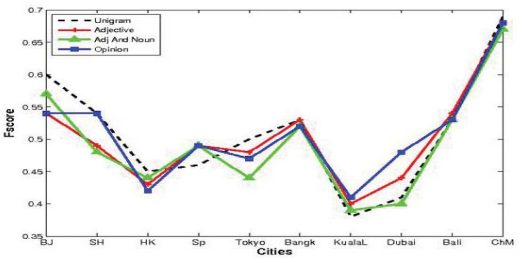
shown in four sub-figures respectively. Each sub-figure includes the results of all clustering methods using one type of features. From the figure, we can observe that the results of hierarchical clustering approaches are generally better than those of non-hierarchical clustering approaches. In particular, HITCC method can get the best Fscore for most types of features.

The reason is possible that there is no obvious distinction between adjacent classes of the tourism review data, which make it very difficult to partition the review into 5 classes with different rating. For instance, it is very difficult to clearly tell the difference between a review rated as 3 and another review rated as 4. Non-hierarchical clustering methods, such as K-means and ITCC, attempt to get all clusters at at one stroke, which is difficult to obtain good clustering results. However, hierarchical clustering approaches apply stepwise exploiting strategy to identify clusters, which can gradually discriminate every class level by level. Hence, these hierarchical clustering methods can obtain more accurate clustering results than non-hierarchical clustering methods for tourism reviews.

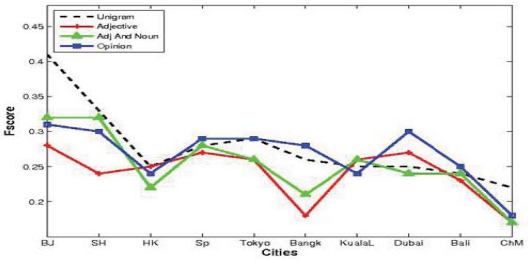
HITCC performs better than other clustering approaches, benefiting from its stepwise partition strategy, co-clustering at each level and feature reduction. Stepwise partition strategy makes HITCC to gradually discover clusters level by level. Co-clustering at each level clusters features and objects simultaneously and the feature clusters can help HITCC reduce feature dimensionality, convey some useful semantic concepts or topics, and identify underlying subspace clusters. Feature reduction removes relatively useless features to reduce dimensionality.

Figure 3.3 shows the comparison results of different features, and each sub-figure is from one clustering method. From these figures, we can observe that different features lead to similar or even better result for most clustering methods. Only using opinion words or adjectives can obtain similar result for most cities as using all the words as features. These results indicate that only some representative

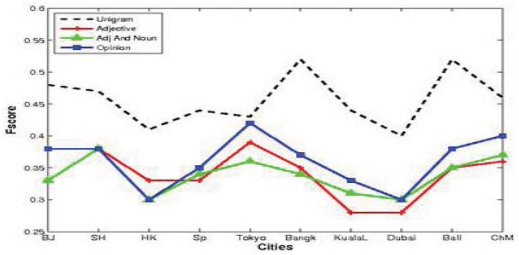
and discriminative features can represent overall reviews. This can reduce the dimensionality substantially. So it is more suitable for HITCC to use the opinion words or adjectives as features to generate ratings from tourism reviews.



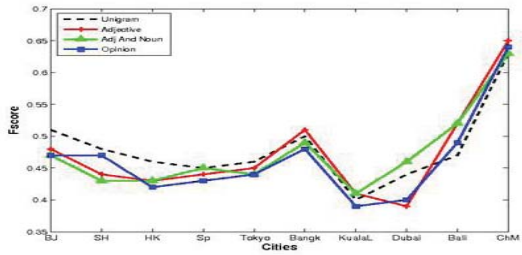
(a) HITCC



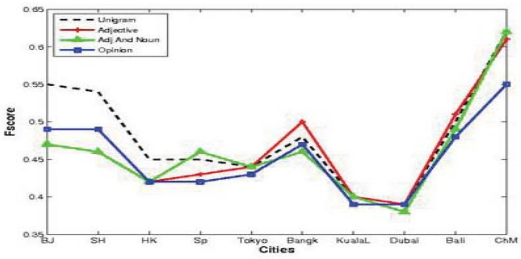
(b) ITCC



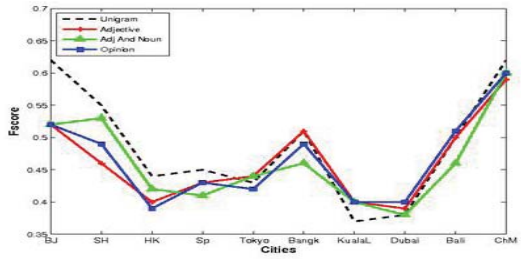
(c) KMeans



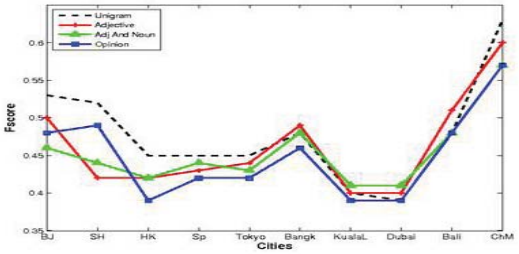
(d) I1



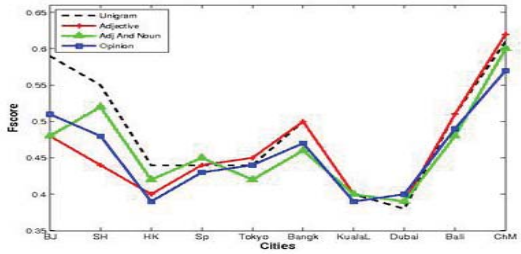
(e) I2



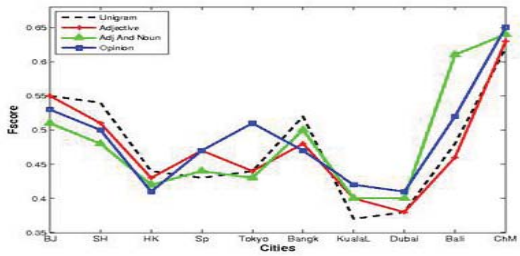
(f) E1



(g) H1



(h) H2



(i) G1

Figure 3.3: Comparison of the result of different features.

3.5 Summary

In this chapter, we compared several clustering methods for sentiment analysis on tourism reviews, to infer numerical ratings from reviews. Moreover, different features were also investigated to choose more suitable one for clustering methods.

Experimental results have shown that hierarchical clustering approaches, including HITCC and six traditional hierarchical methods performs better than non-hierarchical clustering approaches, including K-means and ITCC. The reason is possible that hierarchical clustering methods apply stepwise exploiting partition strategy, but non-hierarchical clustering approaches get clusters at one stroke. Particularly, HITCC outperforms other clustering methods, benefiting from the stepwise partition strategy, co-clustering at each level and feature reduction. The results based on different features manifested that only using Part-of-speech or opinion words as features are better than using all the words as features, because Part-of-speech or opinion data can obtain similar or even better result than those using all the words and at the same time reduce overall feature dimensionality.

Although our rating inference method is helpful to augment ratings for rating-based recommender systems, only using unsupervised approach to do sentiment analysis on tourism review data for inferring rating is very difficult. Moreover, rating inference should be based on textual reviews. However, if travellers are new users who have no ratings or reviews, we can not do sentiment analysis to infer ratings for recommender systems. Then, how to make recommendation?

Chapter 4

Applicability of Demographic Recommender System to Tourist Attractions

4.1 Introduction

Nowadays, increasing number of tourists search information on the internet before a travel to help themselves make travel plans. However, with the explosive growth of internet, tourists are usually overwhelmed by the existence of a large quantity of travel information. Recommender systems can assist tourists in managing the numerous information, and facilitate them to make travel decisions [94].

A recommender system provides a user with suggestions about an item [92]. There has been much work done on recommender algorithms for electronic commerce [105, 40]. Typical examples include recommendation of CDs, books,and other products at Amazon.com [59], and movies, DVDs, and VHS videos at Movielens [73]. Various approaches for performing recommendation have been developed, making use of either rating, content, demographic or other knowledge. Most of them are [12]: collaborative filtering, content-based, knowledge-based and demographic recommender systems. Collaborative filtering systems make user recommendations of items based on the user's previous preferences and the those of other users who are

similar to the user [104], as measured by the ratings. Content-based systems provide a user with recommendations on items which have similar features to the ones this user has liked before [1]. Knowledge-based systems make recommendations based on the knowledge about users' requirements on items [12]. Demographic recommender systems generate recommendations based on demographic features [87]. Compared to the other approaches, the advantage of this type of system is that it only uses the demographic data of user such as gender, age, education, etc, and may not need the history of user rating, the textual description or the knowledge of item. Therefore, new users can get recommendations before they rate any items.

In the tourism domain, many recommender systems have been developed. Most of them are content-based and knowledge-based [94, 64, 96], which need sufficient descriptions about items, historical ratings or extra knowledge. In some tourism communities, these information may be difficult to obtain. Moreover, most existing work confined to making recommendations at the destination level. For instance, VacationCoach's Me-Print and Triplehop's TripMatcher [94] use content-based method to make recommendations on destinations, which need description content about destinations. The knowledge-based systems proposed in [64, 96] deduce the preferences and requirements of the tourist based on the knowledge obtained from conversional dialog. Tourism recommender systems are still in the stage of development, without achieving the level of success as in the domains of books or movies. Hence, there remains much room for improvement.

In this chapter, we investigate the applicability of demographic recommender algorithms for the prediction of ratings of tourist attractions. We examine the integration of different machine learning methods with the system, aiming at determining whether these approaches and demographic information alone are useful and effective to make predictions of the rating.

The rest of this chapter is structured as follows. Chapter 4.2 presents the

demographic systems used on attractions in TripAdvisor. Chapter 4.3 describes the experimental work. Chapter 4.4 includes some concluding remarks.

4.2 Methodology

The objective of this study is to investigate how the demographic information of the tourists impact on the prediction of attractions' ratings. Focusing on the attractions on TripAdvisor, we present different demographic recommender systems by using different machine learning approaches to do the prediction.

4.2.1 Machine Learning Approaches

The machine leaning approaches used in this chapter include Naive Bayes, Bayesian network and Support Vector Machines, which have been proved to be effective classifiers.

Naive Bayes

Naive Bayes classifier is the most popular probabilistic classifier based on Bayes' theorem. It uses a training data set to generate a probabilistic model, assuming that the attributes of the cases are independent. Despite being very simple, it performs well in many complex situations. Let d be a case, and c be a class. According to the Bayes' theorem, the probability that a case d belongs to the class c can be calculated as follows:

$$P(c|d) = \frac{P(c)P(d|c)}{P(d)} \tag{4.1}$$

In equation (4.1), $P(c)$ is the prior probability of class c, $P(d|c)$ denotes the conditional probability of case d given c, and $P(d)$ is the probability of d. To classify case d, the class with the highest probability is assigned:

$$c_{predict} = \underset{c_i}{argmax}\, \frac{P(c_i)P(d|c_i)}{P(d)} \tag{4.2}$$

Bayesian Network

Bayesian network is a probabilistic graphical model which represents the relationship among variables. The nodes in the graph represent variables and the directed arcs represent probability correlations. The start of the arc is the parent node, and the end is the child node. The child node is dependent on its parent node, but conditionally independent of other nodes. Bayesian network is a powerful knowledge representation and reasoning approach based on Bayes' theorem.

Support Vector Machine

Support Vector Machine (SVM) approach is a state-of-the-art classification technique. It is also considered one of the most accurate and robust approaches among all well-known methods [122]. The principle of SVM is that it constructs separating hyperplane in the data space and aims at making a good separation realized by the hyperplane that has the largest margin to the nearest data point of any class.

4.2.2 Demographic Recommender System

The demographic recommender system is built on the data from TripAdvisor, the largest travel community online. Its primary function is to collect and disseminate tourist generated content, including reviews, ratings, photos and videos. To build a demographic recommender system, we obtain the demographic information of tourists and the ratings that the tourists have given to the attractions. There are several types of attractions on Tripadvisor, such as "Theater, Art and Performance", "Parks, Gardens and Nature", and "Museums, Zoos and Aquariums". As a case study, this chapter focuses on the tourists who have rated the set of attractions belonging to "Museums, Zoos and Aquariums", with the aim to predict how a new tourist will rate on this type of attractions at a destination. Demographic information of a tourist is extracted from the tourist's home page, represented as a vector and

then used to do classification with machine learning methods.

Machine learning algorithms introduced above are conducted to obtain demographic categorizations with ratings as classes. The system assumes that tourists in the same category have the same tastes or preferences. So the tourists who have rated the same type of attractions at a destination are used to train the classifier, and those who are classified into the same class are assumed to have the same attitude (rating) to the attractions. Then the trained classifier issues recommendation for a new tourist by first identifying the class that this tourist belongs to and then determining the rating according to the attitude of other tourists in the same class. In addition, a baseline method that just classifies new tourist into the class with the largest number of tourists is used for comparison with the machine learning methods.

4.3 Experimental Results

This section provides data description, experimental implementation and results. The experiments were conducted on six data sets which were collected from TripAdvisor. We compared different demographic recommender systems utilizing three different machine learning approaches. According to the analysis of the results, we discuss the impact of demographic information of tourist on the prediction of ratings.

4.3.1 Data Sets

For the experiment, we developed a web crawler to collect data from Tripadvisor. Focusing on six cities: New York City, Paris, London, Rome, Chicago and Berlin, it started on June 1, 2012 and downloaded the ratings and demographic information of tourists who have rated the attractions belonging to the type of "Museums, Zoo, and Aquariums". The ratings {1,2,3,4,5} were used as multi-class labels. Since a

Table 4.1: Structure of the tourist demographic vector

Features	**Feature values**
Age	$\{age \leqslant 12\}$; $\{13 \leqslant age \leqslant 17\}$; $\{18 \leqslant age \leqslant 24\}$; $\{25 \leqslant age \leqslant 34\}$; $\{35 \leqslant age \leqslant 49\}$; $\{50 \leqslant age \leqslant 64\}$; $\{65 \leqslant age\}$
Gender	{male}; {female}
Travel stye	{roughing it}; {on a tight budget}; {middle of the road}; {splurge occasionally}; {nothing but the best}
Travel for	{work}; {fun}
Great vacation	{museums/cultural/historical sites}; {theme/amusement parks}; {outdoor/sports} {great food/shopping}; {concerts/music festivals}
Travel with	{myself}; {spouse/significant other}; {family/pets}; {friend/collegues} {large group or tour}

Table 4.2: The descriptions of data sets

Data Set	**Destination**	**No. of Tourist**
1	New York City	6360
2	Paris	6890
3	London	11800
4	Chicago	2640
5	Rome	2100
6	Berlin	3250

tourist may rate multiple attractions, only the newest rating was used as the input to the machine learning methods.

To prepare the data sets for the experiment, demographic information of tourist needs be represented as a vector. The information includes “Age”, “Gender”, “Travel style”, “Travel for”, “Great vacation” and “Travel with”. And each of them has more than one possible values. In order to transform the demographic information to a vector with numeric value, the binary encoding approach was used to map the categorical information to higher dimensional features with the value as 0 or

1. The tourist demographic information was encoded as a vector with 26 features, as explained in Table 4.1. Meanwhile, we filtered out the tourists who have no demographic information in their home page. The final data is shown in Table 4.2, summarizing the number of tourists at the level of the set of attractions belonging to "Museums, Zoo, and Aquariums" in a city. And for the tourist who rated more than one attractions, only the newest rating was included in the data set.

4.3.2 Experimental Implementation

The SVM method was implemented by the software $LIBSVM$ [13]. For our multi-classes data, this software can realize multi-class classification with "one-against-one" approach, which decouples the multi-class problem to binary class problem, and applies a voting strategy to determine the class of a data point. To implement the Bayesian network and Naive Bayes methods, a tool called Netica[1] was used. Given the network structure, this tool can learn conditional probability table of the Bayesian network automatically based on the training data. In this chapter, we built our Bayesian network structure according to the study in [39], which gives some relationships between the demographic information and estimates the tourist's preferred activities by Bayesian network. Figure 4.1 shows our Bayesian network structure. The structure of the naive Bayes method is just a special model of the Bayesian network, which connects all the demographic nodes to the rating node. To evaluate our system, we applied the 10-fold cross-validation approach to estimate how accurately it performs in practice. The evaluation metric we selected is Mean Square Error (MSE), which is defined as follows:

$$MSE = \frac{\sum_{i=1}^{n} (\hat{r}_i - r_i)^2}{n} \tag{4.3}$$

[1] `http://www.norsys.com/`

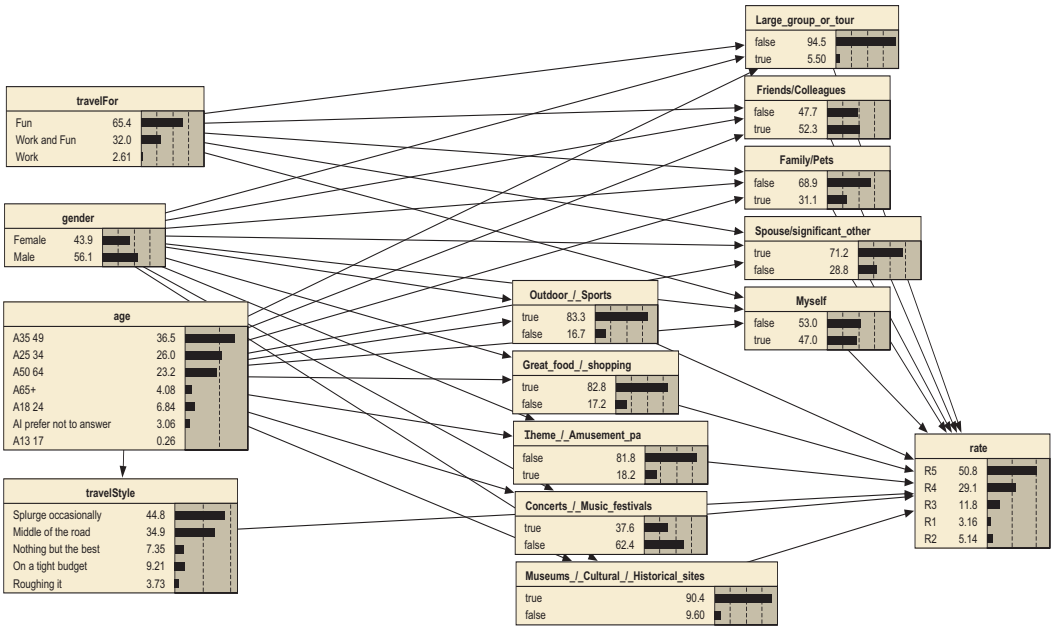


Figure 4.1: Bayesian network structure

where $\hat{r}_i$ is the predicted rating value by the ith tourist, and r denotes the actual rating value. The lower the MSE, the more accurate the recommendation result is.

4.3.3 Results and Discussion

In this subsection, we present the experimental results of applying different machine learning methods in the demographic recommender system to generate predictions.

Table 4.3 shows the average MSE of 10-fold cross-validation on six data sets. We can observe that all of the machine learning approaches perform better than the baseline method. Meanwhile, it can be seen from Table 4.4 that all of the machine learning methods are more stable than the baseline method. In general, SVM performs a little better than naive Bayes method, and both of them outperform the Bayesian network method. Therefore, the best results using demographic data

Table 4.3: Average MSE of the 10-fold cross-validation results

Cities	**Baseline**	**Naive Bayes**	**Bayesian Network**	**SVM**
New York	1.420	1.202	1.352	1.176
Paris	0.913	0.759	0.886	0.781
London	1.800	1.474	1.651	1.467
Chicago	1.433	1.100	1.324	1.089
Rome	1.100	1.077	1.121	0.803
Berlin	1.688	1.493	1.256	1.204

Table 4.4: Variance of MSE of the 10-fold cross-validation results

Cities	**Baseline**	**Naive Bayes**	**Bayesian Network**	**SVM**
New York	0.192	0.146	0.169	0.150
Paris	0.156	0.125	0.146	0.113
London	0.122	0.106	0.109	0.097
Chicago	0.173	0.191	0.163	0.159
Rome	0.282	0.248	0.236	0.211
Berlin	0.310	0.258	0.183	0.233

Table 4.5: Percentage of the tourists classified correctly by SVM

Cities	$Rating_1(\%)$	$Rating_2(\%)$	$Rating_3(\%)$	$Rating_4(\%)$	$Rating_5(\%)$
New York	5.37	15.99	24.61	24.05	92.91
Paris	18.31	16.97	14.26	21.36	96.40
London	7.62	21.16	16.44	28.28	92.74
Chicago	1.00	18.02	26.63	37.00	89.46
Rome	12.22	11.17	25.43	21.73	95.45
Berlin	25.57	12.39	39.87	35.41	89.56

are those obtained by SVM method. Details of the 10-fold cross-validation results can be seen in the Figure 4.2.

Although the demographic recommender systems using machine learning methods are indeed useful to make the prediction of ratings, the results are not so accurate. Table 4.5 shows the percentage of the tourists classified correctly by SVM. We can

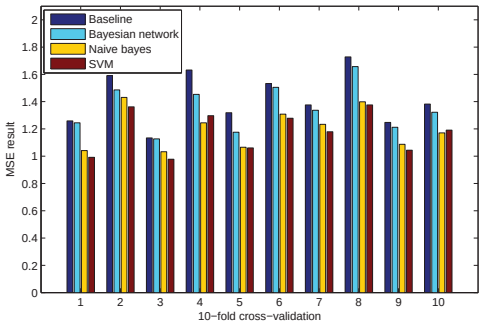
Table 4.6: Number of tourists in different rating classes

Cities	$Rating_1$	$Rating_2$	$Rating_3$	$Rating_4$	$Rating_5$
New York	110	297	742	1632	3579
Paris	55	174	531	1721	4409
London	399	640	1443	3270	6048
Chicago	37	123	334	749	1397
Rome	35	57	180	518	1310
Berlin	105	139	410	916	1680

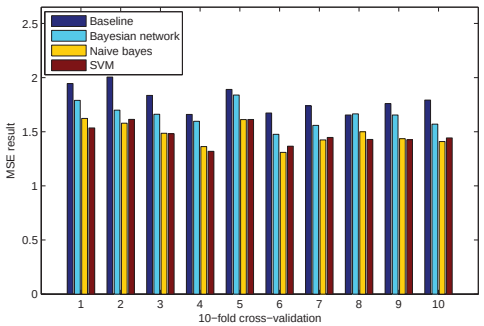
see that SVM identified around 90% or even more tourists that belong to the class rated as 5. However, for other classes, this method only identified 1% - 39.87% tourists. There are two possible reasons why the demographic recommender systems using machine learning methods can not obtain very accurate results. Firstly, the rating data sets on attractions of TripAdvisor are very unbalanced. Table 4.6 shows the distribution of tourists ratings. Most of the tourists give ratings of 5 or 4, and very few tourists give ratings of 1 and 2. For each data set, more than 50% tourists give ratings of 5, but no more than 5% tourists give ratings of 1. For the machine learning methods, because of the imbalance of the data, it is difficult to identify the tourist vectors which belong to the smaller classes, such as the class rated as 1 and 2. All of them have a tendency to classify the tourist into the largest class rated of 5. Furthermore, as shown in Figure 4.2, the results for 10-fold cross validation are not very stable because different testing sets include different number of tourists belonging to the class rated as 5. Secondly, the demographic features are not so representative and discriminative to distinguish five classes, especially the smaller classes. They alone are not sufficient to do accurate prediction of ratings.

Additionally, the fact that the Bayesian network approach obtained worse result than the naive Bayes approach indicates that the structure of our Bayesian network was unable to capture the potential relationships between different features and ratings. It's also possible that the correlation between the demographic features is

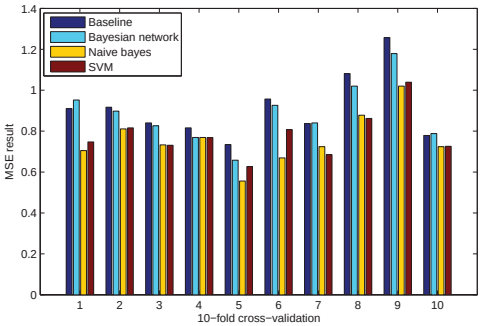
weak.



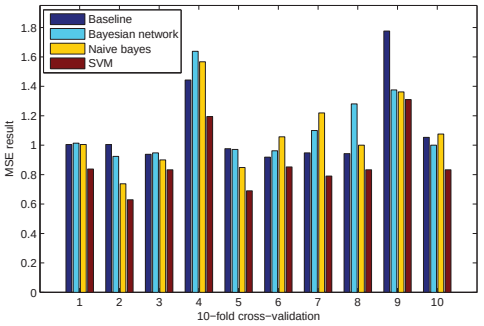
(a) New York



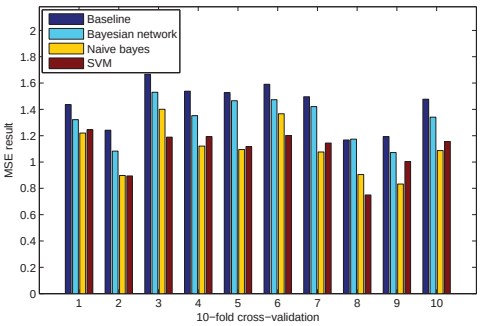
(b) London



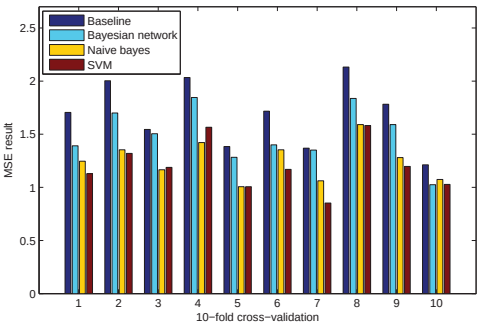
(c) Paris



(d) Rome



(e) Chicago



(f) Berlin

Figure 4.2: Details of 10-fold cross-validation results

4.4 Summary

In this chapter, we applied demographic recommender system by integrating it with machine learning methods to make recommendation of attractions on TripAdvisor. Three different machine learning methods were investigated to determine whether this approach and demographic information of tourists are suitable to predict the ratings that tourists give to the attractions. Experimental results have shown that three machine learning methods based on demographic information performed better than the baseline method, especially the SVM method. These preliminary results also suggest that demographic information alone is not sufficient to do accurate prediction of ratings, though more detailed experiment is needed to confirm our results, such as handling the unbalanced data, considering other type of attractions.

Although demographic recommender system is helpful to alleviate the cold-start issue, it is unable to handle the uninformative, bias or false information, which usually lead to unreliable recommendation. Travellers will make bad travel plans based on unreliable recommendation. Therefore, it is very important to improve tourism recommender systems by addressing information credibility issue.

Chapter 5

Quantifying Reviewer Credibility in Online Tourism

5.1 Introduction

More and more travellers like to post reviews online to share experience and opinions, which has become one important source of information [111, 51, 60]. However, the explosive growth of reviews and the presence of uninformative, biased or even false information make it very time consuming and challenging for the travellers to find credible reviews [53, 70].

Some researchers have investigated into several cues that influence the perception of the credibility of reviews in tourism [51, 34, 124], and their findings provide travelers with some guidelines to judge credible reviews. However, these work rely only on the survey method to explore the cues but did not develop a method to search for credible reviews automatically. Credibility assessment online has been investigated from three perspectives: message credibility, source credibility and medium credibility [71, 27, 98]. In the tourism domain, the review can be considered a message, the reviewer a source and the tourism website a medium, as shown in Figure 5.1. Some literatures [31, 98] have pointed out that credibility assessment of sources and messages are fundamentally interlinked. In the light of this insight,

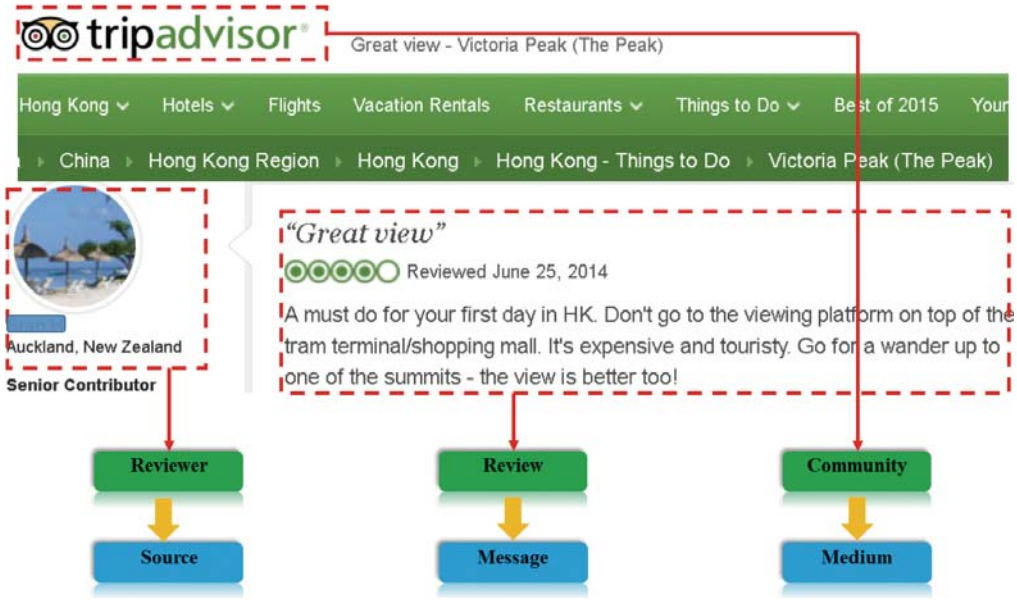


Figure 5.1: Linkage of the components between tourism domain and information assessment

assessing the credibility of reviewers can assist travellers in finding credible reviews. Hence, this chapter focuses on measuring the credibility of the reviewer to help travellers search for credible reviews.

There has been some work on measuring the credibility of reviewers in the tourism domain. Some researchers have applied the survey method to explore the cues affecting the evaluation of the expertise and trustworthiness of the reviewer in online tourism [108, 34, 118, 128], which are also two key dimensions of source credibility [37, 27]. However, these studies fail to make a quantitative evaluation of the reviewer credibility. Besides, Lee et al. [53] have used the average helpful vote (Average RHR), which is the total number of helpful votes received received by a reviewer divided by the total number of reviews posted by the reviewer, to represent the reviewer credibility in TripAdvisor. Although this approach can evaluate reviewer credibility quantitatively, it tends to favor the reviewers who have posted very few reviews possibly implying a narrow range of expertise but nevertheless obtained high

Average RHR which imply high trustworthiness.

In this chapter, we present a method that quantitatively measure the credibility of reviewers in TripAdvisor, which is the most popular travel community in the world. An Impact Index is proposed to compute the reviewer credibility by evaluating expertise and trustworthiness jointly, based on the number of reviews posted by the reviewer and the number of helpful votes received by the review. Reviewers who have a high Impact Index are those who have posted more reviews, implying their expertise, and each of the review having obtained more helpful votes, implying their trustworthiness. Compared to the previous average helpful vote (Average RHR), the Impact Index considers expertise and trustworthiness simultaneously, and does not emphasize one dimension only. To better represent the multi-faceted nature of credibility, the Impact Index is further improved into the exposure-Impact Index by considering in addition the number of destinations on which a reviewer has posted reviews. Then, we examine the effectiveness of the Impact Index and the exposure-Impact Index by comparing them to previous measure of Average RHR.

The rest of this chapter is organized as follows. Chapter 5.2 presents the Impact Index and the exposure-Impact Index. In Chapter 5.3, comparison experiments are presented to demonstrate the effectiveness of the Impact Index and the exposure-Impact Index measurement. Chapter 5.4 includes our conclusions.

5.2 Quantifying the Credibility of Reviewers

In this section, we present a method and its improved approach to compute the credibility of the reviewer.

5.2.1 Reviewer Credibility

This chapter focuses on measuring reviewer credibility by considering two key dimensions: expertise and trustworthiness. In previous literatures [37, 29, 16], the

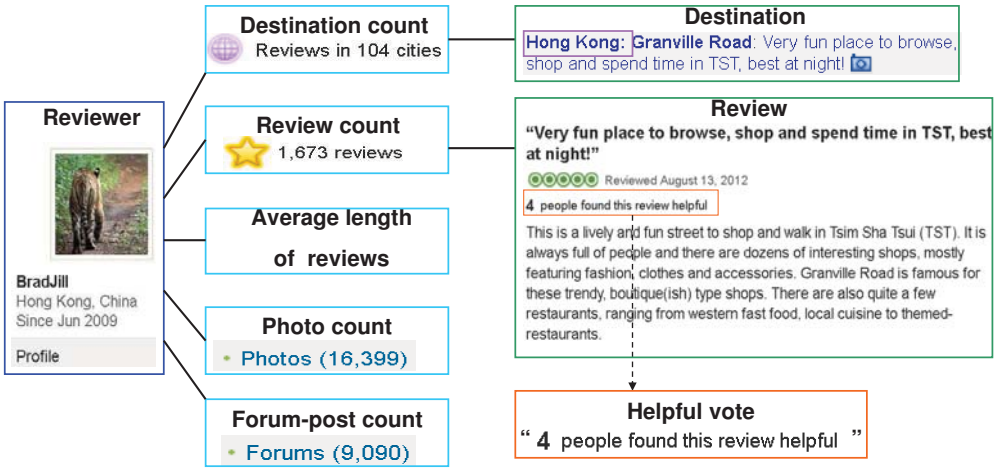


Figure 5.2: A reviewer in TripAdvisor and his contribution factors and helpful votes

expertise of source refers to source's knowledge, ability or skill to know the truth and provide valid information, and it is usually described by the terms "experienced", "knowledgeable", and "competent". The trustworthiness of the source refers to the source's willingness, moral indication or motivation to tell the truth, and it is commonly described by the terms "truthful", "well-intentioned", and "unbiased".

Table 5.1: Contribution factors of reviewers in TripAdvisor

Contribution Factors	Descriptions
Destination Count	Number of destinations that the reviewer has visited
Review Count	Number of reviews that the reviewer has posted
Average Length of Reviews (ALR)	Total number of words of all reviews posted by a reviewer divided by Review Count
Photo Count	Number of photos that the reviewer has uploaded
Forum-post Count	Number of posts that the reviewer has posted in forum

Based on previous studies [34, 108, 118, 128], the experience of reviewers in tourism can be used to represent expertise because expertise increases as experience increases. The experience can be extracted from reviewers' contribution history [108]. TripAdvisor records reviewers' contribution factors, as shown in Figure 5.2.

The main factors include destination count, review count, average length of reviews, photo count and forum-post count, which are described in Table 5.1. In TripAdvisor, reviews posted by reviewers also received feedbacks from other travelers. Helpful vote is the feedback from a traveller who considers the review helpful and reliable. Previous works have pointed out that the number of the helpful vote can signal the quality of online review, and serve as reputation proxy of the reviewer [53]. Therefore, in this chapter, the contribution factors and the number of helpful vote are used as the indicators to evaluate the expertise and trustworthiness of reviewer credibility in TripAdvisor.

5.2.2 Impact Index

This subsection proposes an approach to compute reviewer credibility by evaluating expertise and trustworthiness simultaneously based on reviewer's contribution factors and helpful votes.

We are inspired by the idea of H-Index [36] to measure expertise and trustworthiness simultaneously. H-Index is one of the most popular indicators to assess the scientific output of scientists or researchers. Compared to other single-number indicators, such as total number of citations, total number of papers, and number of highly cited papers, H-Index performs better because it simultaneously measures quantity (productivity) and quality (impact) of the published work of a scientist or researcher [36, 19]. Moreover, H-Index is robust because it is not sensitive to the long tails of the citation-rank distribution and the outstandingly highly cited papers. H-Index has more than 30 variants, such as G-Index, M-Index, S-Index and so on [19, 25, 24, 10]. These variants are classified into five classes by considering in addition different factors: field dependence, self-citation, multi-authorship, career length, and weights of highly cited papers. We can not be inspired by the idea of the first four classes because we only consider developing a measurement of reviewer

credibility based on contribution factors and helpful votes in the tourism domain and do not need to deal with those factors. The fifth class of variants improve H-Index by considering the weights of highly cited papers. Among them, G-Index is one of the most interesting improvements belonging to the fifth class[25, 24]. Compared to H-Index, the advantage of G-Index is that it is more sensitive to evaluate selective scientists who show intermediate-low quantity but high impact. However, it suffers from occasional "big hit" issue that one or few papers with a large number of citations can not represent the average research performance of a scientist.

In this work, we aim at searching for more credible reviewers with higher level of both expertise and trustworthiness, which means that these reviewers are more experienced and trustful. Therefore, we need to find the top ranked reviewers who show high quantity and quality (impact). Moreover, our measurement should not suffer from occasional "big hit" issue. Hence, the idea of G-Index and other indices similar to it is not suitable to our work and only the idea of H-Index can enlighten our work. Inspired by the idea of H-Index [36], we develop a method to measure both the expertise and trustworthiness of reviewer. The H-Index is computed based on a set of the most cited papers and the number of citations these papers have received. Among different contribution factors of reviewers in TripAdvisor, the number of reviews can directly reflect the quantity of reviewer's contribution. Moreover, the number of helpful vote can represent the impact of reviews. Hence, we propose an Impact Index to measure the credibility of reviewers based on the number of reviews and the number of helpful votes. It is defined as:

Definition 5.1. *Assume a reviewer has posted n* ***reviews****, each of which has received* h_i *helpful votes (*$i = 1, 2...n$ *), where* $h_i \geqslant h_{i+1}$*. The Impact Index of the reviewer is said to be L, if* $L \leqslant n$*, and* $h_i \geqslant L$ *(for* $1 \leqslant i \leqslant L$*),* $h_j \leqslant L$ *(for* $L < j \leqslant n$*).*

Reviewers with higher Impact Index should satisfy two conditions: first, the

reviewers have posted more reviews, which show that they have experienced more things about their travels, such as different attractions, restaurants or hotels and. That means, they have obtained more experience and knowledge, which make them relatively more competent than those who have posted few reviews. Therefore, more reviews posted by the reviewer can be considered as an indicator of the high level of expertise of the reviewer; Second, there are sufficient reviews obtained more helpful votes. It implies that more travellers believe the reviews are helpful and reliable, and the reviewers who have posted these reviews tend to be well-intentioned, truthful, and unbiased. So, more helpful votes received by the reviews posted by reviewers can indicate high level of trustworthiness. The algorithm for computing the Impact Index of a reviewer is as follows:

Algorithm 5.1 The algorithm for computing Impact Index

Initialize: the Impact Index: L=0; The number of reviews: N.
Input: a list of reviews and the number of helpful votes;

1. Rank the reviews based on their number of helpful votes, from the most, to the least, and get their ranked orders, from 1 to N. The number of the helpful votes of the i'th review is denoted as H(i), and H(i)>H(i+1) from $i = 1$ to $(N - 1)$;

for $i = 1$ **to** N **do**
  if $H(i) >= i$ **then**
    $L = i$;
  else
    break;
  end if
end for
Output: the Impact Index L.

For instance, as shown in Figure 5.3, a reviewer has posted 7 reviews, which are ranked according to their number of helpful votes, from the most to the least, with the ranked indexes as {1,2,...,7}. For each of the review, from the 1'st to the 4'th, the number of helpful votes is larger than its ranked index. But for the 5'th review, its number of the helpful votes is smaller than the ranked index. Therefore, the Impact Index L of the reviewer is 4. A reviewer cannot have a high Impact Index without posting a substantial number of reviews. Meanwhile, these reviews need to receive

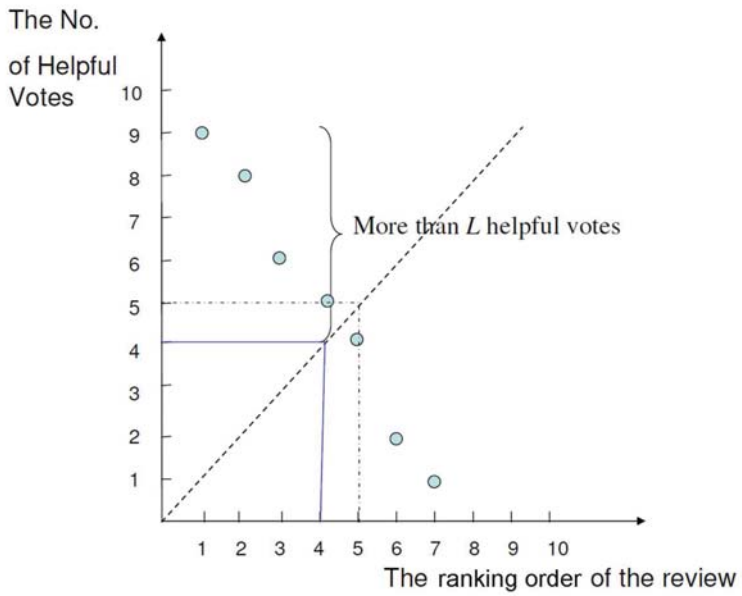


Figure 5.3: Geometrical representation of Impact Index of a reviewer

more helpful votes from travellers in order to count for the Impact Index.

Although the Impact Index measures the credibility of reviewer by considering both expertise and trustworthiness of the reviewer, only using the number of the reviews is not enough to represent the multi-faceted nature of credibility. For instance, a reviewer may have posted a lot of reviews on only one destination, which indicate that this reviewer is knowledgable about only one destination, and her or his knowledge and experience is limited rather than diverse and broad. Therefore, we need to consider more contribution factors as indicators to evaluate the reviewer credibility.

5.2.3 Exposure-Impact Index

We further consider in addition the number of destinations on which a reviewer has posted reviews as another indicator of the credibility of reviewers. Posting reviews on many destinations is an indication that the reviewer tends to have high exposure and has obtained rich experiences and broad knowledge at diverse destinations,

which enable the reviewer to provide relatively more comprehensive and reliable information. Therefore, the Impact Index is improved into Exposure-Impact Index by taking advantage of the number of destinations on which the reviewer posted reviews and the number of helpful votes received by the reviews on each destination to evaluate the expertise and trustworthiness of the reviewer credibility simultaneously. It is defined as:

Definition 5.2. *Assume a reviewer has posted reviews on n* ***destinations****, and the reviews on each destination have received h_i helpful votes ($i = 1, 2...n$), where $h_i \geqslant h_{i+1}$. The Exposure-Impact Index of the reviewer is said to be E, if $E \leqslant n$, and $h_i \geqslant E$ (for $1 \leqslant i \leqslant E$), $h_j \leqslant E$ (for $E < j \leqslant n$).*

If the Exposure-Impact Index of reviewers is higher, two conditions should be satisfied: on one hand, they have posted many reviews on more destinations and therefore have higher exposure, which indicates their richness of experience and the breadth of knowledge, and further implies the wider range of expertise; on the other hand, the reviews on each destination have received more helpful votes, which manifests that reviewers have posted more helpful and reliable reviews, indicating their higher level of trustworthiness. Hence, the number of destinations is a direct element to evaluate expertise, and the number of reviews is considered indirectly. The algorithm to compute the Exposure-Impact Index is presented as Algorithm 5.2:

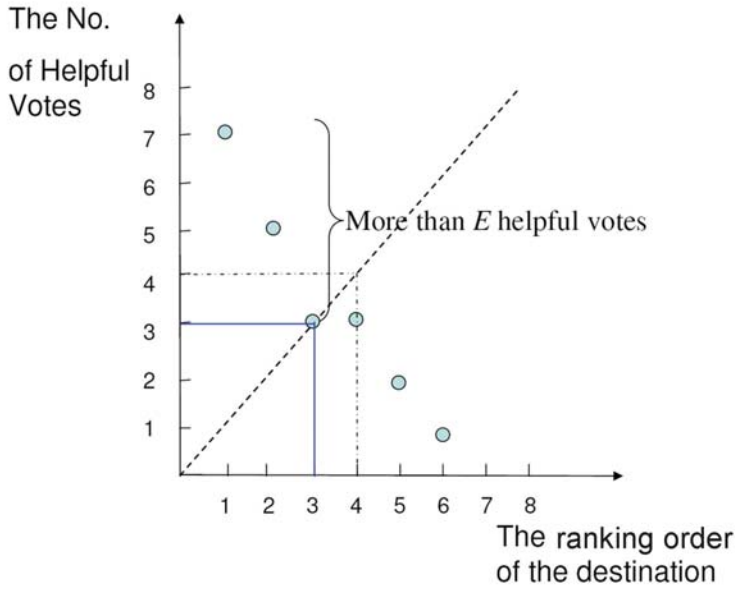


Figure 5.4: Geometrical representation of Exposure-Impact Index of a reviewer

Algorithm 5.2 The algorithm for computing Exposure-Impact Index

Initialize: The Exposure-Impact Index: E=0; The number of the destinations the reviewer has visited: N;
Input: a list of destinations, corresponding reviews and the number of helpful votes;

1. For each of the destination, count the total number of the helpful votes received by all the reviews on it;
2. Rank the destinations based on their total number of helpful votes, from the most, to the least, and get their ranked orders, from 1 to N. And the total number of the helpful votes of the i'th destination is $H(i)$, and $H(i) > H(i+1)$, from $i = 1$ to $(N-1)$;

for $i = 1$ **to** N **do**
  if $H(i) >= i$ **then**
    $E = i$;
  else
    break;
  end if
end for
Output: Exposure-Impact Index E.

Figure 5.4 shows an example to get the Exposure-Impact Index of a reviewer, who has been to 6 destinations. The destinations are ranked based on their total number of helpful votes, which is the sum of the number of helpful votes received by all the reviews on each destination. The ranked index of the 6 destinations are {1,2,...,6}.

For each destination, from the 1'st to the 3'rd, the number of the helpful votes is larger than or equal its ranked index. However, for the 4'th destination,its number of the helpful votes is smaller than the ranked index. So the Exposure-Impact Index E of the reviewer is 3. Reviewers cannot have high Exposure-Impact Index if they have not posted some reviews on a substantial number of destinations. Meanwhile, the reviews on each of destination need to receive more helpful votes in order to count for the Exposure-Impact Index.

5.3 Evaluation

The experiments were carried out on three data sets collected from TripAdvisor to compare our impact index and exposure-impact index to previous Average RHR. Firstly, we conducted a survey that let some human raters assess the credibility of the reviews posted by the reviewers with high impact index, exposure-impact index and Average RHR, in order to investigate the effectiveness of these methods on helping sear for credible reviews. Secondly, The linear regression analysis was applied to examine the relationship between the contribution factors and the credibility of the reviewer qualified by different methods.

5.3.1 Data Collection

Table 5.2: The descriptions of data sets

Data Sets	No. of reviewers	Descriptions
D-HongKong	4205	The reviewers who have posted review on Hong Kong
D-NewYork	21879	The reviewers who have posted review on New York
D-London	21375	The reviewers who have posted review on London

We developed a web crawler to collect data from TripAdvisor. The crawler started running on Aug 10, 2012, and downloaded the information of three groups of reviewers who have posted reviews on three destinations: Hong Kong, New York, and

London. These destinations are the famous tourist destinations located in different continents. Also, there are sufficient reviews in English, making data collection easier. General description of the data sets is shown as Table 5.2. The overlap of the reviewer between any two groups was less than 5%. And the number of the reviewers who have posted reviews on all of these three destinations was only 20. Hence, we obtained three different data sets to evaluate the methods. And each data set includes the contribution factors and helpful votes of reviewers.

5.3.2 Design and Implementation

Human Evaluation

This chapter focuses on quantifying reviewer credibility to help travellers find credible reviews. Therefore, to assess the effectiveness our methods, some human raters were invited to evaluate the credibility of reviews posted by the reviewers with high value of impact index, exposure-impact index and Average RHR. Previous work [71, 27, 98, 30, 97, 69] have investigated several criteria in evaluating the credibility of messages or reviews, we summarized them into three dimensions, including organization, information and reliability. The organization dimension was used to judge if the review is written in well-organized structure, clear topic and fluent language that make it easy to read and understand [71, 27, 30]. The information dimension was used to judge if the review contains diverse, detailed and sufficient relevant information about what it reviews on [71, 27, 97, 69]. And the reliability dimension was used to judge if the review is telling truth, and expressing unbiased opinion based on reviewer's personal experience [71, 97, 69]. The 5-point scale was used to evaluate each dimension, which is described in detail as Table 5.3, 5.4 and 5.5.

We recruited 15 human raters who are graduate students, and divided them into three groups equally. For each group of the data set, one group of human raters were

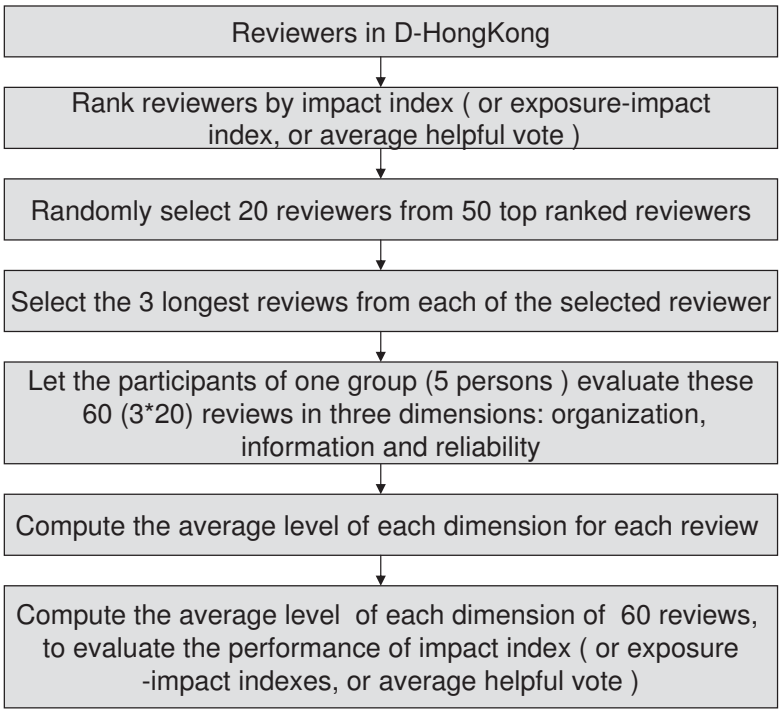


Figure 5.5: Implementation flow of human evaluation

invited to check the credibility of the reviews posted by the reviewers with high value of one measurement. Figure 5.5 shows the human evaluation procedures, using the data set D-HongKong as an example.

Linear Regression Analysis

The experiment applied the linear regression analysis method to estimate the relationship between the contribution factors and the reviewers' impact index, exposure-impact index and Average RHR, respectively. The independent variables were the contribution factors, and the dependent variable was the value of reviewers measured by each method. For each group of the data, the linear regression analysis investigated the relationship between the dependent and independent variables, in order to find the contribution factors which are strongly related to reviews' impact

Table 5.3: The descriptions of each level in the organization dimension of the review

Scales	Description of organization
5	Its structure is clear and easily recognized; Most of the paragraphs have clear topic with detailed description; Its language is fluent and easy to read and understand.
4	It is paragraphed clearly, but not organized in a clear structure; Some of the paragraphs have clear topic with detailed description; Its language is generally fluent and easy to read and understand.
3	It is not paragraphed clearly or just has one long paragraph; Most of the paragraphs do not have clear topic. If it has one paragraph, the topic is not clear but with detailed description; Its language has no obvious problem for reading and understanding.
2	It has only one paragraph; It has no clear topic or detailed description; Its language is not easy to read and understand.
1	It has only one paragraph with few sentences; It has no topic or detailed description; Its language is very poor, and difficult to read and understand.

index, exposure-impact index and Average RHR.

5.3.3 Results and Analysis

In this subsection, human evaluation results are presented to show the comparison of review credibility posted by the reviewers with higher impact index, exposure-impact index and Average RHR. Then, we present the results of linear regression analysis.

Human Evaluation Results

For each group of data set, the evaluation results of the reviews chosen from the reviewers with high value of three methods were collected from the corresponding group of human raters. The average level of the reviews in each dimension posted the reviewers ranked top by three methods for each group of data are shown as Figure 5.6. From the figure, we can observe that the reviews posted by the reviewers with high impact index and exposure-impact index obtain much higher level in each dimension than those with high Average RHR. For instance, the level of reviews posted by the reviewers with high impact index and exposure-impact index is higher in the

Table 5.4: The descriptions of each level in the inforamtion dimension of the review

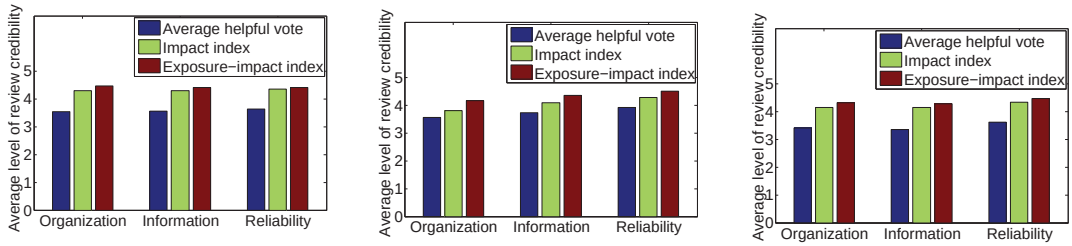
Scales	Description of information
5	The information it describes is relevant to the reviewed object; It includes sufficient basic information and some unique information (e.g. something the reviewer experienced personally); The information about the object is detailed and useful (e.g. it describes several different aspects in detail).
4	Most of the information it describes is relevant to the reviewed object; It includes some basic information about the object; The information bout the object is generally detailed and useful.
3	Some of the information it describes is relevant to the reviewed object; It includes a little basic information about the object; The information about the object is general. (e.g. it describes only one or no aspect in detail)
2	A little information it describes is relevant to the reviewed object; It includes just one piece of basic information about the object; Or it includes some basic information about the object, which is too common and easily obtained.
1	The information it describes are not relevant to the reviewed object.

organization dimension than that with high Average RHR, by 7.00%-22.00% and 16.81%-26.98%, respectively. Moreover, the level of reviews posted by the reviewers ranked top by the exposure-impact index is higher in each dimension than that by the impact index.

Human evaluation results manifest that the reviews posted by the reviewers with high impact index and exposure-impact index are more credible than those with high Average RHR. Therefore, our methods work more effectively than the Average RHR to help find credible reviews. That's because both the impact index and exposure-impact index methods evaluate reviewer credibility by considering two dimensions, including expertise and trustworthiness, while the Average RHR methods tend to emphasize the reviewers who have posted few reviews possibly implying a narrow range of expertise, but nevertheless obtained high Average RHR. Particularly, the exposure-impact index method performs better than the impact index method, because it assesses the expertise of the reviewer by considering not only the review count indirectly, but also the the destination count directly.

Table 5.5: The descriptions of each level in the reliability dimension of the review

Scales	Description of reliability
5	The information of the object includes comprehensive and convincing specifics, examples, or data, and can be accepted as truth; The opinion is fair and unbiased, with detailed personal experience (e.g. including date, time, person, or what happened) as evidences, which can support the opinion.
4	The information of the object includes some convincing specifics, examples, or data, and can be generally accepted as truth; The opinion is generally fair, with some detailed personal experience as evidence, which can generally support the opinion, though not sufficiently. And there may be a few biased opinions.
3	The information of the object includes a few convincing specific, example, or data, and can be generally accepted as truth; Some of the opinions are generally fair, with a few personal experiences as evidence. And there are some biased and unfair opinions.
2	The information of the object is very general without any detail. And it is difficult to accept the information as truth; The opinion is biased and unfair, or the personal experience can not support the opinion.
1	There is little or no basic information of the object. The information is seems to be false; The opinion is expressed in a very emotional and extreme way, and is unfair without any evidence. Its purpose is to boast of or attack the reviewed object.



(a) (b) (c)

Figure 5.6: The average level of reviews in each dimension posted by reviewers ranked top by three methods on three data sets (a) D-HongKong (b) D-NewYork (c)D-Longdon

Linear Regression Analysis Results

The results obtained by linear regression analysis are presented in Figure 5.7 which shows the most related contribution factors to reviewers' impact index exposure-impact index and Average RHR. Each sub-figure, such as Figure 5.7(a), represents

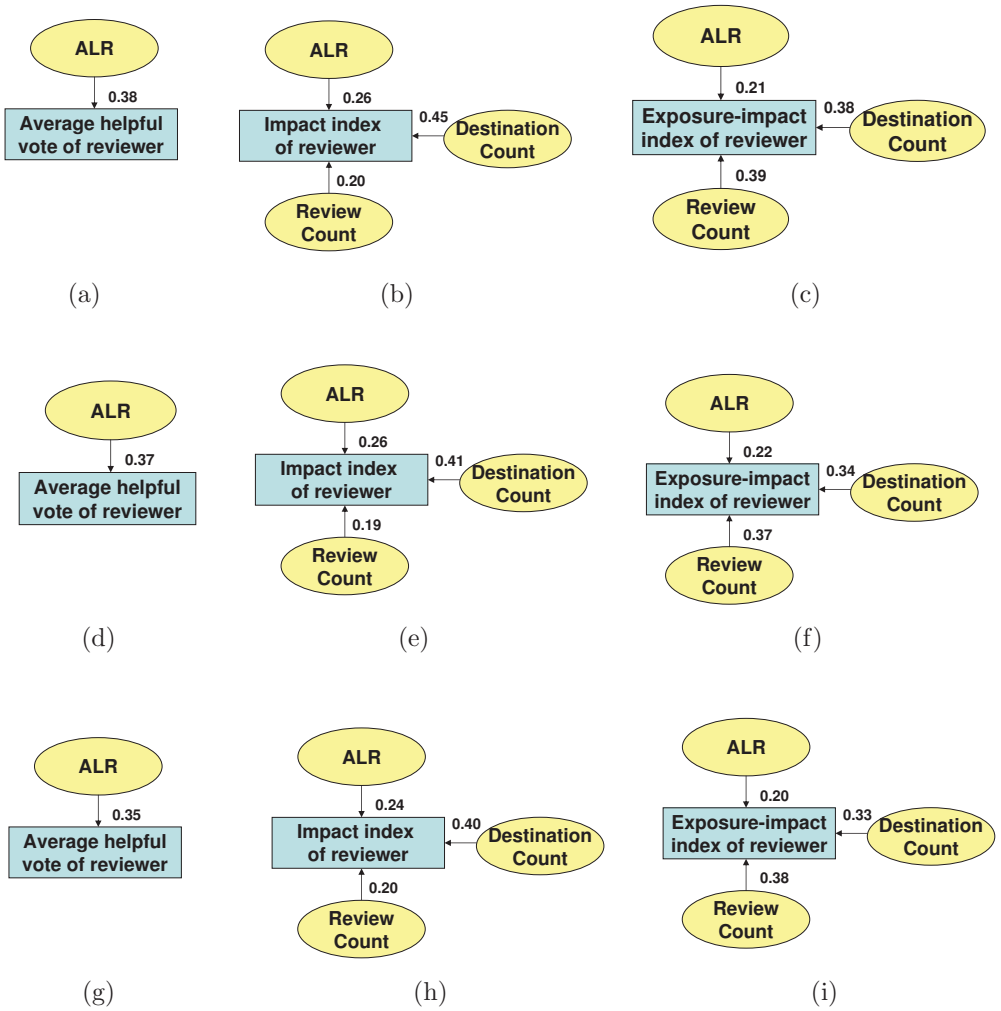


Figure 5.7: Results of linear regression analysis between contribution factors and reviewer's Average RHR, Impact Index, and Exposure-Impact Index. (a)-(c) D-HongKong; (d)-(f) D-NewYork; (g-i) D-London.

the model generated by linear regression which fits the data well. And the weights of the edges are the values of Beta , which is the standardized coefficients of the model. They are the estimates of the correlations between the independent variables and dependent variable that have been standardized with the variance as 1.

From the figure, we can observe that the Average RHR of reviewers is strongly related to the average length of review (ALR), while the impact index and exposure-

Table 5.6: Two reviewers of D-HongKong who are ranked lower by Average RHR

Reviewer name	Ranking order by Average RHR	Ranking order by I-Index	Ranking order by EI-Index	Destination Count	Review Count
ct-cruisers	908	15	5	125	416
Fiver75	3266	28	6	87	983

impact index are positively related to the Destination Count, Review Count and ALR. Although the Average RHR is able to find credible reviewers with high value of Average RHR, it misses some reviewers who have been to many destinations and posted numerous reviews showing high level of expertise, but the value of Average RHR is not so high. For instance, Table 5.6 shows two reviewers' information and their ranking orders according to the value of different methods. Both of them have been to many destinations and posted many reviews with higher exposure and expertise. They would not have been discovered by the Average RHR method due to the lower ranking, but could be discovered by our method with much higher ranking.

5.4 Summary

To help travellers search credible reviews online, we proposed one measurement and a variant to quantify the credibility of reviewers in TripAdvisor. The Impact Index was proposed to evaluate the reviewer credibility by considering expertise and trustworthiness based on the number of reviews and the number of helpful votes. To represent the the multi-faceted nature of credibility, the Impact Index was further improved into Exposure-Impact Index by considering in addition the number of destinations on which a reviewer has posted reviews. Experimental results have shown that both the Impact Index and the Exposure-Impact Index work more effectively than average RHR to quantify the credibility of reviewers to help find credible reviews. Additionally, the Impact Index and Exposure-Impact Index can discover some credible reviewers that the average RHR missed.

Chapter 6

Validating Reviewer Credibility Quantification Across Diverse Travel Communities

6.1 Introduction

In Chapter 5, we proposed an Impact Index (I-Index) and Exposure-Impact Index (EI-Index) that quantitatively measures reviewer credibility to help travellers search for credible information from online travel communities. Both methods measure reviewer credibility by evaluating the expertise and trustworthiness simultaneously based on one of the contribution factors of reviewer (the number of reviews posted by the reviewer or the number of destinations on which the reviewer posted reviews) and the number of helpful votes received by the reviews. Previous results on the data from TripAdvisor have shown that both the Impact Index and the Exposure-Impact Index work more effectively than the average helpful vote (Average RHR) to quantify the credibility of reviewers. Additionally, the Impact Index and Exposure-Impact Index can discover some credible reviewers that the Average RHR missed.

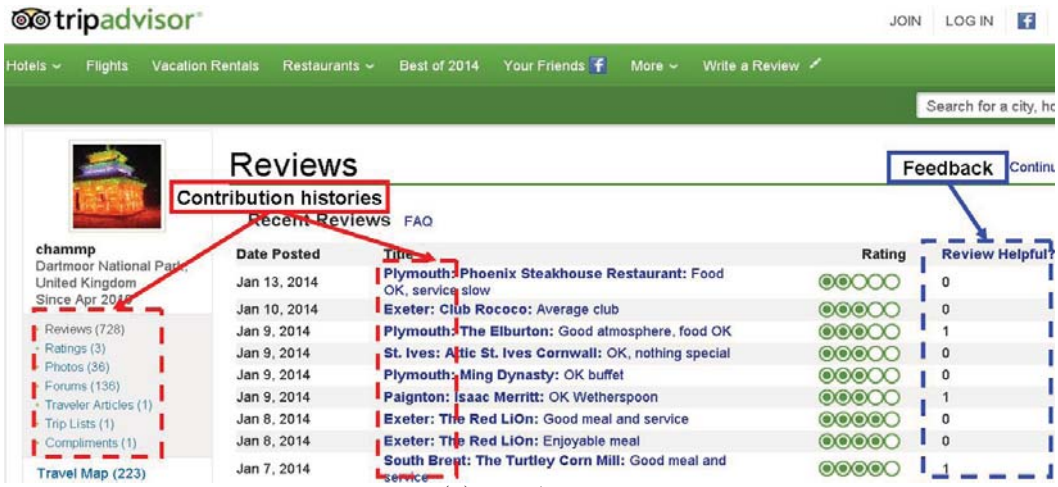
To make further investigation into the effectiveness and applicability of the proposed methods across diverse travel communities, this chapter evaluates them on several data sets collected from two rather different online travel communities:

TripAdvisor, the world's largest travel community, and Qunar, one of the most popular travel communities in China. Differences between the two communities are also compared and analyzed in this chapter to validate that our methods are applicable to diverse travel communities.

The rest of this chapter is organized as follows. Section 6.2 compares the differences between two travel communities. Section 6.3 provides quantitative comparison of two data sets collected from different travel communities. In Section 6.4, experimental results are presented to demonstrate the effectiveness and applicability of the Impact Index and the Exposure-Impact Index across diverse travel communities. Finally, we conclude this chapter by giving some directions for future work.

6.2 Analysis of Travel Communities Online

Recently, searching for travel information has become one of the most popular online activities and such information is increasingly used by travellers to facilitate decision making [33]. Hence, Lonely Planet, IgoUgo, TripAdvisor, and other similar travel communities become well known among international travellers. Meanwhile, several Chinese travel communities, such as Qunar, Ctrip and eLong are becoming popular in China. In this section, we analyze and compare two representative travel communities, TripAdvisor, the largest travel community in the world, and Qunar, one of the most popular Chinese travel communities.



(a) TripAdvisor

(b) Qunar

Figure 6.1: Contribution histories of a reviewer on TripAdvisor and Qunar respectively

Both TripAdvisor and Qunar conform to the reviewer-review-feedback mode. They allow reviewers to post reviews, upload photos, give ratings, and answer questions. They also keep a record of these contribution histories, as shown in Figure 6.1. The contribution histories are mainly composed of several contribution factors, as listed in Table 5.1 in section 5.2. In addition, both communities allow other travellers to vote on the helpfulness of reviews as a kind of feedback, and require

that travellers can only vote once for each review.

Even though both communities follow the same mode, there are significant differences between them, as shown in Table 6.1. First, the scale of TripAdvisor is much bigger than Qunar. One possible reason is that it was founded earlier and has accumulated more reviews. TripAdvisor has become the world's most prominent online travel platform since 2007 [33, 72, 78, 116]. But Qunar claimed that they built the world's largest Chinese hotel review system with 1 million reviews in 2010[1]. Second, reviewers from TripAdvisor spread all over the world, most of whom are Europeans and Americans (60%-70%)[2], while Qunar is only focused on China. Compared to developed western countries such as the US, tourism development in China is still at an early stage [114]. Third, reviews from TripAdvisor are mainly written in English, and Qunar Chinese. Fourth, destinations in TripAdvisor on which reviewers posted reviews are around the world, but those from Qunar tend to be in China. Although international travel from China is already growing, its development level remains relatively low[3]. Fifth, reviewers from TripAdvisor can post reviews on diverse targets, including hotels, restaurants, and attractions, but Qunar's reviews only hotels.

In terms of those points above, reviews, destinations and helpful votes of TripAdvisor are different from those of Qunar. For TripAdvisor, the number of reviews posted by reviewers, the number of destinations on which the reviewers has posted reviews and the number of helpful votes received by reviews are generally much higher than those of Qunar.

[1] http://www.qunar.com/site/zh/zhMilestones.shtml

[2] http://www.onecaribbean.org/wpcontent/uploads/CTOTripAdvisorCWeek2013Paganelli.pdf

[3] http://en.cnta.gov.cn/html/2012-7/2012-7-26-10-50-21009_1.html; http://sete.gr/_fileuploads/entries/Online%20library/GR/140128_BCG_Winning_the_Next_Billion_Asian_Travelers_Dec_2013.pdf

Table 6.1: Comparison between TripAdvisor and Qunar

Aspect	TripAdvisor	Qunar
Founded Time	Founded in 2000, became popular earlier	Founded in 2005, became popular later
Reviewer	Reviewers are spread all over the world	Reviewers are from China
Main language	English	Chinese
Destination	Destination is throughout the world	Destination is mainly in China
Reviewing target	Reviews are related to hotels, restaurants and attractions	Reviews are only related to hotels
Reward	Reviewers only obtain different star badges according to the number of the reviews they contributed	Reviewers can obtain gifts or free hotels according to their contributions.

6.3 Quantitative Comparison of Two Diverse Travel Communities

We collected different data sets from TripAdvisor and Qunar to examine the effectiveness and applicability of I-Index and EI-Index. Moreover, we provided a detailed study on the differences between the data sets collected from diverse communities, and then cleaned them by removing some possible manipulators.

6.3.1 Data Collection

Table 6.2: Description of the original data sets

Data sets	No. of reviewers	No. of reviews
T-Beijing	6833	260038
T-Shanghai	5893	241565
T-HongKong	12636	494677
Q-Beijing	48245	200819
Q-Shanghai	40076	194960
Q-HongKong	21178	133292

We developed two web crawlers to collect data from TripAdvisor and Qunar respectively. From Aug to Nov 2013, the crawler for TripAdvisor collected all reviewers who have posted reviews on three destinations: Beijing, Shanghai, and Hong Kong. Then, all reviews of these reviewers written in English were downloaded

from TripAdvisor. At the same time, reviewers on the same destinations and their reviews were downloaded from Qunar. These three destinations were chosen because they are popular travel destinations located in China, ensuring that sufficient reviewers can be acquired from both TripAdvisor and Qunar. Along with each review, the crawlers also downloaded metadata of reviewers, such as the number of helpful votes and the contribution factors. Statistics of the downloaded data sets are shown in Table 6.2.

6.3.2 Quantitative Comparison of Data Sets

We investigated the distributions of the number of reviews, destinations on which reviewers posted reviews, as well as the number of helpful votes received by reviews, to present the differences between the data from TripAdvisor and Qunar.

Reviews Posted by Reviewers

Figure 6.2 shows the distribution of the number of reviews with y-axis (number of reviewers) on logarithmic scale. More than 70% of TripAdvisor reviewers have posted more than 10 reviews and around a quarter of reviewers even more than 50 reviews. However, only about 10% of Qunar reviewers have posted more than 10 reviews. Moreover, the distribution of the number of reviews from TripAdvisor is unimodal, while Qunar bimodal, with a minor mode composed of a small number of reviewers strangely distributed between 25 and 50 reviews. We would come back to this issue of bimodality in next section. The distributions reflect that the reviewer from TripAdvisor tends to post much more reviews than that from Qunar. One possible reason is that TripAdvisor was developed earlier than Qunar [33, 72, 78], so the reviewer has accumulated more reviews. Another possible reason is that most reviewers from TripAdvisor are Europeans and Americans (60%-70%)[4] who travel

[4] TripAdvisor Trends: `http://www.onecaribbean.org/wp-content/uploads/CTOTripAdvisorCWeek2013Paganelli.pdf`

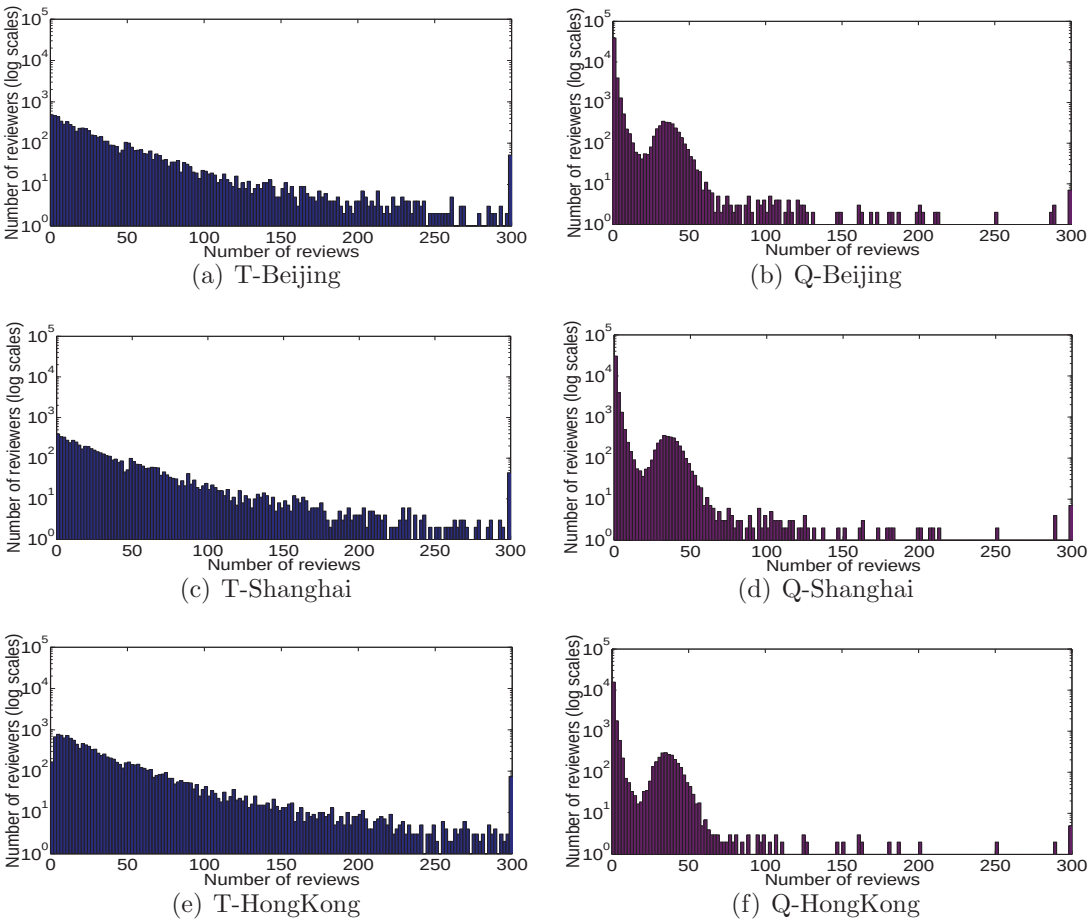


Figure 6.2: Distributions of the number of reviews posted by each reviewer (the tail of x-axis, including less than 1% reviewers, is truncated for clear illustration)

more than Chinese, and tourism development in China is still in an early stage, compared to western countries [114]. In addition, reviewers from TripAdvisor can post reviews on different things (hotels, restaurants, and attractions), but reviewers from Qunar only hotels.

Destinations on Which Reviewers Posted Reviews

Figure 6.3 shows the distribution of the number of destinations on which each reviewer posted reviews, with y-axis in logarithmic scale. The number of destinations on which reviewers from TripAdvisor posted reviews is intuitively distributed in a

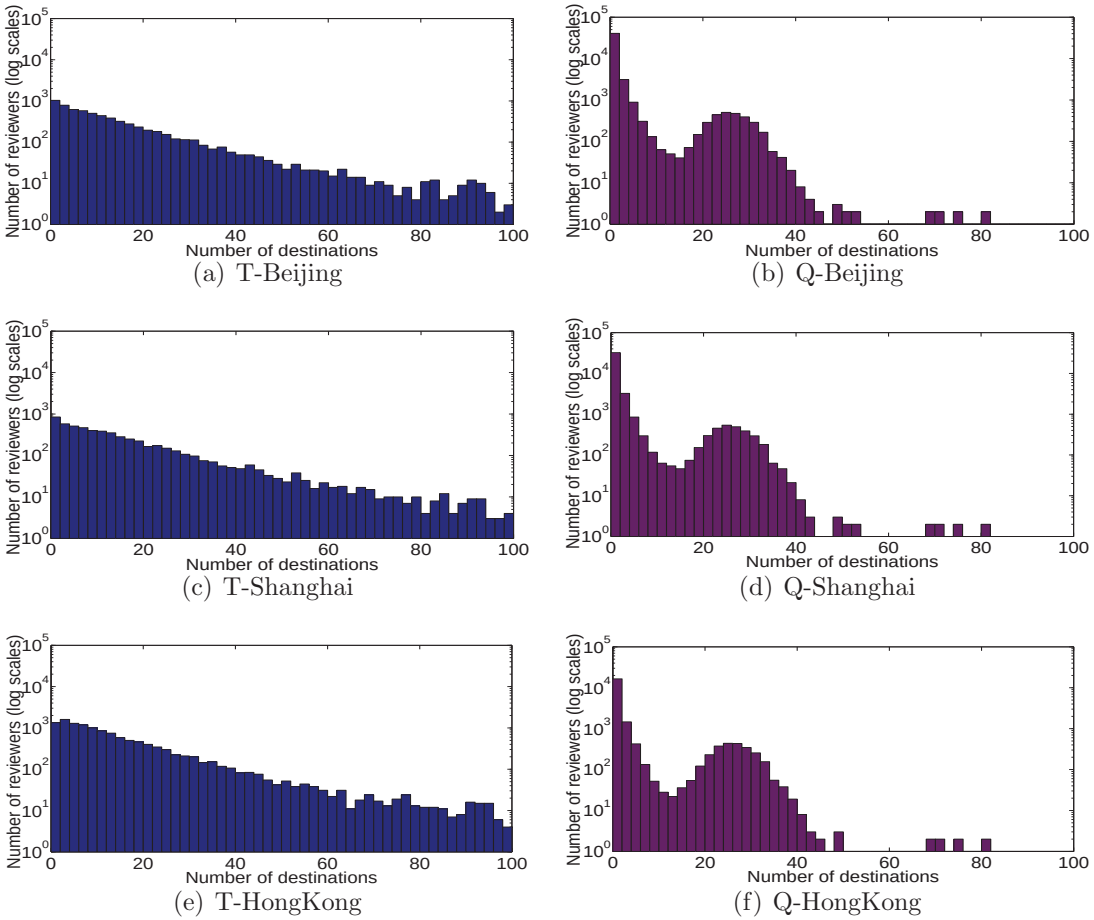


Figure 6.3: Distributions of the number of destinations on which each reviewer posted reviews (the tail of x-axis, including less than 1% reviewers, is truncated for clear illustration)

unimodal manner, while Qunar strangely bimodal. We would further investigate into this phenomenon in next section. Most TripAdvisor reviewers (around 70%) have posted reviews on more than 5 destinations, but only a small number of Qunar reviewers (about 10%) reach this level. Therefore, reviewers from TripAdvisor tend to post reviews on many destinations, but those on Qunar only one or two destinations. One possible reason is that destinations covered by TripAdvisor reviewers are around the world, but Qunar tend to be in China. Besides, reviewers from TripAdvisor can post reviews on diverse targets about travel, but Qunar only hotels, which may also

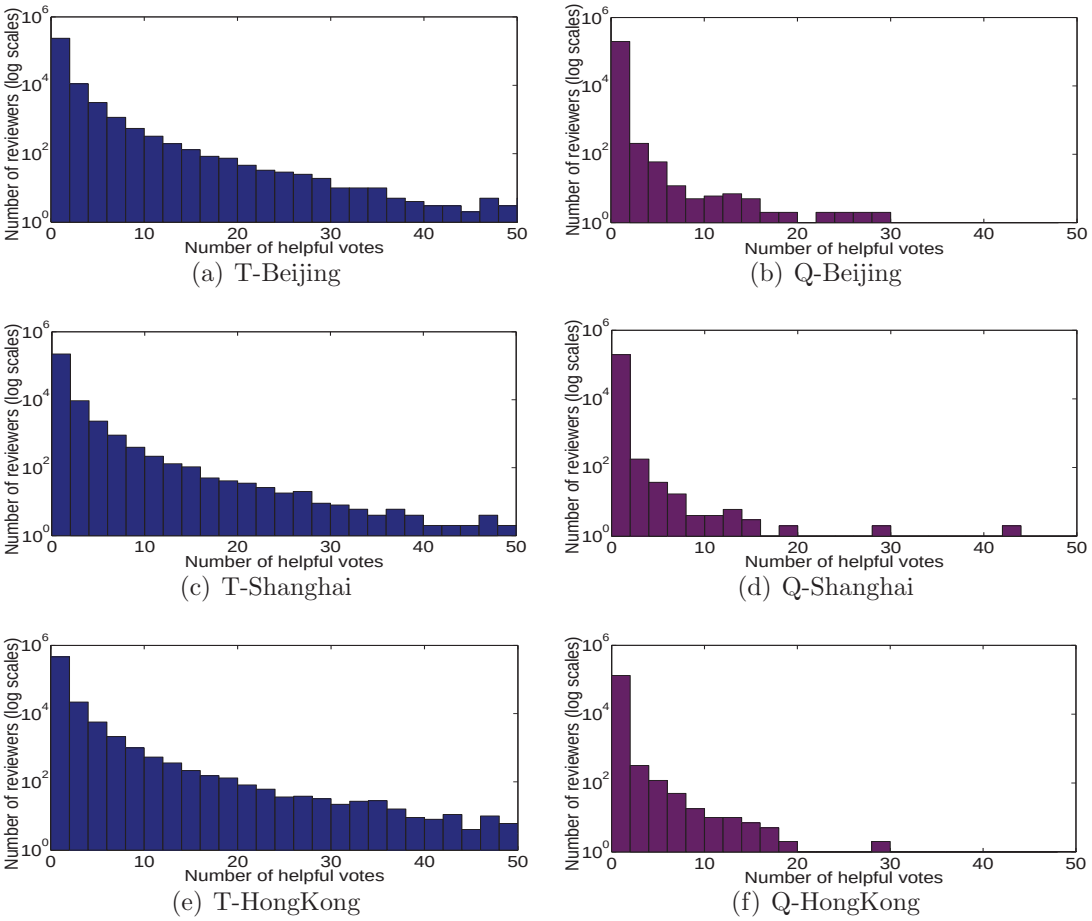


Figure 6.4: Distributions of the number of helpful votes received by each review (the tail of x-axis, including less than 0.3% reviewers, is truncated for clear illustration)

affect the number of destinations.

Helpful Votes Received by Each Review

Figure 6.4 shows the distribution of the number of helpful votes received by each review, with y-axis in logarithmic scale. As shown in Figure 6.4, above 30% of TripAdvisor reviews have received helpful votes, but only around 3% of Qunar reviews have obtained helpful votes. Thus, reviews from Qunar received much fewer helpful votes than those from TripAdvisor. One possible reason is that TripAdvisor is the world's largest travel community, attracting more travellers all over the world

to use it, but Qunar only famous in local China. For instance, TripAdvisor had nearly 260 million monthly unique visitors in 2013[5], while Qunar around 75 million [26]. Therefore, reviews from TripAdvisor may attract more travellers than those from Qunar to read and give helpful vote.

6.3.3 Data Cleansing

In most travel communities, anyone can post a review, rate a hotel or give a helpful vote without factual verification, leading to the occurrence of manipulation behaviors. Many travel communities have made some efforts to detect manipulation behavior, but the issue is till problematic for tourism communities, such as TripAdvisor [44, 67, 112]. In this work, we discovered that the issue is much worse in Qunar.

In this work, our main concern is quantifying reviewer credibility rather than handling all manipulation behaviors. Hence, we removed the data suspected of exhibiting manipulation behaviors on the posting of reviews and helpful votes. First, we developed a helpful vote-cleansing method to remove the reviewer who has received an abnormal number of helpful votes. Second, we removed the reviewer with a suspicious behavior that she/he always posted reviews with a same timestamp. For the data from TripAdvisor, the second cleansing work was not conducted because the detailed posted time of reviews is not shown to public. On the other hand, the data from TripAdvisor show very few signs of suspicious behavior. Although we cannot remove those reviewers, it is not as serious as in Qunar. Those reviewers have not contributed significantly to a bimodality anomaly.

Reviewer Cleaning Based on Helpful Votes

In the study of citations received by published academic papers, researchers have observed that a small number of papers have received a large number of citations,

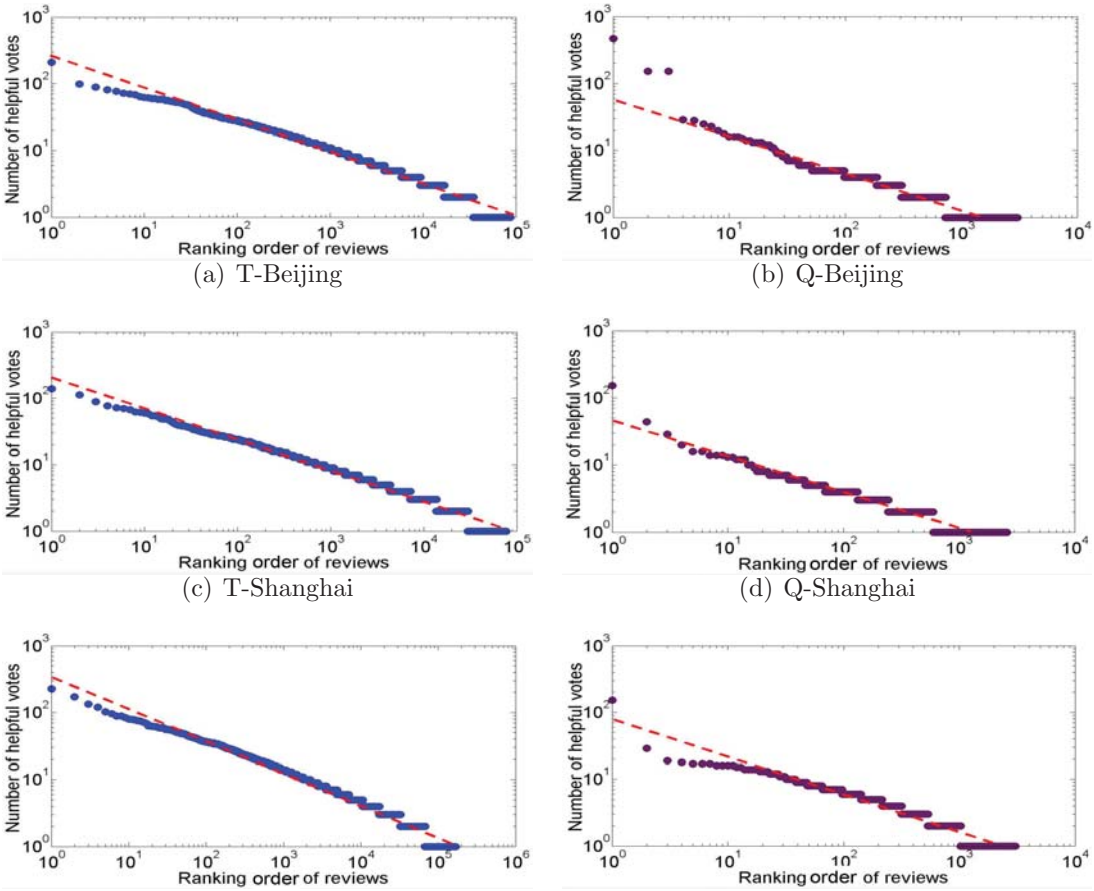
[5] Source: Google Analytics, worldwide data, July 2013

and most papers have received a small number of citations [17, 91]. The distribution of citations with reference to the ranked papers by the number of their citations actually follows Zipf's law [17, 91]. We found that the distributions of the number of helpful votes versus ranked reviews by the number of helpful votes behave in a similar manner as those of citations versus ranked papers, and the plot of the distribution in logarithmic scale follows quite closely a straight line, as shown in Figure 6.5, which indicates that the distribution of helpful votes also follows Zipf's law [17, 77], with formulation as follows:

$$f(x) = \frac{C}{x^{\theta}} \tag{6.1}$$

where $f(k)$ denotes the frequency of occurrence of the event at rank k, the parameter C is a constant, and θ is a positive parameter, which is also the slope of the straight line in the loglog plot.

To detect the manipulation behavior on helpful votes, we assumed that the distribution of the number of helpful votes versus ranked reviews follows Zipf's law and used Mean Square Error (MSE) to estimate the conformance to Zipf's distribution. Lower MSE means that the model better fits the data. Too high MSE indicates that there are some reviews received abnormal number of helpful votes, which makes the distribution of the number of helpful votes deviate far from Zipf's law, indicating possible manipulation behavior on helpful votes. In this work, we only aimed to filter out the reviewers with obvious manipulation behavior rather than all of the possible manipulators so that the impact of the manipulation on our evaluation can be reduced. Therefore, to remove more obvious manipulators, we ranked reviewers by MSE from high to low, and then removed those reviewers who were ranked in top 5%, indicating the distributions of the number of helpful votes of those reviewers deviate further than those of other reviewers who were ranked lower.

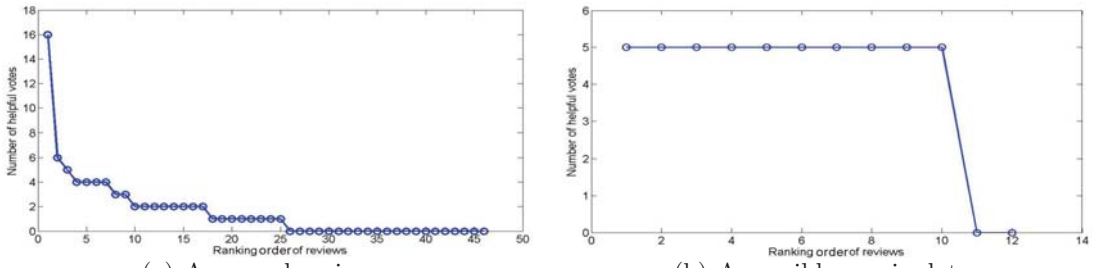


(a) T-Beijing (b) Q-Beijing

(c) T-Shanghai (d) Q-Shanghai

(e) T-HongKong (f) Q-HongKong

Figure 6.5: Loglog plot of the number of helpful votes versus ranking orders of reviews



(a) A normal reviewer (b) A possible manipulator

Figure 6.6: Distributions of the number of helpful votes received by a normal reviewer and a possible manipulator judged by Zipf's law

The initial value of θ (the slope of the straight line)is estimated by the linear regression method, because the log-log plot of the distribution of the number of

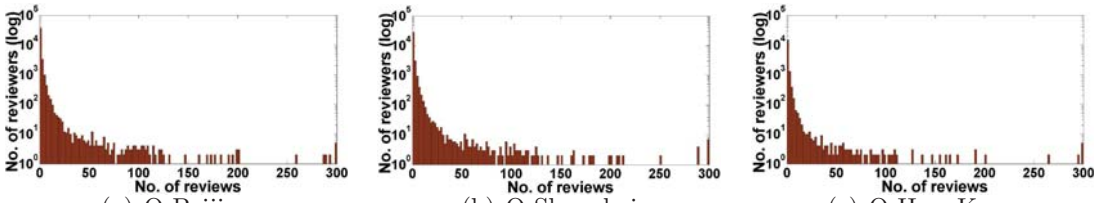
helpful votes versus ranking order of reviews should be a straight line according to our assumption. Then, we explored different values of θ around the estimated value based on the range [0, 2θ] and the step 0.02, to further minimize MSE. The selected value of θ is the one that best fits the data. For a reviewer, if the selected value of θ is very low or even zero, that means the number of helpful votes received by each review is very similar or even miraculously the same. Then, the reviewer was considered as more obvious manipulator. To filter out them, we ranked reviewers based on the value of θ from the lowest to the highest, and removed those who were ranked in top 5%, manifesting that the number of helpful votes of each review posted by those reviewers is too similar.

Figure 6.6 shows examples of a normal reviewer (Figure 6.3.3) and a possible manipulator (Figure 6.3.3) distinguished by the helpful vote-based manipulation detection. The normal reviewer has a small number of reviews received many helpful votes, and most reviews received a small number of helpful votes. However, the possible manipulator has 10 reviews (out of 12) which received the same number of helpful votes.

Reviewers Cleansing Based on Review-posting Timestamp

We found that some reviewers on Qunar always posted a lot of reviews with the same timestamp ("00:00:00") on different days. This phenomenon is very weird because there were other reviewers who posted reviews at normal time on the same day, manifesting that it is not caused by some database or system problem. Moreover, Qunar did not report that there were some system failures leading to this problem.

To decrease the impact of this possible manipulation behavior on our evaluations, those reviewers who have many reviews (over 50%) posted at "00:00:00" were filtered out. After that, the distributions of the number of reviews become unimodal as expected, as shown in Figure 6.7. Therefore, it is this possible manipulation that



(a) Q-Beijing (b) Q-Shanghai (c) Q-HongKong

Figure 6.7: Distributions of the number of reviews posted by each reviewer after data cleansing

leads to the bimodal distribution of original number of reviews and destinations of Qunar data, as shown in Figure 6.2 and Figure 6.3. Moreover, we found that most of the reviews with timestamp as "00:00:00" were posted in 2009 and 2010. The reason for this issue is out of our concern in this work. We only know that Qunar began to recruit "Try to sleep member" in 2009 and required the members to write reviews, providing them with very high salary and free hotels, which attracted nationwide attention. Then they claimed to be the world's largest online Chinese-language hotel review platform in 2010[6].

Cleaned Data Sets

According to an article of CNN[7], one of the signs of a fake review is that the review is the reviewer's only review. Moreover, the reviewers who posted only one or two reviews can not be chosen as credible reviewers by our methods because reviewers can not have high I-Index or EI-Index without posting a substantial number of reviews. Therefore, we also filtered out the reviewers who just posted one or two reviews. Furthermore, old reviews posted before 2008 were not considered, since there are very few reviewers who posted reviews in Qunar before 2008. The summary of data sets after cleaning is shown in Table 6.3. More than half (over 55%) of reviewers

[6] http://investor.qunar.com/phoenix.zhtml?c=252141&p=irol-govmilestone_pf

[7] http://edition.cnn.com/2014/03/09/travel/tripadvisor-decoded/index.html?hpt=hp_c5

Table 6.3: Data sets after cleansing

Data set	No. of reviewers	No. of reviews
T-Beijing	4049	115293
T-Shanghai	3270	97004
T-HongKong	9898	276131
Q-Beijing	2450	29083
Q-Shanghai	2003	26265
Q-HongKong	915	12689

from TripAdvisor were left and reviewers' average number of reviews is close to 30. However, only around 5% reviewers from Qunar were left and reviewers' average number of reviews is around 13.

6.4 Evaluation

The effectiveness and applicability of I-Index and EI-Index across travel communities was evaluated by comparing their performance against Average RHR on data sets collected from TripAdvisor and Qunar, taking into account the differences between the data sets.

Previous work [31, 98] pointed out that the prediction of source credibility and message credibility are fundamentally interlinked and influenced by each other. That means credible sources are likely to generate credible messages and credible messages are likely to originate from credible sources [31]. Furthermore, Fragale and Health [31] mentioned that individuals may use their evaluation of the message itself to infer source credibility, which indicates that believing a message makes people think the source is credible. Based on these observations, we evaluate the credibility of reviewers (sources) in terms of the quality of reviews (messages) posted by the reviewers. A team of human raters were invited to rate the quality of reviews posted by the reviewers with high values of I-Index, EI-Index and Average RHR. The rating of a reviewer is represented by the average rating of reviews posted by the reviewer.

Higher average ratings manifest that the quality of reviews posted by the reviewers are higher and the reviewers are more credible, so the measurement that returns these reviewers are more effective. Therefore, this evaluation by human raters forms the basis of the investigation into the effectiveness of I-Index and EI-Index.

To further examine the applicability of our measurements on diverse data sets, we evaluate the quality of rankings of reviewers returned by different measurements. The Spearman's rank correlation coefficient (Spearman's rho) is adopted to compare the correlation between the rankings of reviewers generated by each measurement and those obtained from human rating results. Higher positive Spearman's rho indicates that the ranking generated by the measurement is more consistent with the human ranking based on rating results, indicating that the reviewers ranked higher by the measurement are relatively more credible than those ranked lower, and therefore the corresponding credibility measurement is better.

6.4.1 Evaluation by Human Raters

We invited human raters to evaluate the credibility of reviews posted by the reviewers with high value of I-Index, EI-Index or Average RHR, because the credibility of reviewers (sources) can be assessed in terms of the quality of reviews (messages) [31, 98].

Human Rating Evaluation Method

Based on previous investigations [15, 27, 30, 69, 71, 97, 98] into the criteria in evaluating the credibility of messages or reviews, three dimensions were considered to rate the credibility of reviews, including organization, information and reliability, which are defined in section 5.3.2. We applied the 5-point Likert scale to evaluate each dimension, and the detailed description of each rating level in each dimension is shown in Table 5.3, 5.4 and 5.5.

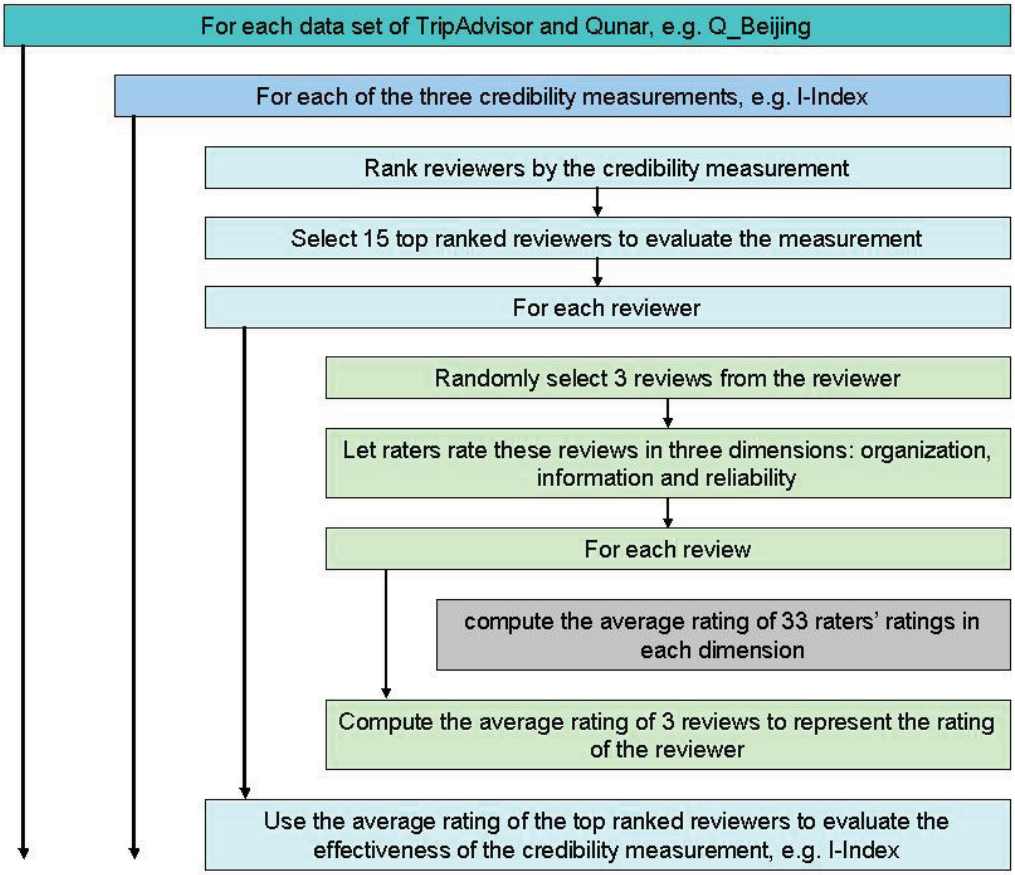
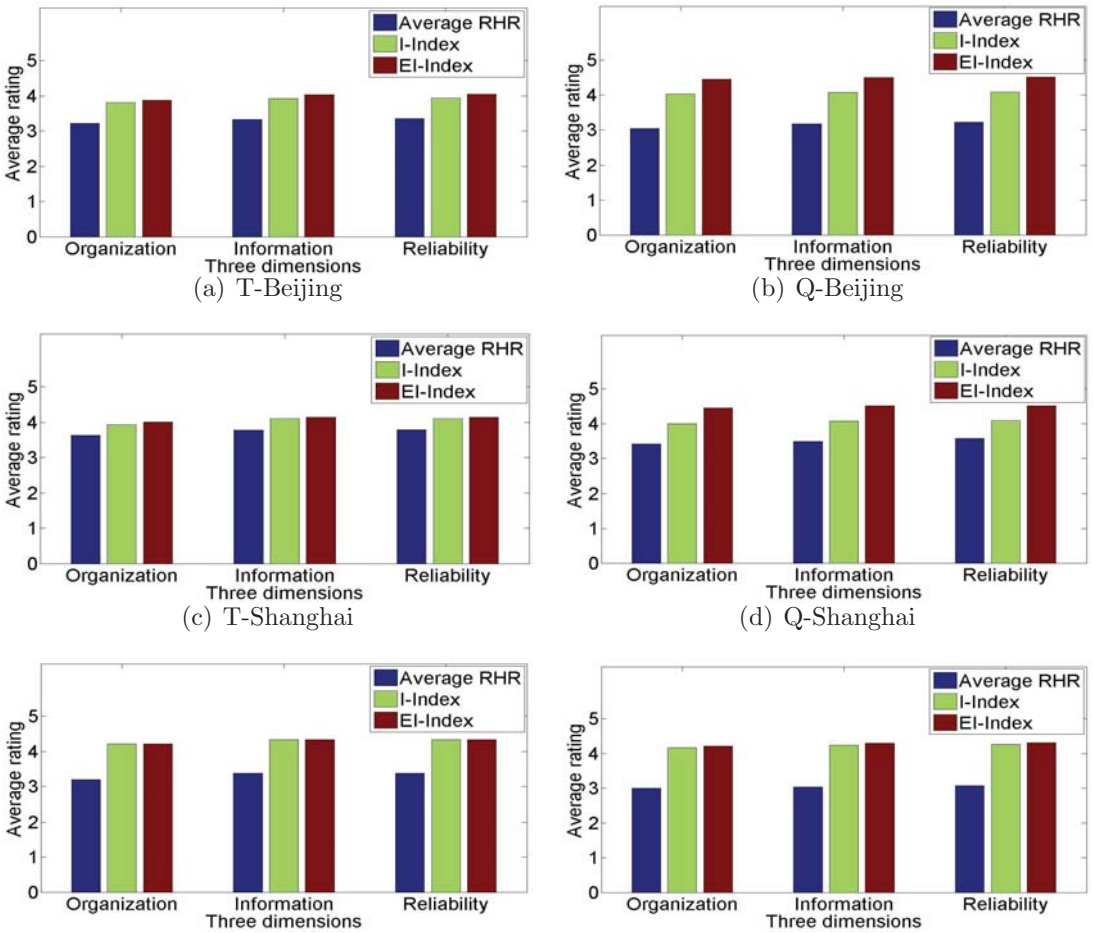


Figure 6.8: Implementation flow of human evaluation

We recruited 33 research students as human raters who are familiar with both English and Chinese, and let them understand the description of each rating level before rating work. The flow of human evaluation procedures is shown in Figure 6.8, using the data set Q-Beijing and the method of I-Index as an example. For a reviewer, three reviews were randomly chosen to be rated, and the rating of each review was denoted by the average value of 33 raters' ratings. Then, the average rating of three reviews was used to represent the rating of the reviewer.

Human Rating Results

Evaluation results for all data sets from TripAdvisor and Qunar were collected from human raters. As shown in Figure 6.9, the average rating of the top ranked reviewers



(a) T-Beijing (b) Q-Beijing

(c) T-Shanghai (d) Q-Shanghai

(e) T-HongKong (f) Q-HongKong

Figure 6.9: Distributions of the number of reviews posted by each reviewer

returned by I-Index and EI-Index are higher in each dimension for each data set than those returned by Average RHR method. Moreover, the advantage of our method on the data sets from Qunar is more obvious than that from TripAdvisor. For instance, the average rating of reviewers on TripAdvisor returned by I-Index and EI-Index is higher in the information dimension than those returned by Average RHR, by 18.2% and 19.7% on average of three data sets, respectively, while that on Qunar by 28.4% and 37.6% on average of three data sets, respectively. The human rating results suggest that reviews posted by the reviewers returned by I-Index and EI-Index are more credible than those returned by Average RHR for the data sets

from TripAdvisor and Qunar, especially Qunar. They further indicate that reviewers ranked high by our indices are more credible than those by the Average RHR for the data sets collected from both travel communities. Thus, our indices work more effectively on the data sets collected from both travel communities than Average RHR to measure reviewer credibility. We believe that this is because both I-Index and EI-Index measure reviewer credibility by considering expertise and trustworthiness dimensions simultaneously, while Average RHR tends to favor the reviewers who have posted few reviews, which manifest relatively few experiences, implying a narrow range of expertise, but nevertheless got high Average RHR.

Furthermore, the average rating of reviewers returned by EI-Index is a little bit higher in each dimension than that returned by the I-Index. Different from I-Index which evaluates the expertise of reviewers based on the number of reviews posted by the reviewer, EI-Index assesses the expertise by directly considering the number of destinations, and it emphasizes the credible reviewers who have better exposure, richer experience and broader knowledge.

6.4.2 Evaluation of Ranking Quality of Reviewers

To gain further insight into the effectiveness of I-Index and EI-Index across different travel communities, we evaluate them by assessing the quality of reviewer rankings they returned. Rankings of reviewers based on human ratings were viewed as benchmark. Human raters expect that a reviewer ranked higher provides better reviews. The Spearman's rank correlation coefficient (Spearman's rho) was applied to measure the correlation between the ranking returned by each method and benchmark in each dimension. Higher positive Spearman's rho corresponds to stronger positive correlation between the ranking of the measurement and the benchmark, which indicates that the ranking of the measurement is more consistent with the expectation of human raters. Lower positive or even negative Spearman's

rho means weaker positive or negative correlation between the ranking of the measurement and the benchmark, implying the ranking result deviate farther from the expectation of human raters.

The software of SPSS[8] was used to implement the analysis of Spearman's rho. As shown in Table 6.4, generally, the Spearman's rho between the benchmark and the ranking of either I-Index or EI-Index is significantly positive in each dimension, which indicates that the rankings of both I-Index and EI-Index are relatively consistent with the benchmark. It further implies that reviewers ranked higher by them tend to provide better reviews than those ranked lower, as the expectation of human raters. However, the Spearman's rho between the ranking of Average RHR and the benchmark is obviously negative for most data sets, which manifests that the ranking of Average RHR deviates far away from the expectation of human raters, implying that most reviewers ranked higher by Average RHR are unable to provide better reviews than those ranked lower. Therefore, the rankings of the top ranked reviewers returned by both I-Index and EI-Index are more consistent with the expectation of human raters than those returned by Average RHR. Moreover, the Speraman's rho between the ranking of EI-Index and the benchmark is a little bit higher than that between the ranking of I-Index and the benchmark, which suggests that the rankings of EI-Index is slightly more closer to the human judgments.

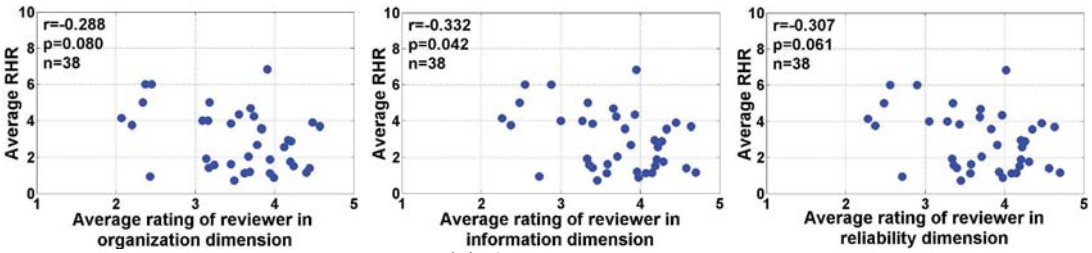
Figure 6.10 and Figure 6.11 provide several scatter plots of "T-Beijing" and "Q-Beijing" as examples, to further show the correlation between the ranking of reviewers returned by each measurement and benchmark. An increasing monotonic trend between average rating of reviewers and either I-Index or EI-Index can be easily observed from Figure 6.10(b)-6.10(c) and Figure 6.11(b)-6.11(c). However, Average RHR tends to decrease when the average rating of reviewer increases, as shown in Figure 6.10(a) and Figure 6.11(a).

[8] `http://www-01.ibm.com/software/analytics/spss/`

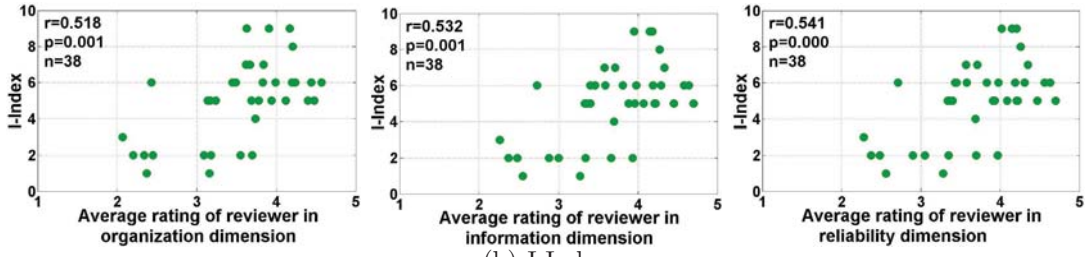
Table 6.4: Spearman's rho between the ranking of reviewers returned by each method and benchmark in each dimension

Data sets	Average RHR			I-Index			EI-Index		
	Org	Info	Reli	Org	Info	Reli	Org	Info	Reli
T-Beijing	-0.288	**-0.332**	-0.307	**0.518**	**0.532**	**0.541**	**0.603**	**0.606**	**0.612**
T-Shanghai	-0.082	-0.115	-0.089	0.294	**0.405**	**0.404**	**0.460**	**0.407**	**0.413**
T-HongKong	**-0.517**	**-0.487**	**-0.495**	**0.643**	**0.623**	**0.632**	**0.652**	**0.635**	**0.644**
Q-Beijing	**-0.554**	**-0.502**	**-0.519**	**0.412**	**0.361**	0.353	**0.595**	**0.540**	**0.552**
Q-Shanghai	**-0.554**	**-0.549**	**-0.533**	0.319	**0.385**	0.351	**0.562**	**0.627**	**0.628**
Q-HongKong	**-0.752**	**-0.750**	**-0.770**	**0.628**	**0.657**	**0.637**	**0.707**	**0.731**	**0.725**

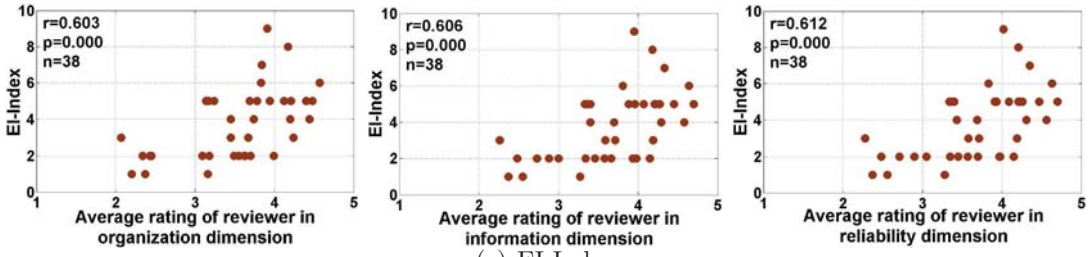
Bold values manifest significant correlation



(a) Average RHR

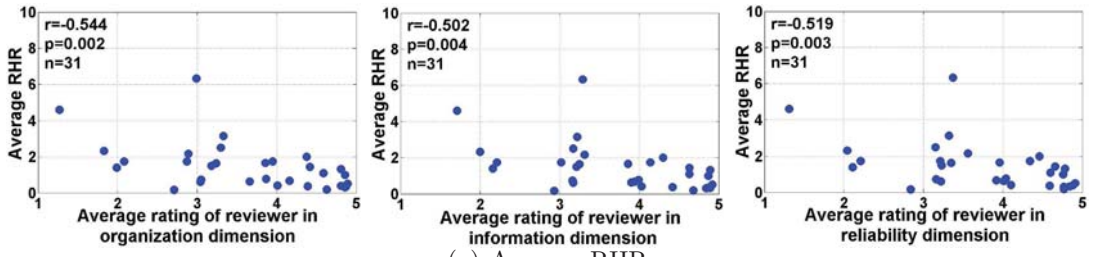


(b) I-Index

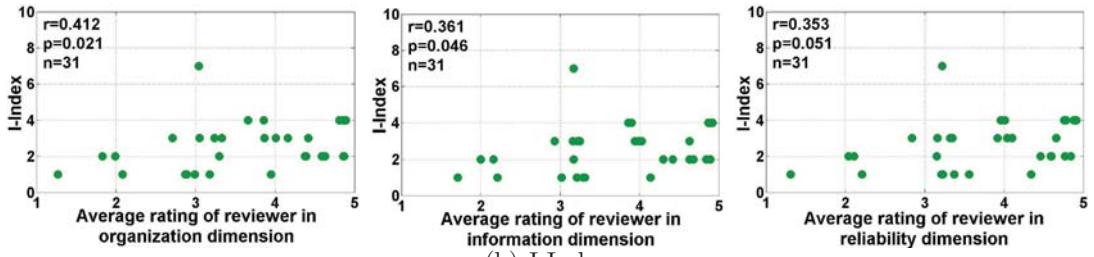


(c) EI-Index

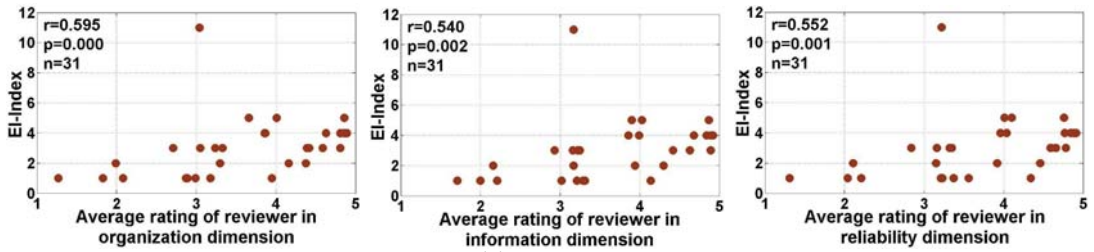
Figure 6.10: Spearman correlation between the ranking returned by each measurement and benchmark on the data set of T-Beijing



(a) Average RHR



(b) I-Index



(c) EI-Index

Figure 6.11: Spearman correlation between the ranking returned by each measurement and benchmark on the data set of Q-Beijing

The result of Spearman's rho revealed that both I-Index and EI-Index can provide good ranking of reviewers, which is consistent with human judgments. However, we found a ranking paradox [5] from the ranking result of Average RHR. The ranking paradox is a Simpson-like paradox, in which a trend appears in the ranking of individual data sets, but the reverse trend appears in the ranking of the aggregate of data sets [5]. According to the study conducted by Lee et al. [53], reviewers with higher Average RHR should be more credible and generate better reviews, and thus the ranking of reviewers returned by Average RHR was assumed to be positively correlated to benchmark. On the contrary, we got a negative correlation, which is a phenomenon of ranking paradox.

One reason for the paradoxical finding about Average RHR is that only considering the ranking of the top ranked reviewers is different from that of the overall data set. Those top ranked reviewers are easily to be dominated by some random noises that are not so credible and unable to provide good reviews. This is because the Average RHR favors the reviewers who posted fewer reviews, but obtained higher value of Average RHR. In the top ranked list, the higher the random noises were ranked, the more serious the dominant issue is. Considering the overall data set, the ranking of Average RHR is more likely positively correlated to benchmark.

For instance, the reviewer R_1, shown in Table 6.5, was ranked very high (the first) by Average RHR, but just posted 4 reviews on 4 destinations, which indicate that the experiences of this reviewer are very few, implying lower level of expertise. Some more credible reviewers were missed by Average RHR method. If a reviewer has posted a lot of reviews on many destinations and received a fair number of helpful votes, she/he will be ranked higher by both I-Index and EI-Index, such as the reviewer R_2 shown in Table 6.5, who have contributed 119 reviews on 44 destinations and obtain 350 helpful votes, manifesting higher exposure, expertise and trustworthiness. But this reviewer was ranked relatively lower by the Average RHR method.

Table 6.5: Rankings of example reviewers of T-Beijing returned by each method

Information	R_1	R_2	R_3	R_4
Ranking order by **Average RHR**	1	41	651	165
Ranking order by **I-Index**	1317	2	3	22
Ranking order by **EI-Index**	1148	2	1932	6
No. of reviews	4	119	750	56
No. of destination	4	44	5	56
No. of helpful votes	24	350	802	106
Average human rating in organization	2.45	4.17	3.63	3.94
Average human rating in information	2.88	4.18	4.15	4.21
Average human rating in reliability	2.90	4.21	4.15	4.21

Reviewers ranked higher by I-Index tend to have posted more reviews, no matter

how many destinations they posted on. For instance, the reviewer R_3 shown in Table 6.5 has posted 705 reviews on only 5 destinations. Although the exposure is not high, this reviewer was ranked high by the I-Index because she/he is knowledgeable about these particular destinations. Different from I-Index, EI-Index considers the number of destinations on which the reviewer has posted reviews. As shown in Table 6.5, the reviewer R_4 was ranked higher by EI-Index because she/he has posted reviews on 56 destinations, and tends to have higher exposure, indicating rich experiences and broad knowledge on diverse destinations. Moreover, both R_3 and R_4 were also rated higher by human raters. However, they were ranked lower by Average RHR.

These results suggest that Average RHR will provide travellers some credible reviewers but it misses some more credible reviewers who have higher level of expertise, but not so high Average RHR. Fortunately, both I-Index and EI-Index can discover these reviewers from both TripAdvisor and Qunar, especially when the data set is not so good, like the data set from Qunar. I-Index is able to discover credible reviewers who have posted a lot of reviews and received many helpful votes, manifesting high level of expertise and trustworthiness. EI-Index represents another side of nature of reviewer credibility in the tourism domain, and considers in addition the exposure of reviewers to discover credible reviewers with wide range of expertise. Furthermore, both I-Index and EI-Index can provide a better ranking of reviewers, to make it easier for travellers to find more credible reviewers.

6.5 Summary

To validate that both I-Index and EI-Index can be effectively applied to diverse travel communities, we evaluated their effectiveness on data sets collected from two diverse communities, by taking account into the differences in the language, the scale and the distributions of the data. In the experiment, we provided a detailed study on

the differences between the data sets collected from two communities and cleaned them by removing some possible manipulators. Then, we examined the effectiveness and applicability of our indices on those data sets. Experimental results show that both I-Index and EI-Index lead to results more consistent with human judgments than previous Average RHR on the data sets collected from both communities, especially when the data is not good. Our measurements can not only discover more credible reviewers missed by the Average RHR, but also provide a better ranking of reviewers. It was demonstrated that our indices for quantifying reviewer credibility are applicable to diverse travel communities with a reviewer-review-feedback mode.

Actually, our methods can also be extended into other domains, such as movies or books. But new domain should have a reviewer-review-feedback mode like the tourism domain, which record reviewers' history contribution factors, such as the number of reviews the reviewer has posted and the number of helpful votes the reviews have received, which can be used to measure two key dimension of reviewer credibility, including expertise and trustworthiness.

Impact Index method can be directly applied to measure reviewer credibility in movie or book communities, if the communities record the number reviews posted by reviewers and the number of helpful votes received by these reviews. For instance, in book domain, if reviewers have posted many reviews on books, that means the reviewers have bought and read many books, which manifests that they are more knowledgeable and competent to provide helpful and credible information about books. Therefore, these reviewers have high level of expertise. Moreover, if the reviews of reviewers have receive sufficient number of helpful votes, it implies that more users think the reviews are helpful and reliable. Hence, the reviewers have high level of trustworthiness.

Exposure-Impact Index can not be directly applied into other domains, because it makes use of the number of destinations to measure the expertise dimension

of reviewer credibility, but other domains, such as movies and books, have no "destination" factor. Therefore, other factor should replace "destination" to measure expertise dimension. For instance, the idea of Exposure-Impact Index can be applied to movie or book domain by using the number of categories of movies or books on which reviewers have posted reviews instead of the number of "destination" to evaluate expertise of reviewers.

Chapter 7

Conclusions and future work

This chapter draws conclusions on the thesis, and provides some possible future works related to the work done in this thesis.

7.1 Conclusions

The work of this thesis focused on improving existing tourism recommender systems, such as TripAdvisor and Qunar. We addressed in this thesis the crucial challenges in tourism recommender systems, in particular the problems of sparseness and cold-start recommendation and the credibility of information. Especially, we investigated the credibility of reviewers in detail.

- We applied rating inference method to augment ratings for alleviating data sparseness issue. Sentiment analysis on tourism reviews is used to infer ratings. Different from existing research, our work investigated several popular clustering methods integrated in different features to do unsupervised sentiment analysis. The preliminary results showed that hierarchical clustering algorithms (six traditional hierarchical and hierarchical co-clustering) obtain more accurate clustering results than flat clustering (K-means and co-clustering) for tourism data sets. Especially, the hierarchial co-clustering method outperform other methods, benefiting from its hierarchy, feature clusters and feature

reduction at each level, which may reduce the dimensionality. The results also suggested that only using Part-of-speech or opinion words obtained similar or even better results than those using all the words. This can reduce the overall dimensionality for review data. What need further improvement is that only using these unsupervised approaches with these features for sentiment analysis are not enough to divide reviews into so fine scales.

- We proposed to apply demographic recommender system for predicting ratings on attraction, aiming at overcoming the cold-start problem. Based on the features extracted from travellers' demographic information, we examined differen machine learning methods to determine the applicability of these methods and the demographic information for tourism recommendation. Experimental results show that the demographic information with machine learning methods are applicable to predict traveller ratings. What we still need to improve is that using demographic along only achieve limited accuracy.

- It's very challenging for existing recommender systems to handle uninformative, biased and even false information. We proposed a method, Impact Index, and its variant, Exposure-Impact Index to quantify the credible of reviewers, with the purpose of searching for more credible information. Both Impact Index and Exposure-Impact Index measure reviewer credibility by evaluating the expertise and trustworthiness simultaneously based on the contribution factors of reviewers and helpful votes of reviews posted by the reviewers. Our experimental results show that both Impact Index and Exposure-Impact Index perform more effectively than previous Average helpful vote (Average RHR) to find credible reviewers. Moreover, these methods can discover some credible reviewers missed by existing method.

  Furthermore, we evaluated Impact Index and Exposure-Impact Index on

several data sets collected from TripAdvisor and Qunar, to validate the applicability and effectiveness of them across diverse travel communities. We compared the differences of TripAdvisor and Qunar in detail. We also found some manipulation issues, which are even worse in Qunar. Then, two manipulation detection methods were used to clean the data sets. One is based on the distribution of helpful votes, and the other one is based on the timestamps of reviews. Experimental results on the cleaned data show that both Impact Index and Exposure-Impact Index obtained results much closer to the expectation of human raters for the data sets from TripAdvisor and Qunar, despite the differences in language, the scale and the distribution of the data. Therefore, both of them are very promising measurements of the credibility of reviewers and hence their reviews.

7.2 Future Work

Related topics for the future research work are listed below.

Firstly, we found that using demographic information only is insufficient to make accurate recommendations. As we have mentioned, all of the known recommender systems have their own strengths and weaknesses. Many researchers have proposed that hybrid recommender systems combining two or more recommendation techniques can get better performance by making use of the advantages of the systems. Therefore, in our future work, we will focus our research on hybrid tourism recommender system. The hybrid system will incorporate historical ratings, sentiment analysis results form textual reviews, reviewer profile, description information about attractions or hotels.

Secondly, the proposed Exposure-Impact Index has put in equal emphasis on the number of destinations and the number of helpful votes. However, different weighting

schemes maybe more appropriate for different purposes. Therefore, in the future, we will investigate the impact of adjusting the weights of the two dimensions for the Exposure-Impact Index, and then develop more effective methods to evaluate the credibility of reviewers in tourism for helping travellers search for credible reviews.

Thirdly, this thesis has presented two methods to quantify reviewer credibility independent from any particular destination. We only aimed at helping traveller search for credible reviewers according to reviewers' overall contribution. However, traveller may need to find expert or credible reviewers on one particular destination. In the future work, we will develop a system to find destination-specific expert or credible reviewers to satisfy the different requirements of travellers.

Fourthly, the Impact index and Exposure-Impact Index only consider contribution factors and helpful votes. However, the time factor is also important to measure reviewer credibility. For instance a newer review posted by a reviewer is generally more up to date, and provides travellers with the latest information, which is more useful and reliable to help travellers make travel decisions. Therefore, we will consider in addition the time factor to improve credibility measurement.

Finally, based on our investigation into the data sets collected from TripAdvisor and Qunar, the data show some signs of suspicion manipulation behavior. Since our main concern in Chaper 5 and Chapter 6 was quantifying reviewer credibility rather than handling all manipulation behaviors, we only developed two methods to remove some data with a suspicious manipulation behaviors. In the future, will focus our work on manipulation detection in online tourism.

Bibliography

[1] Adomavicius, G. and Tuzhilin, A. Toward the next generation of recommendersystems: A Survey of the state-of-the-art and possible extensions. *IEEE Transactions on Knowledge and Data Engineering*, 17(6), 734–749, 2005.

[2] Annett, M., and Kondrak, G. A comparison of sentiment analysis techniques: Polarizing movie blogs. *Advances in Artificial Intelligence*, pp. 25–35, 2008.

[3] Armstrong, C. L., and Nelson, M. R. How Newspaper Sources Trigger Gender Stereotypes. *Journalism & Mass Communication Quarterly*, 82(4), 820–837, 2005.

[4] Balabanovic, M., and Shoham, Y. Fab: content-based, collaborative recommendation. *Communications of the ACM*, 40(3), 66–72, 1997.

[5] Bargagliotti, A. E., Greenwell, R. N. Statistical significance of ranking paradoxes. *Communications in Statistics-Theory and Methods*, 40(5), 916–928, 2011.

[6] Basu, C., Hirsh, H., and Cohen W. Recommendation as classification: Using social and content-based information in recommendation. *In Proceedings of the 15th National Conference on Artificial Intelligence*, Madison, WI, pp. 714–720, 1998.

[7] Berger, C. R., and Calabrese, R. J. Some explorations in initial interaction and beyond: Toward a developmental theory of interpersonal communication. *Human communication research*, 1(2), 99–112, 1975.

[8] Billsus, D., and Pazzani,M. User Modeling for Adaptive News Access. *User-Modeling and User-Adapted Interaction* 10(2–3), 147–180, 2000.

[9] Boiy, E., and Moens, M. F. A machine learning approach to sentiment analysis in multilingual Web texts. *Information retrieval*, 12(5), 526–558, 2009.

[10] Bornmann, L., Mutz, R., Hug, S. E., and Daniel, H. D. A multilevel meta-analysis of studies reporting correlations between the h index and 37 different h index variants. *Journal of Informetrics*, 5(3), 346–359, 2011.

[11] Burke, R., Hammond, K., and Young, B. The FindMe approach to assisted browsing. *IEEE Expert*, 12(4), 32–40, 1997.

[12] Burke, R., Hybrid web recommender systems. *The Adaptive Web*, pp. 377–408, 2007.

[13] Chang, C. C., and Lin, C. J. LIBSVM: A library for support vector machines. *Software available at* , `http://www.csie.ntu.edu.tw/~cjlin/libsvm`, 2001.

[14] Chen, C., Ibekwe-SanJuan, F., SanJuan, E., and Weaver, C. Visual analysis of conflicting opinions. *In IEEE Symposium on Visual Analytics Science and Technology*, pp. 59–66, 2006.

[15] Chen, C. C., and Tseng, Y. D. Quality evaluation of product reviews using an information quality framework. *Decision Support Systems*, 50(4), 755–768, 2011.

[16] Cho, J., Kwon, K., and Park, Y. Q-rater: A collaborative reputation system based on source credibility theory. *Expert Systems with Applications*, 36(2), 3751–3760, 2009.

[17] Clauset, A., Shalizi, C. R., and Newman, M. E. Power-law distributions in empirical data. *SIAM review*, 51(4), 661–703, 2009.

[18] Claypool, M., Gokhale, A., Miranda, T., Murnikov, P., Netes, D., and Sartin, M. Combining content-based and collaborative filters in an online newspaper. *SIGIR 99 Workshop on Recommender Systems: Algorithms and Evaluation*, Berkeley, CA, 1999.

[19] Costas, R., and Bordons, M. Is g-index better than h-index? An exploratory study at the individual level. *Scientometrics*, 77(2), 267–288, 2008.

[20] Dave, K., Lawrence, S., and Pennock, D. M. Mining the peanut gallery: Opinion extraction and semantic classification of product reviews. *Proceedings of the 12th International Conference on World Wide Web*, pp. 519–528, 2003.

[21] Dhillon, I., Kogan, J., and Nicholas, C. Feature selection and document clustering. *In Survey of text mining*, pp. 73–100, 2004.

[22] Dhillon, I., Mallela, S., and Modha, D. Information-theoretic co-clustering. *In Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 89–98, 2003.

[23] Efron, M. Cultural orientation: Classifying subjective documents by Cociation analysis. *In Proceedings of the AAAI Fall Symposium on Style and Meaning in Language, Art, Music, and Design*, pp. 41–48, 2004.

[24] Egghe, L. An improvement of the h-index: The g-index. The International Society for Informetrics and Scientometrics Newsletter, 2(1), 8–9, 2006.

[25] Egghe, L. Theory and practices of the g-index. *Scientometrics*, 69(1), 131–152, 2006.

[26] Financial Tear Sheet, Qunar, `http://investor.qunar.com/Tearsheet.ashx?c=252141`

[27] Flanagin, A. J., and Metzger, M. J. Digital media and youth: Unparalleled opportunity and unprecedented responsibility. *The John D. and Catherine T. MacArthur Foundation Series on Digital media and Learning*, pp. 5–27, Cambridge, MA: MIT Press, 2008.

[28] Flanagin, A. J., and Metzger, M. J. The perceived credibility of personal Web page information as influenced by the sex of the source. *Computers in Human Behavior*, 19(6), 683–701, 2003.

[29] Fogg, B. J., and Tseng, H. The elements of computer credibility. *In Proceedings of the SIGCHI conference on Human factors in computing systems: the CHI is the limit*, pp. 80–87, 1999.

[30] Fogg, B. J., Soohoo, C., Danielson, D. R., Marable, L., Stanford, J., and Tauber, E. R. How do users evaluate the credibility of web sites? A study with over 2,500 participants. *Proceedings of the 2003 Conference on Designing for User Experiences*, pp. 1–15, 2003.

[31] Fragale, A. R., and Heath, C. Evolving information credentials: The (mis) attribution of believable facts to credible sources. *Personality and Social Psychology Bulletin*, 30(2), 225–236, 2004.

[32] Goldberg, D., Nichols, D., Oki, B. M., and Terry, D. Using collaborative filtering to weave an information tapestry. *Communications of the ACM*, 35(12), 61–70, 1992.

[33] Gretzel, U., and Yoo, K. H. Use and impact of online travel reviews. *Information and communication technologies in tourism*, pp. 35–46, 2008.

[34] Gretzel, U., Yoo, K. H., and Purifoy, M. Online travel review study: Role and impact of online travel reviews. *Laboratory for Intelligent Systems in Tourism, Texas A&M University*, `www.tripadvisor.com/pdfs/OnlineTravelReviewReport.pdf`, 2007.

[35] Hill, W., Stead, L., Rosenstein M., and Furnas, G. Recommending and evaluating choices in a virtual community of use. *In Conference on Human Factors in Computing SystemsCHI 95*, Denver, May, 99. 194–201, 1995.

[36] Hirsch, J. E. An index to quantify an individual's scientific research output. *Proceedings of the National academy of Sciences of the United States of America*, 102(46), 16569–16572, 2005.

[37] Hovland, C. I., Janis, I. L., and Kelley, H. H. Communication and persuasion: Psychological studies of opinion change. *New Haven, CT: Yale University Press*, 1953.

[38] Hovland, C. I., Weiss, W. The influence of source credibility on communication effectiveness. *Public Opinion Quarterly*, 15, 635–650, 1951.

[39] Huang, Y., and Bian, L. A Bayesian network and analytic hierarchy process based personalized recommendations for tourist attractions over the internet. *Expert Systems with Applications*, 36(1), 933–943, 2009.

[40] Huang, Z., Zeng, D., and Chen, H. A Comparison of Collaborative-filtering recommendation algorithms for E-commerce. *IEEE Intelligent Systems*, pp. 68–78, 2007.

[41] Husain, W., and Dih, L. Y. A framework of a personalized location-based traveler recommendation system in mobile application. *International journal of multimedia and ubiquitous engineering*, 7(3), 11–18, 2012.

[42] Inversini, A., Marchiori, E., Dedekind, C., and Cantoni, L. Applying a conceptual framework to analyze online reputation of tourism destinations. *Information and Communication Technologies in Tourism*, pp. 321–332, 2010.

[43] Jacoby, J., Jaccard, J. J., Currim, I., Kuss, A., Ansari, A., and Troutman, T. Tracing the impact of item-by-item information accessing on uncertainty reduction. *Journal of Consumer Research*, pp. 291–303, 1994.

[44] Jeacle, I., and Carter, C. In TripAdvisor we trust: Rankings, calculative regimes and abstract systems. *Accounting, Organizations and Society*, 36(4), 293–309, 2011.

[45] Jensen, C., Davis, J., and Farnham, S. Finding others online: reputation systems for social online spaces. *In Proceedings of the SIGCHI conference on Human factors in computing systems*, pp. 447–454, 2002.

[46] Kamps, J., and Marx, M. Words with attitude. *In Proceedings of the 1st International Conference on Global WordNet*, pp. 332–341, 2002.

[47] Kang, H., Yoo, S. J., and Han, D. Senti-lexicon and improved Naive Bayes algorithms for sentiment analysis of restaurant reviews. *Expert Systems with Applications*, 39(5), 6000–6010, 2012.

[48] Kim, S. M., and Hovy, E. Determining the sentiment of opinions. *Proceedings of the 20th International Conference on Computational Linguistics*, 2004.

[49] Kiousis, S. Public trust or mistrust? Perceptions of media credibility in the information age. *Mass Communication & Society*, 4, 381–403, 2001.

[50] Konstan, J., Miller, B., Maltz, D., Herlocker, J., Gordon, L., and Riedl, J. GroupLens: applying collaborative filtering to usenet news. *Communications of the ACM*, 40(3), 77–87, 1997.

[51] Kusumasondjaja, S., and Shanka, T. Credibility of online reviews and initial trust: the Roles of reviewer's identity and review valence. *Journal of Vacation Marketing*, 18(3), 185–195, 2012.

[52] Lawrence, R. D., Almasi, G. S, Kotlyar, V., Viveros, M. S., and Duri, S. S. Personalization of supermarket product recommendations. *Data Mining and Knowledge Discovery*, pp. 11–32, 2001.

[53] Lee, H., and Law R., Murphy, J. Helpful reviewers in TripAdvisor: An online travel community. *Journal of Travel & Tourism Marketing*, 28(7), 675–88, 2011.

[54] Lee, J., Kim, J., and Moon, J. Y. What makes Internet users visit cyber stores again? Key design factors for customer loyalty. *In Proceedings of the SIGCHI conference on Human factors in computing systems*, pp. 305–312, 2000.

[55] Lee, W.S. Collaborative Learning for Recommender Systems. *Int'l Conf. Machine Learning*, pp. 314–321, 2001.

[56] Leung, C. W. K., Chan, S. C. F., and Chung, F. L. Integrating collaborative filtering and sentiment analysis: A rating inference approach. *Proceedings of the ECAI 2006 Workshop on Recommender Systems, in conjunction with the 17th European Conference on Artificial Intelligence*, pp. 62–66, 2006.

[57] Leung, C. W. K., and Chan, S. C. F. Sentiment analysis of product reviews. *J. Wang, (Eds.), Encyclopedia of data warehousing and mining-Second Edition, Information Science Reference*, pp. 1794–1799, 2008.

[58] Leung, C. W. K., Chan, S. C. F., Chung, F.L., and Ngai, G. A probabilistic rating inference framework for mining user preferences from reviews. *World Wide Web*, 14(2), 187–215, 2011.

[59] Linden, G., Smith, B., and York, J. Amazon.com recommendations: Item-to-item collaborative filtering. *IEEE Internet Computing*, 7(1), 76–80, 2003.

[60] Litvin, S. W., Goldsmith, R. E., and Pan, B. Electronic word-of-mouth in hospitality and tourism management. *Tourism Management*, 29(3), 458–468, 2008.

[61] Liu, B. Sentiment analysis and subjectivity. *Handbook of Natural Language Processing, Second Edition, (editors: N. Indurkhya and F. J. Damerau)*, pp. 627–666, 2010.

[62] Liu, B., Hu, M., and Cheng, J. Opinion observer: analyzing and comparing opinions on the web. *In Proceedings of the 14th international conference on World Wide Web*, pp. 342–351, 2005.

[63] Loda, M. D. Comparing web sites: An experiment in online tourism marketing. *International Journal of Business and Social Science*, 2(22), 70–78, 2011.

[64] Lorenzi, S., Saldana, F., Saldana, R., and Licthnow, D. A tourism recommender system based on collaboration and text analysis. *Information Technology and Tourism*, 6(3), 157–165, 2004.

[65] MacQueen, J. B. Some methods for classification and analysis of multivariate observations. *Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability*, pp. 281–297, 1967.

[66] Maltz, D., and Ehrlich, K. Pointing the way: Active collaborative filtering. *In Proceedings of ACM CHI95 Conference on Human Factors in Computing Systems*, pp. 202–209, 1995.

[67] Mayzlin, D., Dover, Y., and Chevalier, J. A. Promotional reviews: An empirical investigation of online review manipulation. *National Bureau of Economic Research*, 2012.

[68] Melville, P., Gryc, W., and Lawrence, R. Sentiment analysis of blogs by combining lexical knowledge with text classification. *In Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 1275–1284, 2009.

[69] Metzger, M. J. Making sense of credibility on the web: Models for evaluating online information and recommendations for future research. *Journal of the American Society for Information Science and Technology*, 58(13), 2078–2091, 2007.

[70] Metzger, M. J., Flanagin, A. J., and Medders, R. Social and heuristic approaches to credibility evaluation online. *Journal of Communication*, 60(3), 413–439, 2010.

[71] Metzger, M. J., Flangin, A. J., Eyal, K., Lemus, D. R., and McCann, R. M. Credibility for the 21st century: integrating perspectives on source, message, and media credibility in the contemporary media environment. *Communication Yearbook*, 27, 293–335, 2003.

[72] Miguens, J., Baggio, R., and Costa, C. Social media and tourism destinations: TripAdvisor case study. *Advances in Tourism Research*, 2008.

[73] Miller, B. N., AlbertI., Lam, S. K., Konstan, J. A., and Riedl, J. MovieLens unplugged: experiences with an occasionally connected recommender systems. *In Proceedings of Intelligent User Interfaces*, pp. 263–266, 2003.

[74] Montaner, M., Lopez, B., and de la Rosa, J. L. A taxonomy of recommender agents on the internet. *In Submitted to Artificial Intelligence Review*,19(4), 285–330, 2002.

[75] Mooney, R. J., and Roy, L. Content-based book recommending using learning for text categorization. *In Proceedings of the fifth ACM conference on Digital libraries*, pp. 195–204, 2000.

[76] Morinaga, S., Yamanishi, K., Tateishi, K., and Fukushima, T. Mining product reputations on the web. *In Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 341–349, 2002.

[77] Newman, M. E. Power laws, Pareto distributions and Zipf's law. *Contemporary physics*, 46(5), 323–351, 2005.

[78] O'Connor, P. User-generated content and travel: A case study on Tripadvisor. com. *Information and communication technologies in tourism*, pp. 47–58, 2008.

[79] Ohana, B., and Tierney, B. Sentiment classification of reviews using SentiWordNet. *In 9th IT and T Conference*, pp. 13, 2009.

[80] Okanohara, D., and Tsujii, J. Assigning polarity scores to reviews using machine learning techniques. *In Proceedings of the Second International Joint Conference on Natural Language Processing*, pp. 314–325,2005.

[81] Pang, B., and Lee, L. A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. *In Proceedings of the 42nd Annual Meeting of the Association for Computation Linguistics*, pp. 271–278, 2004.

[82] Pang, B., and Lee, L. Opinion mining and sentiment analysis. *Foundations and trends in information retrieval*, 2(1-2), 1–135, 2008.

[83] Pang, B., and Lee, L. Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. *In Proceedings of the Association for Computational Linguistics (ACL)*, pp. 115–124, 2005.

[84] Pang, B., Lee, L., and Vaithyanathan, S. Thumbs up?: sentiment classification using machine learning techniques. *In Proceedings of the ACL02 conference on Empirical methods in natural language processing Volume 10, Association for Computational Linguistics*, pp. 79–86, 2002.

[85] Pazzani, M., and Billsus, D. Content-based recommendation systems. *The Adaptive Web: Methods and Strategies of Web Personalization*, pp. 325–341, 2007.

[86] Pazzani, M., and Billsus, D. Learning and Revising User Profiles: The Identification of Interesting Web Sites. *Machine Learning*, 27, 313–331, 1997.

[87] Pazzani, M. J. A Framework for collaborative, content-based and demographic filtering. *Artificial Intelligence Review*,

[88] Pornpitakpan, C. The persuasiveness of source credibility: A critical review of five decades' evidence. *Journal of Applied Social Psychology*, 34(2), 243–281, 2004.

[89] Rabanser, U., and Ricci, F. Recommender systems: do they have a viable business model in e-tourism? *Information and Communication Technologies in Tourism*, pp. 160–171, 2005.

[90] Read, J. Using emoticons to reduce dependency in machine learning techniques for sentiment classification. *In Proceedings of the ACL Student Research Workshop*, pp. 43–48, 2005.

[91] Redner, S. How popular is your paper? An empirical study of the citation distribution. *The European Physical Journal B-Condensed Matter and Complex Systems*, 4(2), 131–134, 1998.

[92] Resnick, P., and Varian, H. Recommender systems. *Communications of the ACM*, 40(3), 56–58, 1997.

[93] Resnick, P., Iacovou, N., Suchak, M., Bergstorm, P., and Riedl, J. GroupLens: An open architecture for collaborative filtering of Netnews. *In Proceedings of ACM 1994 Conference on Computer Supported Cooperative Work*, ACM, pp. 175–186, 1994.

[94] Ricci, F. Travel recommender systems. *IEEE Intelligent Systems*, 17(6), 55–57, 2002.

[95] Ricci, F., and Del, F. Missier: Supporting Travel Decision Making through Personalized Recommendation. *In C-M Karat, J. Blom, and J. Karat (eds.), Designing Personalized User Experiences for eCommerce*, Kluwer Academic Publisher, pp. 221–251, 2004.

[96] Ricci, F., and Nguyen, Q. N. Critique-based mobile recommender systems. *OEGAI Journal*, 24(4), 2005.

[97] Rieh, S. Y. Judgment of information quality and cognitive authority in the web. *Journal of the American Society for Information Science and Technology*, 53(2), 145–161, 2002.

[98] Rieh, S. Y.,and Danielson, D. R. Credibility: A multidisciplinary framework. *Annual Review of Information Science and Technology*, 41(1), 307–364, 2007.

[99] Robin, B. Hybrid recommender systems: Survey and experiments. *User Modeling and User-Adapted Interaction*, 12, 331–370, 2001.

[100] Sarwar, B., Karypis, G., Konstan, J., and Riedl, J. Analysis of recommendation algorithms for e-commerce. *In Proceedings of the 2nd ACM conference on Electronic Commerce*, pp. 158–167, 2000.

[101] Sarwar, B., Karypis, G., Konstan, J., and Riedl, J. Application of dimensionality reduction in recommender systems a case study. *In Proceedings of the ACM WebKDD Workshop, in conjunction with the ACM-SIGKDD Conference on Knowledge Discovery in Databases*, 2000.

[102] Sarwar, B., Karypis, G., Konstan, J., and Riedl, J. Item-based collaborative filtering recommendation algorithms. *In Proceedings of the 10th International World Wide Web Conference*, pp. 285–295, 2001.

[103] Savolainen, R. Judging the quality and credibility of information in Internet discussion forums. *Journal of the American Society for Information Science and Technology*, 62(7), 1243–1256, 2011.

[104] Schafer, J. B., Frankowski, D., Herlocker, J., and Sen, S. Collaborative filtering recommender systems. *The Adaptive Web*, pp. 291–324, 2007.

[105] Schafer, J. B., Konstan J. A., and Riedl, J. E-Commerce recommendation applications. *Data Mining and Knowledge Discovery*, pp. 115–153, 2001.

[106] Shardanand, U., and Maes, P. Social information filtering: Algorithms for automating 'Word of Mouth'. *In Proceedings of ACM CHI95 Conference on Human Factors in Computing Systems*, pp. 210–217, 1995.

[107] Sharma, A., and Dey, S. A comparative study of feature selection and machine learning techniques for sentiment analysis. *In Proceedings of the 2012 ACM Research in Applied Computation Symposium* pp. 1–7, 2012.

[108] Sidali, K. L., Schulze, H., and Spiller, A. The impact of online reviews on the choice of holiday accommodations. *Information and communication technologies*, pp. 87–98, New York, NY: Springer Wien, 2009.

[109] Smyth, B., and Cotter, P. A personalized TV listings service for the digital TV age. *Knowledge-Based Systems*, 13, 53–59, 2000.

[110] Somasundaran, S., Ruppenhofer, J., and Wiebe, J. Detecting arguing and sentiment in meetings. *In Proceedings of the SIGdial Workshop on Discourse and Dialogue*, 6, 2007.

[111] Sparks, B. A., and Browning, V. The impact of online reviews on hotel booking intentions and perception of trust. *Tourism Managemen*, 32(6), 1310–1323, 2011.

[112] Steven, M. TripAdvisor Under Fire for 'Real Traveller' Contribution Claim. *The Guardian*, http://www.guardian.co.uk/media/2012/feb/01/tripadvisorcriticise-honest-contribution-claim, 2012.

[113] Sundar, S. S., and Nass, C. Conceptualizing sources in online news. *Journal of Communication*, 51(1), 52, 2001.

[114] Suomi, R., and Li, H. Internet adoption in tourism industry in China. *In Towards Sustainable Society on Ubiquitous Networks*, pp. 197–208, 2008.

[115] Tan, S., and Zhang, J. An empirical study of sentiment analysis for Chinese documents. *Expert Systems with Applications*, 34(4), 2622–2629, 2008.

[116] TripAdvisor.com. Fact Sheet. Accessed online (September 5, 2007) at: http://www.tripadvisor.com/pressCenter-c4-Fact_Sheet.html, 2007.

[117] Turney, P. D. Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews. *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)*, PP. 417–424, 2002.

[118] Vermeulen, I. E., and Seegers, D. Tried and Tested: The impact of online hotel reviews on consumer consideration. *Tourism Management*, 30(1), 123–127, 2009.

[119] Wang, Y., Ye, Y., Li, X., Ng, M. K., and Huang, J. Hierarchical information-theoretic co-clustering for high dimensional data. *International Journal of Innovative Computing, Information and Control*, 7(2), 487–500, 2011.

[120] Wietsma, R. T. A., and Ricci, F. Product reviews in mobile decision aid systems. *In Proceedings of the Conference on Information and Communication Technologies in Tourism (ENTER)*, pp. 15–18, 2005.

[121] Williams, A. M, and Balaz, V. Tourism Risk and Uncertainty: Theoretical Reflections. *Journal of Travel Research*, 0047287514523334, 2004.

[122] Wu, X., Kumar, V., Quinlan, J., GhoshJ., Yang, Q., Motoda, H., McLachlan, G. J., Ng, A., Liu, B., and Yu,P. S., et al. Top 10 Algorithms in data mining. *Knowledge Information Systems*, 14(1), 1–37, 2008.

[123] Xia, R., Zong, C., and Li, S. Ensemble of feature sets and classification algorithms for sentiment classification. *Information Sciences*, 181(6), pp. 1138–1152, 2011.

[124] Xie, H., Miao, L., Kuo, P. J., and Lee, B. Y. Consumers' responses to ambivalent online hotel reviews: The role of perceived source credibility and pre-decisional disposition. *International Journal of Hospitality Management*, 30(1), 178–183, 2011.

[125] Ye, Q., Zhang, Z., and Law, R. Sentiment classification of online reviews to travel destinations by supervised machine learning approaches. *Expert Systems with Applications*, 36(3), pp. 6527–6535, 2009.

[126] Yoo, K. H., and Gretzel, U. Comparison of Deceptive and Truthful Travel Reviews. *In ENTER*, pp. 37–47, 2009.

[127] Yoo, K. H., Lee, K. S., and Gretzel, U. The role of source characteristics in eWOM: What makes online travel reviewers credible and likeable. *In Proceedings of the 14th International ENTER Conference in Ljubljana*, pp. 23–34, 2007.

[128] Yoo, K. H., Lee, Y., Gretzel, U., and Fesenmaier, D. R. Trust in travel-related consumer generated media. *Information and Communication Technologies in Tourism*, pp. 49–60. Springer, New York, NY, 2009.

[129] Zanker, M., Gordea, S., Jessenitschnig, M. and Schnabl, M. A hybrid similarity concept for browsing semi-structured product items. *2006 Lecture Notes in Computer Science*, 4082, 21–30, 2006.

[130] Zhao, Y. and Karypis, G., and Fayyad. Hierarchical clustering algorithms for document datasets. *Data Mining and Knowledge Discovery*, 10(2), 141–168, 2005.

[131] Zhao, Y., and Karypis, G. Evaluation of hierarchical clustering algorithms for document datasets. *In Data Mining and Knowledge Discovery*, pp. 515–524, 2002.