Location Relevance Classification for Travelogue Digests

Mao Ye[1†*], Rong Xiao[2‡], Wang-Chien Lee[1†], and Xing Xie[2‡]
[†]Department of Computer Science & Engineering, The Pennsylvania State University, PA, USA.
[‡]Microsoft Research Asia, Beijing, China.
[1]{mxy177,wlee}@cse.psu.edu [2]{rong,xing.xie}@microsoft.com

ABSTRACT

location relevance, we explore the *textual* (e.g., surrounding words) and *geographical* (e.g., geographical relationship among locations) features of locations to perform location relevance classification for theme location discovery. Finally, we conduct comprehensive experiments on collected travelogues to evaluate the performance of our location relevance classification technique and demonstrate the effectiveness of the travelogue service.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Filtering

General Terms

Experimentation

Keywords

classification, travelogue services

1. INTRODUCTION

With the advantages of Web 2.0 technology, many people are willing to share their *travelogues*, which record travel experiences, on weblogs, forums and social communities for travels. These travelogues usually contain rich travel information, such as the tours, lodging, meals, expenses, weather conditions, and so on, which are highly valuable to a trip planner. In this paper, we propose to develop automatic travelogue mining techniques to convey useful information in a travelogue to help trip planning.

> **Siesta Key** - It is a paradise found in **Sarasota, FL**, located a little over an hour drive south to *Tampa, FL*. I like to spend time there - mainly at beaches and bars ... We got back to *Boston* after one week vocation.

Figure 1: A paragraph in a travelogue for Siesta Key. Relevant locations are in bold, and irrelevant locations are in *italic*.

*This work was done when the author was visiting Microsoft Research Asia.

WWW 2011, March 28–April 1, 2011, Hyderabad, India.
ACM 978-1-4503-0637-9/11/03.

As a travelogue intends to record activities and experiences of its author at locations on a trip (i.e., the whereabout of the trip), a key problem is to find *theme locations*, i.e., locations appeared in a travelogue which are closely relevant to the main themes of the travelogue. However, due to the nature of human language presentation, locations mentioned by a document are not necessarily theme locations of the travelogue. For example, in Figure 1, Siesta Key is very relevant to the theme of this paragraph but Boston is not. Thus, there is a need for location relevance classification for theme location discovery.

2. LOCATION RELEVANCE CLASSIFICATION

Location relevance classification aims to discover the theme locations for a given travelogue. It is essential to identify useful features of a location that can help to assess whether the location is relevant to main themes of the travelogue or not. Since travelogues are textual documents recording the authors' experiences in touring places of interests, it naturally contains textual and geographical features.

One the one hand, for a travelogue, we usually have a title to summarize the theme of the travel. The most relevant locations are usually included there. What's more, interested locations might be mentioned several times in a travelogue, while irrelevant locations appear less frequently. According to this observation, we extract two kinds of features, namely, *is in title?* (denoted as F_1) and *number of appearance* (denoted as F_2), respectively. Besides, surrounding words of a location also provides useful hints. For example, in Figure 1, one may easily understand that "Siesta Key" is the most relevant location name in this paragraph, because important locations are usually heading the paragraphs. Thus for a location in a travelogue, we use its surrounding words to extract *bag-of-word* feature (denoted as F_3) and *syntactic pattern* feature (denoted as F_4) to describe the location. Those above four features are categorized as textual feature.

On the other hand, each location as a physical entity holds geographical properties in the real word. For example, location names referred in a travelogue usually have different location types in location partonomy. Location names with smaller scope usually hold higher relevance, otherwise, travelers would not bother to mention that small place. Thus, we consider the location type to be an important feature, named as *location type* (denoted as F_5). Additionally, relevant locations are usually clustered in the partonomy hierarchy of a location ontology. Moreover, people would likely to stay in the same state or the same city during the trip.

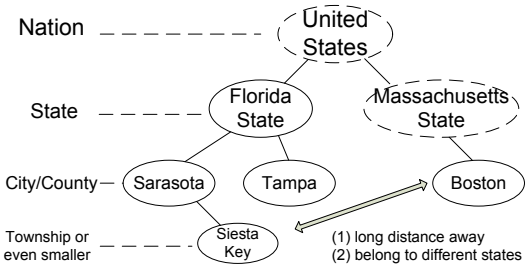

Figure 2: Partonomy hierarchy of locations mentioned in the travelogue shown in Figure 1

Therefore, we consider that partonomy distance among the locations provides important information to leverage relevant locations. More specifically, we extract the partonomy distance among locations as a geographical feature (denoted as F_6). Finally, if two locations are far away, e.g., Siesta Key and Boston in Figure 2, only one of them could mostly be the relevant location for a travelogue. Therefore, we consider geographical proximity between two different locations (mentioned in the same travelogue) to be another important geographical feature (denoted as F_7). The above three features are named geographical features.

3. PERFORMANCE EVALUATION

To evaluate the performance of location relevance classification and the effectiveness of proposed travelogue services, we collected approximately 100,000 travelogues in English (from travel web-site such as *TravelPod*[1], *IgoUgo*[2] and *TravelBlog*[3]) with location labels fallen in United States to form an English Corpus. Nevertheless, instead of relying on the location labels directly, we implemented a location extractor [2] to extract locations mentioned in these travelogues, yielding 18,000 unique locations. Because the task require evaluation by human beings with travel-related background and knowledge, we also built a Chinese Corpus by collecting travelogues from *Ctrip*[4], which consists of 94,000 Chinese travelogues related to around 32,000 locations in China.

In the first experiment, *Precision* and *Recall* are used to evaluate the performance location relevance classification. Note that, location classified as relevant are considered as the theme location to a travelogue. Thus, the performance of location relevance classification also demonstrate the effectiveness of travelogue digests of theme locations. To facilitate the study on impact of extracted features, we form the following three *feature groups*: (1) *baseline* group, containing features F_1 and F_2; (2) *textual* group - we consider all textual features; (3) *textual+geographical* group - we consider both the textual and geographical features. And we labeled 1,000 travelogues in the Chinese Corpus for our performance evaluation. Among those 1,000 labeled travelogues, we randomly select 100 travelogues as the test set and the remaining travelogues for training. Figure 3 shows the precision-recall curves of location relevance classification with different feature groups. We found that the baseline group (i.e., whether the location appears in the title and the count of locations in a travelogue) already provides strong support for deciding relevant locations. However, classifier using baseline features is effective only to some extent. With more textual features, the location relevance classifier gain some improvement. Finally, as shown in Figure 3, geographical features also enhance the performance of location relevance classification.

Next, we would like to evaluate whether location relevance classification can help improve the accuracy for geographical scope identification [1]. Since there is no existing dataset with labeled geographical scope. We use the travelogues labeled with theme locations to derive the ground truth and use labeled theme locations as inputs for our geographical scope computation algorithm. We introduce two performance metrics, namely, *average number of missing scopes* (i.e., the correct scopes not found) and *average number of wrong scopes* (i.e., the scopes found incorrectly) to evaluate the performance. We compare our approach which takes theme locations as input for geographical scope computation with a naive approach that considers all locations as input. As shown in Figure 4, our approach achieves a much better performance than the naive approach in terms of average number of wrong scopes, while remain to be competitive in terms of average number of missing scopes.

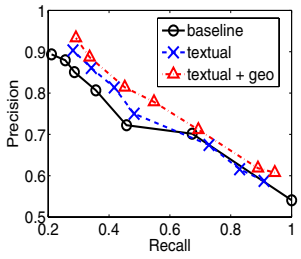

Figure 3: Location relevance classification

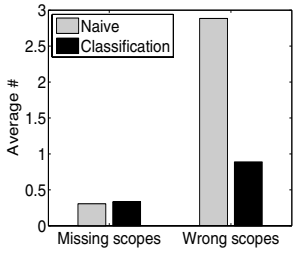

Figure 4: Geographical scope

4. CONCLUSIONS

In this paper, we study the problem of location relevance classification to support the travelogue service, which discovers and conveys travelogue digests such as theme locations and geographical scope to readers. Since travelogues are textual documents recording the authors' experiences in touring places of interests, we explore the *textual* and *geographical* features of locations to realize the location relevance classification. Finally, we conduct comprehensive experiments on collected travelogues to evaluate the performance of our location relevance classification technique and demonstrate the effectiveness of the travelogue service.

5. REFERENCES

[1] E. Amitay, N. Har'El, R. Sivan, and A. Soffer. Web-a-where: geotagging web content. In *SIGIR*, 2004.

[2] T. Qin, R. Xiao, L. Fang, X. Xie, and L. Zhang. An Efficient Location Extraction Algorithm by Leveraging Web Contextual Information. In *GIS*, 2010.

[1]http://www.travepod.com
[2]http://igougo.com
[3]http://www.travelblog.com
[4]http://www.ctrip.com