Neurocomputing 204 (2016) 61–69

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

Get into the spirit of a location by mining user-generated travelogues

Zhu Zhu *, Lidan Shou, Ke Chen

School of Computer Engineering, Zhejiang University, Hangzhou, China

ARTICLE INFO

Article history:
Received 14 March 2015
Received in revised form
24 April 2015
Accepted 25 April 2015
Available online 12 April 2016

Keywords:
Social media mining
User-generated travelogues
Geographic knowledge extraction

ABSTRACT

User-generated travelogues have contributed abundant location representative information (e.g. attracting scenic spots, activities, and local customs), which can greatly facilitate people in trip planning and destination understanding applications. Existing work on location information extraction from user-generated travelogues has primarily focused on the contents, while in this work, we resolve both the contents and structures of travelogues, as well as investigating the interplay of the two. We propose a two-part framework to mine location representative knowledge from travelogues. The first part resolves travelogues in a geographic view. It discovers real location entities subordinated to travel destinations, then decomposes travelogues to acquire corresponding descriptions for each of these entities. Built upon the results of the first part, the second part performs content resolution in a semantic view. By extracting location characterizing concepts and the relatedness among them, it can form a representative concept network for each location entity. Based on a large collection of travelogues, the proposed framework is evaluated using both objective and subjective evaluation methods and shows promising results.

1. Introduction

Travel has already become a fashionable lifestyle for today's human beings. While people are keen on traveling activities, the booming of Internet has encouraged them to share traveling experiences with great passion as well [1]. More and more people record and share traveling experiences by posting travelogues on weblogs, forums or travel communities. Travelogue is mainly composed of free texts describing the author's tourism experience and relevant photos taken during the trips. The content of a travelogue can cover various aspects of a tourism, such as the scene in visited landmarks, interesting things encountered, local custom of the destination, and traveler's personal impression on a place. Offered by experienced travelers, these user-generated contents (UGC) [2] provide fresh information and authentic feedback of travel destinations, which can serve as a powerful reference for users who plan their trips or need to understand locations [3].

We aim to get a deep insight into destinations by mining location representative knowledge from user-generated travelogues. A stream of work has been dedicated to extracting collaborative geographic knowledge from travel-related documents. For example, Wang et al. [4] proposed a probabilistic graphical model to learn the relationship between locations and words from online news and blogs. Hao et al. [5] discovered global and local topics in travelogues, and adopted global topics to generate location representative tags or destination summaries. Pang et al. [6] utilized textual information in user-generated travelogues and visual information in the Web photos to generate comprehensive location overviews. These existing work has focused primarily on content understanding of travelogues. In contrast, our goal is to investigate both the content and the structure of travelogues, as well as the interplay between them.

We will explore travelogues from two different aspects of view.

- *The geographic view*: We will discover finer-grained location entities subordinated to travelogue destinations and retrieve corresponding descriptions of these locations from travelogue text.
- *The semantic view*: We will mine location representative knowledge for each location entity acquired in the geographic view. By describing the characteristics of locations, the knowledge is presented as a concept network composing of the representative concepts and the semantic correlations among these concepts.

The combined result of both views can provide an effective way to thoroughly outline a given destination. Besides, by resolving travelogues from the two above views, we capture not only the relations of geographic entities to textual contents, but also the semantic correlations among mined knowledge components. Our work thus can be utilized not only for the keyword-based location retrieval and recommendation applications, such as helping user to find interesting

* Corresponding author.
E-mail addresses: zzbkynl@163.com (Z. Zhu), should@zju.edu.cn (L. Shou), chenk@zju.edu.cn (K. Chen).

http://dx.doi.org/10.1016/j.neucom.2015.04.129
0925-2312/

destinations, but also to support the location-based cross-media applications, such as enriching the tag annotation of travel photos.

Motivated by the above intuitions, we formulate a framework consisting of two modules. The first is a geographic segmentation process which decomposes the travelogue documents into series of snippets according to the occurrence of real location entities. Real Location Entity (*RLE*) is defined as location entity that has been mentioned in the text of travelogue and is subordinated to travelogue destinations. To identify *RLE*s, a location name detection process is performed on travelogues, then an iterative method combining statistical learning and rule extraction is proposed for location entity disambiguation. By discovering *RLE*s, we can obtain the partonomy structure of a travel destination, and the corresponding descriptions for each *RLE* through the segmentation process.

The second part of our framework is the mining of representative concept network for each *RLE* extracted in the first part. To generate descriptive concepts with integral semantic meaning and clear expression, we utilize natural language parsing techniques to acquire the lexical dependencies between words. Then a graph-based method is proposed to extract descriptive concepts on the linkage among words and the semantic relatedness of these concepts. Finally, to highlight those location representative concepts, a term weighting function is presented.

To summarize, The main contributions of this work are demonstrated as follows:

- We propose a location representative knowledge mining framework which explores user-generated travelogues from both the content and structural aspects. A geographic view and a semantic view are respectively formed and combined to provide an outline for the characteristics of destinations.
- We develop geographic segmentation techniques to decompose travelogues by the correspondence between identified location entities and their descriptions.
- When identifying real location entities, an iterative algorithm is proposed to resolve the ambiguity of the names.
- A graph-based method is presented to generate descriptive concepts and concept networks by utilizing the lexical dependencies between travelogue words.

We describe the detailed solution in Section 2, while Section 3 conducts performance evaluation to validate the proposed framework and algorithms. Section 4 introduces the most related literatures. Finally, Section 5 concludes the paper and discusses the future work.

2. Solution

Our solution is a framework consisting of two main parts. The first part is Geographic Segmentation, in which the travelogue is divided into small textual segments. Each segment describes one or more Real Location Entities (*RLE*) in the travelogue. The second part, which is called Construction of Concept Network, mines geo-informative knowledge from the textual segments obtained for each *RLE* in the first step. The knowledge is represented as a network which characterizes the features of a location with narrative concepts. The semantic correlations among different concepts are also presented. Hence, for each geographic location, we will output a graph-based location concept network, which can then be used for various travel-related applications.

2.1. Geographic segmentation of travelogues

To extract the geographic information on fine-grained levels, we have to refine the raw travelogue to small snippets at first. We call this process Geographic Segmentation. The basis of the segmentation is Real Location Entities, which is described as follows.

Real Location Entity: A Real Location Entity (*RLE*) of a travelogue is a distinct landmark which is a reached place in the tourism. Each *RLE* is not only distinct in the travelogue, but also represents a distinct landmark in real world.

The definition of *RLE* implies some issues: (1) One travelogue includes one or more *RLE*s, the author usually describes the *RLE* he/she went separately in the article sequentially. (2) Different travelogues may have intersecting *RLE*s. (3) There are no inclusion between *RLE*s because we define it as the finest granularity in our zone.

Because of the above characteristics, we make *RLE* as the unit by which the snippets are organized. Ideally each snippet would be associated to one *RLE*. However, due to the diversity of personal writing style, there might be multiple *RLE*s in one sentence. In this case we map the sentence into all the *RLE* included because a finer-grained segmentation is too expensive.

Geographic segmentation: Given a travelogue document d with a list of $RLE_i (i = 0, 1, \ldots, n)$. we divide d into a set of snippets $C_j (j = 0, 1, 2 \ldots, n)$. Each C_j is associated to at least one *RLE* and cannot be further segmented.

The key problem here is to identify the *RLE* in s. The *Location Names* mentioned in travelogue need to be identified and mapped into *RLE* correctly. There are two kinds of main interference here. The first is "Dummy Reference", that means some location names are not *RLE* of this travelogue. For example, The sentence "The magnificent scene here evokes my memories in Greenland" includes one place names "Greenland", which is obviously not a *RLE*. The other interfere is "Named Duplication", as distinct places may have same names. For example when one travelogue mentions the landmark "Rainbow Peak", it could be the Montana/Flathead/Rainbow Peak or the Idaho/Valley/Rainbow Peak or the other twos (see Fig. 1).

We propose a three-step method for geographic segmentation of travelogues. Firstly, we extract the location names from the travelogue. Then a disambiguation process combining statistical learning and rule extraction is performed on these candidate names, producing the *RLE*s. Finally, we segment the travelogue according to the distinguished *RLE*s. We will describe the details of the three steps in the following subsections.

2.1.1. Toponym detection

In this step, we aim to detect location names occurred in travelogue documents. We need to detect multi-granularity toponym that ranges from country-level to landmark-level. The method we used is partly based on work [7], which employs gazetteer for identification of location names. The basic idea is to scan text for occurrences of names appearing in the gazetteer, and examine the spots by calculating a confidence score for each candidates. The gazetteer we used is the Geonames database. We scan the text with a 8-words window. When calculating confidence scores, we take the surrounding context of candidate spots into consideration, as some linguistic patterns are

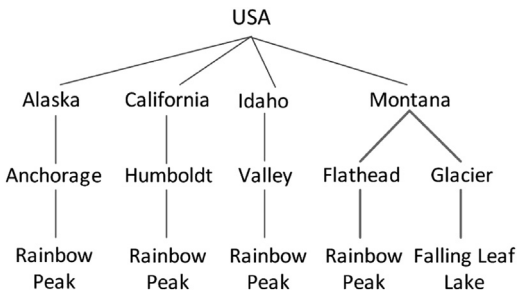

Fig. 1. An example of partonomy.