ON CLASSIFYING DISCUSSION THREADS USING TRAVEL INFORMATION GOAL-ORIENTED MODEL

San-Yih Hwang, Department of Information Management, National Sun Yat-sen University, Kaohsiung, Taiwan, syhwang@mis.nsysu.edu.tw

Roger H.L. Chiang, Department Operations, Business Analytics, and Information Systems, University of Cincinnati, Ohio, USA, roger.chiang@uc.edu

Yung-Lin Hsiao, Department of Information Management, National Sun Yat-sen University, Kaohsiung, Taiwan, man74115@gmail.com

Shanlin Chang, Department of Information Management, National Sun Yat-sen University, Kaohsiung, Taiwan, d004020002@student.nsysu.edu.tw

Abstract

We study how to recommend discussion threads in the tourism domain to meet visitors’ travel information needs. This research-in-progress paper reports the first stage of our research, namely classifying discussion threads into travel goals. We propose an information goal-oriented model, which consists of four goals: Initiation, Attraction, Accommodation, and Route planning, that can be characterized using nine features. Seven of these nine features can be quantified based on lexicons, and the other two can be measured using the named entity recognition technique. Three lexicons can be further enhanced using WordNet. We conduct an experiment in evaluating the impact of these features on goal classification with a data set collected from TripAdvisor.com, the world’s largest travelling website. The experimental results show that our approach generally has comparable or better performance than that of using purely lexical features, namely TF-IDF, for classification.

Keywords: Discussion thread classification, Information goal-oriented model, Text categorization, Ontology, Named entity recognition
1 INTRODUCTION

Information overload has become an imperative problem for today’s Internet users due to the exponentially increasing of information available on the Internet. Information retrieval and recommender systems are the two well-known approaches for tackling this problem. Information retrieval techniques, widely adopted by search engines, allow users to submit queries and receive the most relevant web pages. On the other hand, the aim of recommender systems is to recommend users those items that they are likely to be interested without explicitly expressing their needs. Collaborative and content-based filtering methods are two most prevailing techniques that help users filter out irrelevant information (Adomavicius & Tuzhilin 2005).

Tourism is generally regarded as an information-intensive industry (Sheldon 1997) and nowadays extensively promoted by governments in many countries. Traditionally, tourists rely on travel guides to help plan their trips. More recently, the tourism industry has started to adopt various information techniques to help tourists plan their trips, and many travelling websites are equipped with recommender systems to recommend accommodations (Saga et al. 2008), destinations (Delgado & Davidson 2002), and attractions (Ardissono et al. 2003). However, for many tourists, experiences and advices from their peers are often more useful than direct (attraction/accommodation) suggestions provided by recommender systems. We have embarked on a research project that aims to recommend user-generated content (UGC) such as discussion threads readily available at travelling websites to a trip planner. Despite the many researches that have already been conducted to recommending travel route planning and destinations, there is relatively fewer researches on how to recommend the tourist-generated discussion threads which are available on travelling websites such as TripAdvisor.com.

We study how to recommend discussion threads to tourists according to their travel information needs. In our recommendation scenario, when a tourist needs travel information, she will express the information need as a question together with a set of threads that she has browsed when searching tourist-generated content according to the need. Specifically, we aim to recommend previous discussion threads that can provide potentially useful information to address the tourist’s question. This problem may seem like the traditional Q&A search problem, which aims to rank a given thread by considering the content similarity between the thread and the target question, the timeliness of the thread, and the quality of the thread (Suryanto et al. 2009). However, we observe that tourists at different stages of trip planning may need different types of information. For example, tourists in their initial stage of trip planning often prefer the discussion threads that cover a wider scope, while in later stages such as finding attractions of a particular location, threads that provide more specific travel information will become more useful. Several studies have adopted a process view about travel information search and decision making (Bettman 1979; Correia 2002; Leiper 1990; Vogt & Fesenmaier 1998), and the decision process can be divided into various stages with distinct characteristics. We conjecture that by taking into account different stages of travel information search, the recommender systems will be able to provide more accurate recommendations. In this work, we consider the information need at each stage as a goal fulfilment task and identify features relevant to each goal.

In our recommendation framework, there is a database of discussion threads, each represented as a question followed by a sequence of replies. A tourist, when seeking travel information, may read a number of threads that meet his/her information need and post a question. Four perspectives, namely goal similarity, content similarity, timeliness and quality, are considered when determining the recommendation score of each unseen thread. Figure 1 illustrates our research framework. This paper presents the travel information goal-oriented model and two methods (lexicon-based and WordNet-enhanced) for implementing Goal Identification, shown by the double-lined rectangle in Figure 1. More specifically, we propose an information goal-oriented model that consists of four goals, which can be characterized by nine features, and the proposed methods, with much fewer features, are
shown to achieve better or comparable goal identification result when compared to the baseline method that employs purely lexical features, i.e., TF-IDF.

![Diagram of Recommending Discussion Threads to Tourists]

Figure 1. Architecture of Recommending Discussion Threads to Tourists

The remaining of this paper is organized as follows. Section 2 surveys the travel information search models and presents the information goal-oriented model. In Section 3, we discuss two proposed methods in classifying discussion threads. Sections 4 and 5 discuss the experimental results and identify future research directions, respectively.

2 INFORMATION GOAL-DIRECTED MODEL

There are several process frameworks proposed for travel information search with respect to travel decisions (Bieger & Laesser 2004; Correia 2002; Leiper 1990). For example, Bieger and Laesser (Bieger & Laesser 2004) propose a process framework that divides information sourcing process into three distinctive phases: the pre-decision, the decision, and the post-decision. In the pre-decision phase, people look for “the information processed for decision making before making any constraining decision on a key characteristic of a trip” (Bieger & Laesser 2004). The information need of the decision phase converges toward a focus of a specific destination. In the post-decision phase, people accumulate information for travel preparation after deciding on the destination. However, we observe that the information search and decision making in the travel planning is not a sequential process, yet a process with several feedback cycles. In addition, steps defined in the information search process models are based on their temporal order and may not have unique information genre. To further examine important information types for tourists when they engage in information search on the web, we review Maser and Weiermair’s work that reports a survey result about important information types for travel decision-making (Maser & Weiermair 1998). They identified 13 types of information and presented to individuals who were asked to evaluate each information type in a scale of 1 to 5. General information about the destination was found to be the most important, followed by information about climatic/weather conditions and information on price. The fourth type is location of accommodations.

As a result, we propose a travel information goal-oriented model that consists of four goals: Initiation, Attraction, Accommodation, and Route planning. The Initiation goal is to fulfil the information need of tourists who just begin to plan a trip. The Attraction and Accommodation goals occur when tourists are looking for information about attractions and accommodations, respectively. Although there are
other types of travel destinations, e.g., restaurants and shopping malls, our examination on travelling websites (e.g., TripAdvisor.com) reveals that attractions and accommodations are the two most popular goals. The Route Planning goal is about how to organize the whole trip in an appropriate way, and the quality of a trip depends heavily on its route planning. We argue that these four goals have no absolute temporal order. One may proceed to a goal from any other goal due to personal preference or change of decisions. In the following, we will articulate each goal and its corresponding features.

When planning a trip, the immediate question that comes to people’s mind is "Where to go?" For the Initiation goal, the tourist's information need is undirected, in the sense that they need more diversified types of information. Therefore, specific places could be less often mentioned. Another characteristic of this goal is less price terms because prices are applied to specific places such as hotels, attractions, and restaurants, which are yet to be decided in this goal. The information need of the attraction goal is more directed. For example, tourists might want to know detailed information about several specific attractions in or close to the visited country or city. Attractions could be scenic areas, national parks, recreation areas, or museum campus. The information need of the Attraction goal can be characterized by the high frequency of attraction terms. Another feature is the high frequency of weather terms.

There are plenty of factors that may affect the tourist's decision on accommodation, including price, location, service, facilities, and quality. In addition to accommodation names and price information, an informative discussion thread pertaining to the attraction goal may include quite a few accommodation-related nouns and adjectives (Abbasi, Chen, & Salem, 2008; Xia & Peng, 2009). The quality of the trip depends heavily on its route planning. Because the route planning deals with scheduling, time information usually show up in discussion threads related to route planning. These threads may also involve attraction names and accommodation names. The information need of the route planning goal can be characterized by the high frequencies of date/time term, price terms, transportation terms, accommodation terms and attraction terms in the relevant discussion threads.

In summary, our model includes nine features for these four information search goals, namely place terms (PLACE), attraction terms (ATTRACTION), accommodation terms (ACCOMMODATION), accommodation-related nouns (ACC_NOUN), accommodation-related adjectives (ACC_ADJ), weather terms (WEATHER), transportation terms (TRANSPORTATION), date/time terms (DATE/TIME), and price terms (PRICE), as shown in Figure 2. These nine features are measured by the term frequencies that occurred in the discussion threads.

3 GOAL IDENTIFICATION

3.1 The lexicon-based Method

We have identified nine features to classify discussion threads. In particular, each goal is associated with a classifier because a given discussion thread may meet multiple goals. As the name implies, the lexicon-based method forms a lexicon for each feature except for DATE/TIME and PRICE. A lexicon for a given feature contains a set of terms relevant to the feature. We set up Taiwan as the travel destination to build the model and conduct the experiment. We examined various sources to construct lexicons for Taiwan’s tourism domain. Specifically, the lexicons of PLACE and ATTRACTION are constructed by consulting the glossaries provided by Taiwan Geographic Name Information Systems (http://placesearch.moi.gov.tw/index_en.php) and Taiwan Tourism Bureau (http://eng.taiwan.net.tw/) respectively. With respect to the lexicon of ACCOMMODATION, we consider two sources. One is the articles we crawled from TripAdvisor.com which also contains the information about accommodations in Taiwan. The other one is Taiwan Tourism Bureau, which lists many hotels, hostels, B&B and so on. By combining terms collected from the two information sources, we build the accommodation lexicon. The lexicon of WEATHER and TRANSPORTATION are constructed by
consulting Taiwan Central Weather Bureau (http://www.cwb.gov.tw/eng/index.htm) and glossary of transportation in CIA World Fact Book (http://reliant.teknowledge.com/DAML/Transportation.owl) respectively. The ACC_ADJ and ACC_NOUN lexicons are built by surveying literatures across text mining and tourism about hotel industry (Blair-Goldensohn et al. 2008; Lau et al. 2005; Pekar 2008; Titov & McDonald 2008; Xia & Peng 2009), which list the terms that would be used when people describe their experience of accommodation service.

By calculating the term frequencies of the words in a discussion thread that appear in a feature’s lexicon, the lexicon-based method determines the weight of the feature for a thread. Note that when travelling destination changes (i.e., other than Taiwan), the lexicons of some features such as PLACE, ATTRACTION, ACCOMMODATION and TRANSPORTATION should be reconstructed by consulting other data sources. Table 1 presents the number of terms and examples of these seven lexicons.

<table>
<thead>
<tr>
<th>Feature’s Lexicon</th>
<th>Number of Terms</th>
<th>Example Terms</th>
</tr>
</thead>
<tbody>
<tr>
<td>PLACE</td>
<td>354</td>
<td>Taipei, Taichung, Tainan, Kaohsiung</td>
</tr>
<tr>
<td>ATTRACTION</td>
<td>434</td>
<td>Alishan, Kenting, Sun Moon Lake</td>
</tr>
<tr>
<td>ACCOMMODATION</td>
<td>2,562</td>
<td>Regent Taipei, Evergreen Laurel Hotel</td>
</tr>
<tr>
<td>ACC_NOUN</td>
<td>45</td>
<td>Comfortable, Clean, Decent, Helpful</td>
</tr>
<tr>
<td>ACC_ADJ</td>
<td>51</td>
<td>Room, Beds, Luggage, Bathroom</td>
</tr>
<tr>
<td>WEATHER</td>
<td>108</td>
<td>Sunrise, Cold, Clear sky, Rainy</td>
</tr>
<tr>
<td>TRANSPORTATION</td>
<td>405</td>
<td>Airplane, Railways, MRT, HSR</td>
</tr>
</tbody>
</table>

Table 1. Feature Lexicons in Travel Information Goal-Oriented Model

The terms for the remaining two features, namely DATE/TIME and PRICE are identified using named entity recognition (NER) technique. The NER tool we applied is the Stanford Named Entity Recognizer (Finkel et al. 2005). The software provides a general implementation of linear chain Conditional Random Field (CRF) sequence models, coupled with well-engineered feature extractors for NER. We adopt regular expressions for determining DATE/TIME and PRICE. Specifically, we develop thirteen regular expression rules about date/time format by consulting the Regular Expression
Library, the Internet's first Regular Expression Library (http://www.regxlib.com/). As for the monetary values, we also use numerical regular expression originally from Regular Expression Library by incorporating currency signs to identify the price terms.

As a result, each thread is represented as a 9-tuple, recording the weight (or term frequency) of each of the nine features. Human experts can read some threads and decide their pertaining goals, thereby generating a training data set. We then apply classification techniques on the training data set to build a classification model for each goal. The four classification models will be used to predict the goals of each un-classified thread. More detailed description about the data and the classification techniques will be provided in Section 4.

3.2 The WordNet-enhanced Method

On account of the language peculiarity of English, synonyms are quite common. The WordNet-enhanced method enhances the lexicon-based method by incorporating WordNet, the largest lexical database of English with profound synonyms (Miller 1995). In WordNet, each concept, called a synset, contains a number of synonyms and is linked to its hypernyms and hyponyms, forming an is-a concept hierarchy.

We use WordNet to convert three lexicons, namely ACC_ADJ, ACC_NOUN and TRANSPORTATION, into sets of synsets. All the other lexicons remain the same because they contain mostly proper nouns and their meanings are clear from the text. That is to say, each term in the three lexicons is first manually converted into their corresponding synset, and the term is replaced by the synset and its hypernyms and hyponyms in the new (semantically based) lexicon. Similarly, each sentence has to be converted into a set of synsets. To do so, we first extract nouns, verbs, adjectives and adverbs from sentences of each thread and apply word sense disambiguation techniques (Navigli 2009) to identify the synsets of these terms. Then instead of computing the term frequencies of each feature, the WordNet-enhanced method uses the synset frequency as the weight of the three features. The same approach as described for lexicon-based method is then followed.

4 EMPIRICAL EVALUATION

We have conducted an empirical evaluation by collecting data from TripAdvisor.com, the world's largest travel site in providing reviews and advices on travel-related activities. A Web crawler was developed to extract discussion threads about Taiwan from TripAdvisor.com posted between 1/18/2005 and 02/01/2011. The dataset contains 11,305 forum posts, 54,530 replies and 4,492 hotel reviews. To obtain the training data for classification, we recruited two PhD students who have the profound experience in traveling to label 600 randomly selected threads from our data collection. To observe the consistency between these two labelers, we measured the average Jaccard coefficient (Martin et al. 1995) of the four goals to measure the similarity of their labelling results. After reconciliation, the average Jaccard coefficient increases from 0.83 to 0.96, and we chose only these labels agreed by both labellers. The numbers of threads that were labelled as the goals of Initiation, Attraction, Accommodation, and Route Planning are 21, 71, 220, and 247, respectively.

To verify our conjecture on the pertaining features of each goal as described in Section 2, we performed the chi-square test to measure the correlation between the features and goals, and all the identified hypothesized feature-goal pairs except PLACE-Initiation are significant (i.e., with p-value < 0.05). Failure to show the significance of PLACE-Initiation could be due to the small number of instances labelled as the Initiation goal (21). The result is shown in Table 2.

We then performed a 10-fold cross-validation, by which the dataset is randomly split into ten equal-sized segments and each segment takes turns in serving as the test dataset. As our dataset is quite skewed, training data for each goal has to be carefully prepared. For Initiation and Attraction goals, however, there were insufficient positive examples; thus we employed oversampling and multi-
classifier (Batista et al. 2004) to reduce the negative effect caused by skewed data. We have exercised several classification methods and report the results of Naïve Bayes classification as it achieved the best results.

We have exercised several classification methods and report the results of Naïve Bayes classification as it achieved the best results.

<table>
<thead>
<tr>
<th>Feature</th>
<th>Goal</th>
<th>Initiation</th>
<th>Attraction</th>
<th>Accommodation</th>
<th>Route Planning</th>
</tr>
</thead>
<tbody>
<tr>
<td>PLACE</td>
<td></td>
<td>0.107</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ATTRACTION</td>
<td></td>
<td>0.030*</td>
<td></td>
<td>0.000**</td>
<td></td>
</tr>
<tr>
<td>ACCOMODATION</td>
<td></td>
<td></td>
<td>0.000**</td>
<td>0.000**</td>
<td></td>
</tr>
<tr>
<td>ACC NOUN</td>
<td></td>
<td></td>
<td>0.000**</td>
<td>0.000**</td>
<td></td>
</tr>
<tr>
<td>ACC ADJ</td>
<td></td>
<td></td>
<td>0.000**</td>
<td></td>
<td></td>
</tr>
<tr>
<td>WEATHER</td>
<td></td>
<td>0.050*</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>TRANSPORTATION</td>
<td></td>
<td></td>
<td></td>
<td>0.000**</td>
<td>0.000**</td>
</tr>
<tr>
<td>DATE/TIME</td>
<td></td>
<td></td>
<td></td>
<td>0.000**</td>
<td>0.000**</td>
</tr>
<tr>
<td>PRICE</td>
<td></td>
<td>0.000**</td>
<td>0.000**</td>
<td>0.000**</td>
<td></td>
</tr>
</tbody>
</table>

Table 2. p-vlaues of Features Related to Goals in Travel information goal-oriented model

In our experiments, we choose the well-known Term frequency–inverse document frequency (TF-IDF) as the baseline method. In the TF-IDF method, about 4000 terms that have the highest average TF-IDF values serve as the dimensions, and each thread is represented as a vector of TF-IDF values. Figure 3 shows the precision and recall of two proposed methods and TF-IDF method by applying multinomial Naïve Bayes for building classifiers. The two proposed methods have similar performances as the TF-IDF method for all goals except for Attraction goal. For Attraction goal, the proposed methods have significantly much better performance than TF-IDF. This is because the top several terms identified by TF-IDF method, such as "hotel", "stay", "room", "night", "rest", are incapable in distinguishing Attraction goal from other goals. In addition, despite our effort in remedying the data skewness problem for Initiation and Attraction goals, their relative performance is still inferior. In addition, from our preliminary experiment, we observe that oversampling and multi-classifier for Initiation and Attraction goals are indeed effective and gain almost ten 10% improvements in precision.

Figure 3. (a) Precision (b) Recall of the three methods using Multinomial Naïve Bayes

Overall, the performance of WordNet-enhanced method consistently performs better than the lexicon-based method in both precision and recall, although the difference is small. The experimental result suggests that by collecting comprehensive domain terms, a small set of features is capable of achieving comparable or even higher classification performance than the many lexical features employed by the pure lexical method (i.e., TF-IDF).
5 CONCLUSIONS

The work presented in this paper is the first step toward recommending discussion threads in the tourism domain. To enable effective recommendations, we need to identify the information search goal of a given tourist and then recommend threads pertaining to the goal. Therefore, in the initial research, we focused on identifying four goals in tourism and proposed two methods in classifying discussion threads into these four goals based on nine features. The lexicon-based method utilizes a set of lexicons about Taiwan tourism domain and NER technique to quantify these features. This method is subsequently enhanced using WordNet. We finally use the user-generated content collected from TripAdvisor.com to evaluate the proposed methods. The experimental results show that our approach, with much less features, generally has comparable or better performance than that of using many lexical features, namely TF-IDF, for classification.

References


