International Journal of Information Management xxx (2016) xxx–xxx

Contents lists available at ScienceDirect

International Journal of Information Management

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

Make your travel smarter: Summarizing urban tourism information from massive blog data

Hua Yuan [a,*], Hualin Xu [a], Yu Qian [a], Yan Li [b]

[a] School of Management and Economics, University of Electronic Science and Technology of China, Chengdu 611731, China
[b] School of Management, China University of Mining & Technology (Beijing), Beijing 100083, China

ARTICLE INFO

Article history:
Received 31 May 2015
Received in revised form 27 January 2016
Accepted 23 February 2016
Available online xxx

Keywords:
Blog mining
Geographic term
Tourist location
Word network
Travel route

ABSTRACT

In this work, we propose a research framework to help people summarize tourism information, such as popular tourist locations as well as their travel sequences (routes), for a previously unknown city from massive travel blog with the objective of providing users with better travel scheduling. To do this, we first crawl the massive travel blogs for a targeted city online. Then, we transfer the textual contents of these blogs to a series of word vectors to form the initial data source. Next, we implement the frequent pattern mining method on the data to identify the city's popular locations by their sequenced co-occurrences among the usual tourism activities, which can be visualized into a word network. Finally, we develop a max-confidence based method to detect travel routes from the network. We illustrate the benefits of this approach by applying it to the data from a blog web-site run by a Chinese online tourism service company. The results show that the proposed method can efficiently explore the popular travel information from massive data.

1. Introduction

With the prosperity of tourism and Web 2.0 technologies, most tourism websites such as tripadvisor.com and ctrip.com enable consumers to exchange information and opinions about tourism destinations, products and services in the form of blog (Pan, MacLaurin, & Crotts, 2007). This has led to a massive archive of travel blogs that contain valuable information from which potential tourists are likely to learn efficient travel experiences.

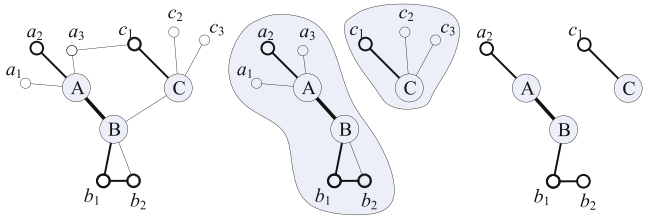
However, Web 2.0 technology enables people with various background to become online writers. As a result, the task of exploring information from online documents may become more difficult when faced with the phenomenon of so-called *information overload* (Park & Lee, 2008) and *data sparsity* (Popat, Balamurali, Bhattacharyya, & Haffari, 2013), referred to the difficulty a person will have in understanding an issue and making decisions in the presence of overly abundant and various information.

In recent times, much attention has been drawn to the great needs for research of smart tourism (Gretzel, Sigala, Xiang, & Koo, 2015). Obviously, some new technologies, such as data mining (Olmeda & Sheldon, 2002), can be useful for extracting information from massive online contents automatically and then for helping people make more efficient and economic travel decisions (Majid et al., 2013). However, literature reviews indicate that research so far has focused on information about tourist sites mined from online multimedia, such as travel photos (Lu, Wang, Yang, Pang, & Zhang, 2010; Vu, Li, Law, & Ye, 2015), Check-in data (Lu, Chen, & Tseng, 2012), and GPS data (Zheng, Zhang, Xie, & Ma, 2009) rather than textual contents. This is largely because the natural language processing (NLP) method is involved in text mining and the correlation analysis knowledge in NLP is typically complex. For example, the documents in traditional NLP are usually represented by a vector space model (Salton, Wong, & Yang, 1975), in which, the position information of words in documents is lost (Soucy & Mineau, 2005; Turney & Pantel, 2010), and all the terms in a document are assumed statistically independent (Robertson, 2004). However, not all the words in a blog are necessarily independent variables (Yang & Wilbur, 1996). Therefore, a lot of noisy results would be generated while the traditional NLP methods are introduced directly in blog mining, especially in terms of correlation analysis.

Inspired by the formation of blog contents, we notice that tourism blogs (in paragraph) are always organized according to a topic, for example, a tourist site. If a tourism topic is discussed in blog text, a set of local things having strong correlation with the topic may also be mentioned and recorded. For instance, if people talk tourism about *New York*, then hot tourism areas such as *Manhattan* could be mentioned with a very high probability. Here, we

* Corresponding author.
E-mail addresses: yuanhua@uestc.edu.cn (H. Yuan), bruce123.xu@gmail.com (H. Xu), qiany@uestc.edu.cn (Y. Qian), liyan@cumtb.edu.cn (Y. Li).

http://dx.doi.org/10.1016/j.ijinfomgt.2016.02.009
0268-4012/



(a) A location network for a city. (b) Partition of tourism areas. (c) Popular routes in each area.

Fig. 1. An example of tourism areas and routs mining from word network.

call *Manhattan* as a local feature of topic *New York*. Further, if *Manhattan* becomes the topic of a piece of blog, then *Wall Street* may be involved as a hot local feature associated to it. Along this line, there may exist some spacial semantic hierarchies among massive tourism terms (Kuipers, 2000), in which, a higher level thing always acts as a topic in blog contents while the lower level things associated to it may recorded nearby in the same contents as its local features. In addition, those locations or features close to each other geographically can be clustered into a *tourism area* (Ye, Xiao, Lee, & Xie, 2011).

However, the distributions of topic words as well as their local features in blogs, which are contributed by ordinary users, are not well-organized as they should be according to their appropriate semantic relations. Therefore, a tourism blog summarization system could be faced with **two technical challenges**: first, how to mine the true topic(s) from a piece of blog text efficiently; second, how to identify the true local features for each mined topic, i.e., identify the master–slave (dependency) relationships between pairs of words.

In this paper, we propose a research framework that can summarize the valuable information like popular locations, tourism areas (clusters of popular locations), and trip routes (travel sequences of popular locations), about a specific city from massive archives of tourist blogs. There are three main steps in this process:

- First, the frequent pattern mining method is introduced to mine a city's tourism profile to help potential tourists understand the tourism information for the city as a whole. The main contents of the profile are the *popular sites* (i.e., the sites mentioned frequently by independent bloggers at a frequency bigger than a given threshold) as well as their correlations (i.e., the geographical relation between two locations that had been visited by tourists successively). In addition, the correlations are visualized as a network (See Fig. 1(a)).
- Then, a max-confidence based method is proposed to partition all the attractions into groups to form a set of *tourism areas* according to their relative geographical position. The term of a *tourism area* refers to an area in which a few geographically closer locations are contained (In Fig. 1(b), if sites "A" and "B" are frequently mentioned simultaneously, then they may be close to each other in geography, and therefore, they can be grouped into the same tourism area).
- Finally, the popular travel sequences of the sites within an area, as well as their popular local features are extracted. For example, "a_2" of location "A", "b_1, b_2" of location "B", and "c_1" of location "C" in Fig. 1(c). This information would help potential tourists make better decisions in travel route planning, especially, when the potential tourist is limited by time, stamina (physical strength), or both.

The summary and data visualizations will potentially supplement the existing online information so that a user will have enriched understanding of the interconnections among the places of interest. The rest of this paper is organized as follows. Section 2 presents the related work. Section 3 sketches out the research framework. Section 4 shows the experimental results. Section 5 discusses some problems about the method. This paper is concluded in Section 6.

2. Related work

2.1. Blog mining

In blog mining, we note that the basic technology used in online text processing is text mining (Cao, Duan, & Gan, 2011; Ghose & Ipeirotis, 2011), in which three main tasks are involved: feature extraction, words correlation analysis and text visualization.

- Feature extraction

In a general text mining task, a text or document is always represented as a bag of words (BOW) (Gabrilovich & Markovitch, 2005), is used to derive insights from user-generated contents. BOW primarily originated in the computer science literature (Hu & Liu, 2004; Pang & Lee, 2008). The BOW representation raises two problems: (1) the feature space has a high dimensionality, and (2) not all the words in a document can be used as the key features.

In literature, feature selection is an important technology used to deal with these problems (Liu, Liu, Chen, & Ma, 2003). It is a process that chooses a subset from the original feature set according to certain criteria. The selected features should retain the original meaning and provide a better (usually more efficient and precise) understanding of the data. There are lots of feature selection metrics that assess the representativeness or importance of different document features, as exemplified by term frequency (TF), term frequency and inverse document frequency (TF-IDF), document frequency (DF), information gain (IG), mutual information (MI), chi-square statistic (χ^2), and term strength (TS) (Boley et al., 1999; Pantel & Lin, 2002; Roussinov & Chen, 1999). Both the strength and the weakness of these metrics are summarized in the following Table 1.

The theoretical basis of these six methods is sound, but their performance on feature selection is different. Both χ^2 and IG often achieve better accuracy than MI and TF-IDF (Yang & Pedersen, 1997). However, the IG is suspected having poor performance on skewed text corpora (Mladenic & Grobelnik, 1999). It has been seen that TF is an efficient and simple algorithm for matching words to documents. Furthermore, encoding TF is straightforward, making it ideal for forming the basis for more complicated algorithms and query retrieval systems (Berger, Caruana, Cohn, Freitag, & Mittal,

Table 1
Metrics used in feature selection.

Metric	Strength	Weakness
TF	Simple in calculation	No order information; Prefer to high frequency terms
TF-IDF	Simple in calculation (Joachims, 1997)	No order information; Inefficiency for high frequency terms in long documents. (Manning & Schütze, 1999)
IG	Good in classification (Liu et al., 2003; Yang & Pedersen, 1997)	Depends much on training data set. (Lee & Lee, 2006)
MI	Low complexity (Estevez, Tesmer, Perez, & Zurada, 2009)	Has bias on rare terms; Sensitive to critical value. (Kalousis, Prados, & Hilario, 2007)
χ^2	Good in classification (Liu et al., 2003; Yang & Pedersen, 1997)	Defects for low-frequency terms. Depends much on training data set.
TS	Independent of predefined information	Complex in calculation. (Liu et al., 2003)

Please cite this article in press as: Yuan, H., et al. Make your travel smarter: Summarizing urban tourism information from massive blog data. *International Journal of Information Management* (2016), http://dx.doi.org/10.1016/j.ijinfomgt.2016.02.009