Towards Summarizing Popular Information from Massive Tourism Blogs

Hua Yuan, Hualin Xu, Yu Qian, Kai Ye
School of Management & Economics, University of Electronic Science and Technology of China, Chengdu 610054, China
Email: yuanhua@uestc.edu.cn, bruce123.xu@gmail.com, qiany@uestc.edu.cn, tingxueye@gmail.com

***Abstract*—In this work, we propose a research method to summarize popular information from massive tourism blog data. First, we crawl blog contents from website and segment each of them into a semantic *word vector* separately. Then, we select the geographical terms in each *word vector* into a corresponding *geographical term vector* and present a new method to explore the *hot tourism locations* and, especially, their *frequent sequential relations* from a set of geographical term vectors. Third, we propose a novel *word vector* subdividing method to collect the local features for each *hot location*, and introduce the metric of *max-confidence* to identify the *Things of Interest* (ToI) associated to the location from the collected data. We illustrate the benefits of this approach by applying it to a Chinese online tourism blog data set. The experiment results show that the proposed method can be used to explore the*hot locations*, as well as their sequential relations and corresponding ToI, efficiently.**

***Keywords*-blog mining; hot tourism locations; things of interest; max-confidence**

I. Introduction

In recent years, tourism has been ranked as the foremost industry in terms of volume of online transactions [1] and most tourism sites (such as blog.tripadvisor.com and www.travelblog.org) enable consumers to post blogs to exchange information, opinions and recommendations about the tourism destinations, products and services within a web-based communities. Meanwhile, some readers are more likely to enjoy a high quality travel experience from others' blogs. By obtaining reference knowledge from these blogs, individuals are able to visualize and manage their own travel plans. For instance, a person is able to find some places that attract him from other people's travel routes, and schedule an efficient and convenient (even economic) path to reach these places.

In this work, by deeming each tourism location in one's targeted destination as a travel "topic", and the ToI as some special interested local features associated with the location, we propose a research framework to summarize the popular tourism information from blogs as a whole. First, we crawl blog contents from website and divide each of them into a semantic *word vector* respectively. Then, we select the geographical terms from each blog vector to form a geographical data set. This data set has two characters: all the elements in the i-th record are the geographical terms that mentioned in the i-th blog; and, each record is transactional. More important, the elements in each record should keep their positional order as that in the original blog so that we can mine the *frequent sequential relationships* for some *hot geographical terms*. In real, such a sequential relation can be seen as a travel route. Third, we propose a vector subdividing method to collect the data set of local features for each *hot location*. The significant result of this method is shielding the impacts of irrelevant word co-occurrences which have very high frequency. Further, we present a new method basing on the measurement of *max-confidence* to identify the ToIs for each *hot location* from its local features.

II. Related Work

A. Blog summarization and text mining

In blog summarization literature, we note that the basic technology used in online text processing is text-mining [2], [3], which is used to derive insights from user-generated contents and primarily originated in the computer science literature [4], [5]. Thus some previous research were focused on automatically extracting the opinions of online contents [6] and the hot topics [7]. These methods used in blog mining not only involves reducing a larger corpus of multiple documents into a short paragraph conveying the meaning of the text, but also is interested in features or objects on which customers have opinions. Especially, some research have been focused on mining tourism blogs for better successors' decision-making [8] .

One important application of blog mining is sentiment analysis, which is to judge whether an online contents expresses a positive, neutral or negative opinion [9]. In recent, sentic computing [10] has brought together lessons from both affective computing and common-sense computing to grasp both the cognitive and affective information (termed semantics and sentics) associated with natural language opinions and sentiments [11], [12], which involves a deep understanding of natural language text by machine.

B. Topic model and feature selection

A topic model is a type of statistical model for discovering the "topics" that occur in a collection of documents and topic modeling is a way of identifying patterns in a corpus. An early topic model was probabilistic latent semantic indexing (PLSI), created by Thomas Hofmann [13] and the Latent Dirichlet Allocation (LDA), perhaps the most common topic model currently in use, is a generalization of PLSI developed

978-1-4799-4274-9/14 $31.00 © 2014 IEEE
DOI 10.1109/ICDMW.2014.29

by David Blei etc. in 2002 [14]. LDA is an unsupervised learning model basing on the intuition that documents are represented as mixtures over latent topics where topics are associated with a distribution over the words of the vocabulary. Therefore, it is good at finding word-level topics [15] and there were lot of proposed latent variable models basing on it [16].

Another type of method is try to look through a corpus for the clusters of words and groups them together by a process of similarity or relevance, for example, frequent pattern analysis [17] and co-expression analysis [18], [19]. In these methods, a text or document is always represented as a bag of words which raises two severe problem: the high dimensionality of the word space and the inherent data sparsity. In literature, feature selection is an important technology used to deal with the problems [20].

However, there are few impressive researches on providing blog readers valuable knowledge that they are personally interested in, i.e., the common topics from massive contents, as well as the special local features of these topics.

III. The Methodology

A. Problem statement

Given a set of tourism blogs $\mathbb{B}$, and assume that each blog in it can be represented by a *word vector*, thus $\mathbb{B}$ can be represented with a set of vectors as $\boldsymbol{B} = \{\boldsymbol{b}_1, ..., \boldsymbol{b}_i, ..., \boldsymbol{b}_{|\mathbb{B}|}\}$, and the total items in $\boldsymbol{B}$ is $\Sigma_{\boldsymbol{B}}$. There are two metrics of:

- $supp(X)$ (with a predefined threshold $mini\_supp$), which is used to measure the frequency of itemset X;
- $\theta_{\{x,y\}}$ (with a threshold θ_0), which is used to measure the dependency of item y on x (or vice versa).

The research problem can be specified as two subtasks:

- for the data set $\boldsymbol{B}$, find out a set of *frequent geographical terms* $\{b^H\} \in \Sigma_{\boldsymbol{B}}$ (i.e., *hot locations* in $\mathbb{B}$) such that $supp(b^H) \geq mini\_supp$. The position relations of these *frequent geographical terms* are studied as well;
- for each term of b^H, find out an appropriate set of terms $\{b_i\} \in \Sigma_{\boldsymbol{B}}$ such that $\theta_{\{b^H b_i\}} \geq \theta_0$.

B. Research framework

The presented research framework in this work is about three parts: *blog extraction and word segmentation* (BEWS), *frequent travel routes mining* (FTRM) and *interesting things detection* (ITD). In BEWS, each piece of blog is segmented into a set of *semantic words* so that it can be transformed into a *word vector*, in which, the elements are only *semantic words* and necessary *punctuation marks* after data cleaning. The FTRM subsystem is introduced to mine the travel route from the blog generated *word vectors*. The ITD subsystem is used to mine the ToIs for each *hot location*.

IV. Blog Contents Extraction

In BEWS subsystem, three subtasks of blog extraction, word segmentation and data cleaning are involved to transform a piece of blog into a *word vector*.

Web crawling technology can help people extracting information from the website. In this wrok, it used to obtain large-scale users generated blogs from a tourism website. All the blogs are crawled into an initial data set of $\mathbb{B}$.

Word segmentation is usually involving the tokenization of the input text into words at the initial stage of text analysis for NLP task [21]. In this work, it is the problem of dividing a string of written language into some component units.

For the work of data cleaning, we put it into an equivalent task on judging the usefulness of a component unit generated by the segmentation process. Here, we simply keep the follows as the useful word segments:

- *Semantic word (phrase)* [22], [23]: Our goal is to find the hot tourism locations and the ToI associated with them. In tourism blogs, almost all of these two things are presented in the form of nouns.
- *Punctuation marks* [24]: In any text-based document, a period, a question mark, or an exclamation mark is a real sentence-ending. Therefore, people could take them as the sentence boundaries. We use symbol "||" to represent all types of these reserved punctuation marks.

After data processing with the BEWS subsystem, finally, a set of *word vectors* as follows can be generated:

$$\boldsymbol{B} = \{\boldsymbol{b}_1,\ \boldsymbol{b}_2,\ ...,\ \boldsymbol{b}_{|\mathbb{B}|}\}. \tag{1}$$

All the items in $\boldsymbol{B}$ is $\Sigma_{\boldsymbol{B}} = \bigcup_{i=1}^{|\boldsymbol{B}|}\{b_{ij}\}$, where b_{ij} is the j-th term in $\boldsymbol{b}_i$.

V. Frequent Travel Routes Mining

A. Non-geographical terms eliminating

To find out the *frequent geographical terms* from the *word vectors*, a *geographical name table* is needed. This table can be extracted temporarily from an official travel guide providing by the local government, or provided by a creditable third part, for example, www.geonames.usgs.gov and maps.google.com.

Given a *geographical name table* denoted by GNT, firstly, we use it to filter out the non-geographical terms from $\boldsymbol{b}_i$ to form a *geographical term vector* as:

$$\boldsymbol{b}_{Gi} = \boldsymbol{b}_i \cap GNT, \tag{2}$$

then, a geographical data set of $\boldsymbol{B}_G$ is generated:

$$\boldsymbol{B}_G = \bigcup_{i=1}^{|\boldsymbol{B}|}\{\boldsymbol{b}_{G_i}\}. \tag{3}$$

Different from the traditional method, in $\boldsymbol{B}_G$, all the elements in the i-th record are the geographical terms that mentioned in the i-th blog, and more important, the elements

in each record should keep their positional order as that in the original blog. All the items in $\boldsymbol{B}_G$ is $\Sigma_{\boldsymbol{B}_G} = \bigcup_{i=1}^{|\boldsymbol{B}_G|}\{b_{ij}\}$, where b_{ij} is the j-th term in $\boldsymbol{b}_{G_i}$.

Example 1: Given a $GNT = \{A,\ B,\ C,\ D,\ E,\ F\}$, and a data set $\boldsymbol{B}$ composed of five word vectors as in Table I. We can obtain a geographical data set with relation (2) as shown in Table II.

Table I
DATA SET $\boldsymbol{B}$.

	Elements
$\boldsymbol{b}_1$	{a1 A a2 ‖ B b1 b2 ‖ C c1 ‖ D}
$\boldsymbol{b}_2$	{A a1 a2 ‖ B b1 ‖ D ‖ E }
$\boldsymbol{b}_3$	{A a1 ‖ b1 B b2 ‖ C c1 A c2}
$\boldsymbol{b}_4$	{B b1 ‖ D d1 E}
$\boldsymbol{b}_5$	{a1 a2 A a3 ‖ B b1 b2 ‖ C c1 ‖ F}

Table II
DATA SET $\boldsymbol{B}_G$.

	Elements
$\boldsymbol{b}_{G_1}$	{A B C D}
$\boldsymbol{b}_{G_2}$	{A B D E }
$\boldsymbol{b}_{G_3}$	{A B C A}
$\boldsymbol{b}_{G_4}$	{B D E }
$\boldsymbol{b}_{G_5}$	{A B C F}

B. Hot tourism location mining

From the perspective of *data set* in database technology, *word vector* $\boldsymbol{b}_i \in \boldsymbol{B}$ is also a transactional data record, so do the *geographical term vector* of $\boldsymbol{b}_{G_i} \in \boldsymbol{B}_G$. Therefore, we can mine the itemsets that appear in $\boldsymbol{B}_G$ frequently as the hot tourism locations for common people.

The frequent n-itemset X in data set $\boldsymbol{B}$ is denoted as:

$$FP^{(n)}(\boldsymbol{B}) = \{X | X \subseteq \Sigma_{\boldsymbol{B}}, |X| = n,\ supp(X) \geq mini\_supp\}, \tag{4}$$

where $supp(X)$ means the *support* of X in data set $\boldsymbol{B}$ and $mini\_supp$ is a predefined threshold.

Similarly, we can define $FP^{(n)}$ on any transaction data set. Especially, the frequent 1-itemsets in $\boldsymbol{B}_G$, i.e., $FP^{(1)}(\boldsymbol{B}_G)$, can be seen as the hot tourism locations. In example 1, we can obtain $FP^{(1)}(\boldsymbol{B}_G)$={A, B, C, D} with $mini\_supp = 60\%$.

Note that, the frequent pattern mining technology is not the main concern in this work, thus any feasible algorithms can be used depending on the contents of $\boldsymbol{B}$.

C. Travel routes generation

From a semantic perspective, the position relationship of all the elements in vector $\boldsymbol{b}_{G_i}$ shows the real travel sequence (location correlations) of blogger i. Thus, the common relations of all these geographical terms should lie in $\boldsymbol{B}_G$. Generally, the basic correlation between two locations is the co-occurrences and the adjacent position relationship of their representative terms in $\boldsymbol{B}_G$. Given two locations of $b_{is} \in \boldsymbol{b}_{G_i}$ and $b_{it} \in \boldsymbol{b}_{G_i}$, the adjacent position relationship means that b_{is} and b_{it} are mentioned frequently in $\mathbb{B}$ and $|Pos(b_{is}) - Pos(b_{it})| = 1$. In the following, we will propose a new method to reveal these location correlations.

First, we filter out all the unpopular locations (infrequent geography terms) from $\boldsymbol{b}_{G_i}$ to get a simplified *frequent geographical term vector*:

$$\boldsymbol{b}_{G_i^H} = \boldsymbol{b}_{G_i} \bigcap FP^{(1)}(\boldsymbol{B}_G). \tag{5}$$

Here, the elements in $FP^{(1)}(\boldsymbol{B}_G)$ indicate the common and frequent concerns of the bloggers (travellers). Thus, relation (5) tells us the *hot locations* mentioned in the i-th blog.

Lemma 1: $\boldsymbol{b}_{G_i^H} \subseteq \boldsymbol{b}_{G_i} \subseteq \boldsymbol{b}_i$.

Proof: According to (2) and (5), lemma 1 holds true. ■

Further, we can calculate

$$\boldsymbol{B}_{G^H} = \bigcup_{i=1}^{|\boldsymbol{B}|}\{\boldsymbol{b}_{G_i^H}\}. \tag{6}$$

All the items in $\boldsymbol{B}_{G^H}$ is $\Sigma_{\boldsymbol{B}_{G^H}}$.

Next, assume that m_i *hot locations* appear sequentially in the i-th blog:

$$\boldsymbol{b}_{G_i^H} = \{b_{i1}^H\ b_{i2}^H\ ...\ b_{im_i}^H\}.$$

To keep the position information of these *hot locations* in $\boldsymbol{b}_{G_i^H}$, we transform the formation of $\boldsymbol{b}_{G_i^H}$ into

$$\tilde{\boldsymbol{b}}_{G_i^H} = \{b_{i1}^H b_{i2}^H,\ b_{i2}^H b_{i3}^H,\ ...\ b_{i(j-1)}^H b_{ij}^H,\ b_{ij}^H b_{i(j+1)}^H,\ ...\}, \tag{7}$$

where "$b_{ij}^H b_{i(j+1)}^H$" in $\tilde{\boldsymbol{b}}_{G_i^H}$ means that two *hot locations* of b_{ij}^H and $b_{i(j+1)}^H$ are mentioned sequentially in blog $\boldsymbol{b}_i$. Similarly,

$$\tilde{\boldsymbol{B}}_{G^H} = \bigcup_{i=1}^{|\boldsymbol{B}|}\{\tilde{\boldsymbol{b}}_{G_i^H}\}. \tag{8}$$

For example 1, if $FP^{(1)}(\boldsymbol{B}_G)$={A, B, C, D}, then the calculation results of $\boldsymbol{B}_{G^H}$ and $\tilde{\boldsymbol{B}}_{G^H}$ are shown in Table III and IV respectively. In Table IV, if we set $mini\_supp = 40\%$, then $FP^{(1)}(\tilde{\boldsymbol{B}}_{G^H}) = \{AB, BC, BD\}$. Any elements in $FP^{(1)}(\tilde{\boldsymbol{B}}_{G^H})$ indicates a correlation between two *hot locations* which are mentioned sequentially and frequently by blogers.

Table III
DATA SET $\boldsymbol{B}_{G^H}$.

Vector	Elements
$\boldsymbol{b}_{G_1^H}$	{A B C D}
$\boldsymbol{b}_{G_2^H}$	{A B D}
$\boldsymbol{b}_{G_3^H}$	{A B C A}
$\boldsymbol{b}_{G_4^H}$	{B D }
$\boldsymbol{b}_{G_5^H}$	{A B C}

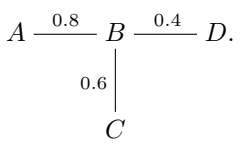
Table IV
DATA SET $\tilde{\boldsymbol{B}}_{G^H}$.

Vector	Elements
$\tilde{\boldsymbol{b}}_{G_1^H}$	{AB BC CD}
$\tilde{\boldsymbol{b}}_{G_2^H}$	{AB BD}
$\tilde{\boldsymbol{b}}_{G_3^H}$	{AB BC CA}
$\tilde{\boldsymbol{b}}_{G_4^H}$	{BD }
$\tilde{\boldsymbol{b}}_{G_5^H}$	{AB BC}

From a network perspective, the *hot locations* in $FP^{(1)}(\boldsymbol{B}_G)$ and their correlations in $FP^{(1)}(\tilde{\boldsymbol{B}}_{G^H})$ can form a *route network* of $G = (V, E)$ by setting the vertex set as $V = FP^{(1)}(\boldsymbol{B}_G)$ and setting the edge set as $E = FP^{(1)}(\tilde{\boldsymbol{B}}_{G^H})$.

In order to facilitate the calculation, the weight of edge $(b_{ij}, b_{i(j+1)})$ can be set as $w_{(b_{ij}^H, b_{i(j+1)}^H)} = supp(\{b_{ij}^H, b_{i(j+1)}^H\})$. For the $FP^{(1)}(\tilde{\boldsymbol{B}}_{G^H})$ generated by the

data in Table IV ($mini\_supp = 40\%$), a weighted *route network* is formed as follows:



VI. ToI Extraction

A. Hot-location-based blog word vector subdividing

It is known to all, if there are too many hot tourism locations and popular interesting things in the same *word vector*, the impact of noise correlations between elements is inevitable when mine the frequent patterns directly from $\boldsymbol{B}$. Fortunately, blogers tend to present something interesting around (or nearby) the appearances of a *hot location* in blogs. Moreover, if the contents of some adjacent sentences are about the same *hot location* (topic), then these sentences may constitute a topic domain for the location. Also, the topic domain are usually bounded by a set of punctuation marks. We can expect the most close local features (ToI) for each *hot location* would exist in such a topic domain and are bounded by a pair of adjacent punctuation marks.

According to lemma 1, the elements in $\boldsymbol{b}_i$ can be classified into two types (overlap is allowed): *hot locations* (i.e., elements also existing in $\boldsymbol{b}_{G_i^H}$) and the others. Therefore, $\boldsymbol{b}_i$ can be subdivided into sub-itemsets according to the topic boundaries of its m_i *hot locations* (see relation (V-C)). For the j-th hot location b_{ij}^H, we just need to cut the adjacent elements which belong to the same topic domain of b_{ij}^H. Since blog sentences are usually ended by a set of punctuation marks, so the divided sub-itemset for b_{ij}^H can be started at the first punctuation mark before b_{ij}^H (the first term denoted by b_{ij_S}) and ended at the first punctuation mark just before the next *hot location* of $b_{i(j+1)}^H$ (the last term denoted by b_{ij_E}):

$$\boldsymbol{b}_i = \{..., \| \underbrace{\ b_{ij_S}, ...,\ b_{ij}^H,\ ..., b_{ij_E}\ }_{j-th\ sub-itemset} \| \ b_{i(j+1)_S}, ...,\ b_{i(j+1)}^H\},$$

where $b_{ij}^H \in FP^{(1)}(\boldsymbol{B}_G), j = 1, ..., m_i$ and symbol "$\|$" denotes the punctuation marks. According to the usual text writing style of blogers, we can expect that the ToI for location b_{ij}^H may lie in the itemset of $\{b_{ij_S}, ...,\ b_{ij}^H,\ ...,\ b_{ij_E}\}$ with a high possibility. More generally, given a *hot location* $b^H \in FP^{(1)}(\boldsymbol{B}_G)$, we define its *cut vector* in $\boldsymbol{b}_i$ as follows:

$$CUT(b^H|\boldsymbol{b}_i) = \{\ b_{ij_S}, ...,\ b^H,\ ...,\ b_{ij_E}\}. \tag{9}$$

All the *cut vectors* for b^H from $\boldsymbol{B}$ is:

$$CUT(b^H) = \bigcup_i \{CUT(b^H|\boldsymbol{b}_i)\}. \tag{10}$$

All the items in $CUT(b^H)$ is $\Sigma_{CUT(b^H)} = \bigcup_{i=1}^{|CUT(b^H)|}\{b_{ij}\}$, where b_{ij} is the j-th term in $CUT(b^H|\boldsymbol{b}_i)$.

Obviously, there are two key elements in a *cut vector*, i.e., a *hot location* term as the semantic key word, and two punctuation marks as the semantic topic boundaries for such a key word. In fact, $CUT(b^H)$ can be deemed as a set of *local contexts* for term b^H which refers to either an ordered sequence or unordered set of other useful words in the same sentence (or in a set of adjacent sentences), such that they co-occur, have syntactic dependencies, or both [25], [26]. Consequently, we can expect that all the potential ToI about location b^H may lie in $CUT(b^H)$.

In example 1, we know that *hot locations* are $FP^{(1)}(\boldsymbol{B}_G)$={A, B, C, D} while $mini\_supp = 60\%$, then the data sets of local features for all the *hot locations* are shown in Tables V-VIII.

Table V
$CUT(A)$.

Vector $\boldsymbol{b}_i$	$CUT(A\|\boldsymbol{b}_i)$
$\boldsymbol{b}_1$	{a1 A a2}
$\boldsymbol{b}_2$	{A a1 a2 }
$\boldsymbol{b}_3$	{A a1 C c1 A c2}
$\boldsymbol{b}_4$	{}
$\boldsymbol{b}_5$	{a1 a2 A a3}

Table VI
$CUT(B)$.

Vector $\boldsymbol{b}_i$	$CUT(B\|\boldsymbol{b}_i)$
$\boldsymbol{b}_1$	{B b1 b2}
$\boldsymbol{b}_2$	{B b1 }
$\boldsymbol{b}_3$	{b1 B b2 }
$\boldsymbol{b}_4$	{B b1}
$\boldsymbol{b}_5$	{B b1 b2}

Table VII
$CUT(C)$.

Vector $\boldsymbol{b}_i$	$CUT(C\|\boldsymbol{b}_i)$
$\boldsymbol{b}_1$	{C c1}
$\boldsymbol{b}_2$	{}
$\boldsymbol{b}_3$	{C c1 A c2}
$\boldsymbol{b}_4$	{}
$\boldsymbol{b}_5$	{C c1}

Table VIII
$CUT(D)$.

Vector $\boldsymbol{b}_i$	$CUT(D\|\boldsymbol{b}_i)$
$\boldsymbol{b}_1$	{D}
$\boldsymbol{b}_2$	{D }
$\boldsymbol{b}_3$	{}
$\boldsymbol{b}_4$	{D d1 E}
$\boldsymbol{b}_5$	{}

B. Dependency between two frequent co-occurred terms

When bloggers record a tourist location b_1 in a city, some of them would tend to review a scenic spot b_2 which has relation with b_1. Under such a case, the appearances of scenic spot ($supp(\{b_2\})$) are dependent on the blogs in which both the location and the scenic spot were reviewed ($supp(\{b_1, b_2\})$). That is to say, the scenic spot has a strong relation (dependency) with the tourism location, if somebody inquires the information about the location b_1, then the information of b_2 should be provided since it is an necessary ToI of b_1. Here, we identify the dependent relation between b_1 and b_2 in the 2-itemset $X = \{b_1, b_2\}$ with a metric of *max-confidence*, which is defined as [27]:

$$\theta_X = \frac{supp(X)}{\min\{supp(\{b_1\}), supp(\{b_2\})\}}. \tag{11}$$

In this work, the *max-confidence* is used to measure the dependency of a candidate ToI on a *hot location*: given a predefined value of $\theta_0 \in [0, 1]$, we called $b \in \Sigma_{CUT(b^H)}$ is a ToI associated to *hot location* b^H, if

$$\theta_{\{b^H, b\}} \geq \theta_0. \tag{12}$$

C. ToI Extraction Algorithm

Given a *hot location* $b^H \in FP^{(1)}(\boldsymbol{B}_G)$ and a potential ToI term b, the ToI extraction processes are mainly focused on calculating the value of $supp(b|CUT(b^H))$ and $min\{supp(b^H), supp(b)\}$ and analyzing the dependency of b on b^H in data set $CUT(b^H)$ with *max-confidence*.

Algorithm 1 ToI Extraction Algorithm

1: **Input**: Word vector set $\boldsymbol{B}$; Geographical word vector set $\boldsymbol{B}_G$; Hot tourism location set $\Sigma_{\boldsymbol{B}_{G^H}}$; θ_0;
2: **Output**: ToI set;
3: $CUT = \phi$;
4: **for** $i = 1$ to $|\boldsymbol{B}|$ **do**
5: $\quad \boldsymbol{b}_{G_i^H} = \boldsymbol{b}_{G_i} \bigcap FP^{(1)}(\boldsymbol{B}_G)$;
6: $\quad$ **for** $j = 1$ to $|\boldsymbol{b}_{G_i^H}|$ **do**
7: $\quad\quad$ Calculate $CUT(b_{ij}^H)$ from $\boldsymbol{b}_i$;
8: $\quad\quad$ $CUT(b^H) \leftarrow CUT(b_{ij}^H)$ where $b^H = b_{ij}^H$;
9: $\quad$ **end for**
10: **end for**
11: $CUT \leftarrow CUT(b^H)$ for all the $b^H \in FP^{(1)}(\boldsymbol{B}_G)$;
12: Calculate $FP^{(1)}(\boldsymbol{B})$;
13: **for** $i = 1$ to $|CUT|$ **do**
14: $\quad$ Calculate $FP^{(1)}(CUT_i)$ and $FP^{(2)}(CUT_i)$;
15: $\quad$ **for** each $b \in FP^{(1)}(CUT_i)$ **do**
16: $\quad\quad$ **if** $\theta_{\{b_i^H, b\}} \geq \theta_0$ **then**
17: $\quad\quad\quad$ **if** $\theta_{\{b,b' \} \in FP^{(2)}(CUT_i)} \geq \theta_0$ **then**
18: $\quad\quad\quad\quad$ $ToI(b_i^H) \leftarrow < b^H, b, b' >$;
19: $\quad\quad\quad$ **else**
20: $\quad\quad\quad\quad$ $ToI(b_i^H) \leftarrow < b^H, b >$;
21: $\quad\quad\quad$ **end if**
22: $\quad\quad$ **end if**
23: $\quad$ **end for**
24: **end for**
25: **return** $\bigcup ToI(b_i^H)$.

Algorithm 1 goes through three phases:

- Finding out the *hot locations* from blog i (Line 5);
- Obtaining *cut vector* for b_{ij}^H in blog i and putting it into the appropriate data set of $CUT(b^H)$ (Line 6-10). All the $CUT(b^H)$ were put into CUT (Line 11).
- Analyzing the dependency of all the elements in $\Sigma_{CUT(b^H)}$ on *hot location* b^H (Line 12-24).

To obtain the complete interdependent relationships in $CUT(b^H)$, we set the max-confidence computation as a progressive process (Line 16-22).

The algorithm shows (1) the analysis of the dependency of potential ToI term b on b^H in data set $CUT(b^H)$ need to scan all the items in $\boldsymbol{B}$; (2) the computation complexity has been approximate to $\#\ of\ hot\ terms \times |FP^{(1)}(CUT(b^H))| \times |FP^{(2)}(CUT(b^H))|$. That is, the efficiency of the algorithm is affected by the *minimum support* threshold in mining frequent 1- and 2-itemset from $CUT(b^H)$.

VII. Experimental Results

A. Experiment setup

The blogs data were extracted from the www.mafengwo.com, one of the most famous blog sites in China for tourism information sharing. Altogether, 450 blogs posted from 10-01-2006 to 01-31-2014 in "Hongkong" (targeted destination) tourism channel were collected with a blog extraction tool. The blogs with empty text (some blogs are pictures only) were removed and 396 valid travel blogs were remained for the following experiments.

The contents of GNT are extracted from the attraction terms on the www.tripadvisor.com.

B. Word segmentation

First, an initial data set of *all terms* is generated after data cleaning. Then, we extract the data set of *nouns* from the *all terms* by removing the non-noun *word segments*. To identify the *hot locations* and their sequential relations efficiently, we further select a special data set of *geographical terms* from the *nouns* according to the GNT. The frequency of terms in the data set of *all terms*, *nouns* and *geographical terms* are sorted in Fig.1.

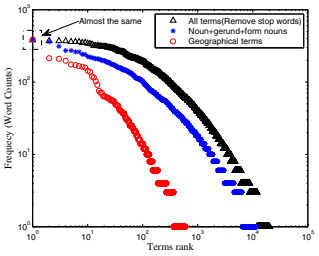


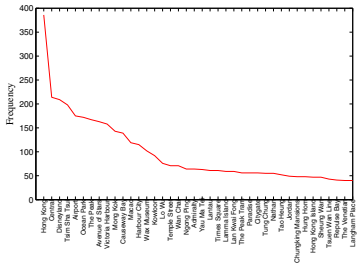
Figure 1. Frequencies of word segment in blogs.

Different from the systems characterized by short text, such as micro-blogs, the frequency of terms in *all terms* to their corresponding ranks does not follow the common power-law distribution. The most likely reason is that the document length of blogs is much longer than that of the contents in BBS or microblogs. This result indicates that the blogging behaviors of bloggers are independent from each other, but they try to sketch out the travel experiences in detail to obtain blog readers' appreciations, which will result in longer document length and various words used in blog contents. On the other hand, the ranks of geographical terms to their corresponding frequencies is of power-law distribution. This illustrates that, there are few geographical terms are very popular while the most of others are not.

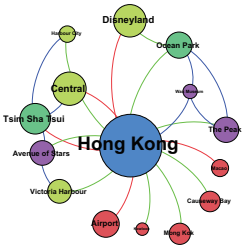
C. Hot tourism locations and travel routes

For the blogs about Hongkong tourism, the ranks of geographical terms to their frequency are shown in Figure 2(a). There are two significant turning points on the

curve: *"Central"* and *"Kowloon"*. In order to reserve as much valuable information, we take the top 15 geographical terms whose frequencies are bigger (and equal) than that of *"Kowloon"* as the *hot locations* in Hongkong.



(a) Frequent geographical terms.



(b) Popular travel routes for Hongkong.

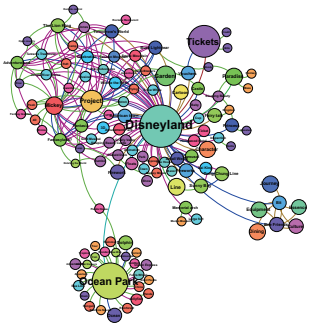
Figure 2. Backbone-nodes-based travel routes for Hongkong.

Using the 15 *hot locations* as network nodes and analyzing their position relations in the blogs, we can sketch out a popular travel routes for blogers' travelling in Hongkong (Figure 2(b)). In which, each node represents a *hot location*, and the node size shows the frequency of the location in all the blogs. Obviously, the larger the node, the tour location it represented is more frequent. The lines in Figure 2(b) show that there exists some travel sequential relationships between the connected nodes. For example, "Ocean park-Disneyland", "Central-Harbour City", and so on.

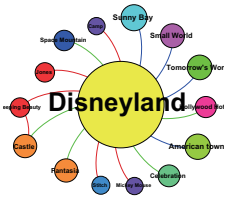
D. ToI extraction

We use the "*Disney land*" as an illustration for the ToI extraction. Firstly, the local features located nearby the term of "Disney land" and bounded by a pair of adjacent punctuation marks are cut from each blog generated *word vector* to form the data set of $CUT(\text{"}Disney\ land\text{"})$. Then, the frequent 1- and 2-itemsets in $CUT(\text{"}Disney\ land\text{"})$ are mined. Finally, the extracted ToI for the *hot location* of "Disney land" are shown in Figure 3.

Setting the threshold of *max-confidence*=0.6, people can obtain more ToIs around Disneyland (Fig 3(a)). Some nearby tourist attractions, such as "*Ocean Park*", and their primary ToI has also been extracted partially because they are described frequently by blogers in the same context of a comparative or associated manner. Interestingly, other matters related to Disney tourism, such as the ticket ("*Tickets*") and



(a) Threshold of Max-confidence=0.6.



(b) Threshold of Max-confidence=0.95.

Figure 3. Extracted ToI for "Disney park".

the public transport ("*Tung Chung Line*") are also presented, which are ToI highly relevant to the Disney tourism and may provide richer information for people's travel planning. However, a relative loose threshold of *max-confidence* would cause the correlations between ToI to become more complex. To obtain the clear relationship between a *hot location* and its most close ToI, thus, a bigger threshold of *max-confidence* is needed (Fig 3(b)). As we can see, these ToI are the most popular tourist projects for "Disney land".

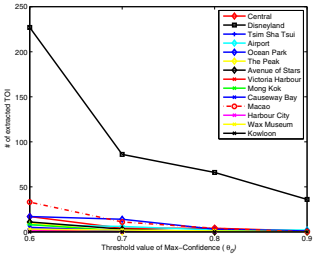


Figure 4. Number of extracted ToI for the top-14 *hot locations*.

At different values of θ_0, the number of extracted ToI for the top-14 *hot locations* in Hongkong (term "Hongkong" is ignored) are shown in Figure 4, in which, we can see that "Disneyland", "Macao" and "Ocean park" are three *hot locations* having more ToI than the others. Which means, when people tour in these locations, they may spend more time and money. This is in accordance with the common sense in nature. For the rest *hot locations* with few ToI, people can bind them into several travel packages according to their geographic relationship so that these locations in the

same package can be visit together.

E. Performance

In this section we discuss the measures used in evaluating the performance of the experiments.

1) Evaluation measure: For the task of blog extracting, people has four possible outcomes for the extracted and inherent ToI, as shown in Table IX.

Table IX
CLASSIFICATION OF THE RESULTS OF A TOI EXTRACTION TASK.

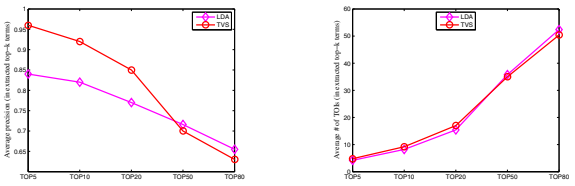
	Extracted	Not extracted
Relevant ToI	True-Positive (tp)	False-Negative (fn)
Irrelevant words	False-Positive (fp)	True-Negative (tn)

In literature of information retrieval, $Precision$ and $Recall$ are used to measure the extraction results:

$$Precision = \frac{\#tp}{\#tp + \#fp}\ ,\ Recall = \frac{\#tp}{\#tp + \#fn}. \quad (13)$$

In the following, we will compare the efficiency of our method, namely *Term Vector Subdividing* (TVS), with some classic methods, and show the comparison results of (1) the average $precision$ and (2) the average number of ToI ($\#tp$) in the extracted top-k terms.

2) Experiment results: Firstly, we conduct the comparison experiment with TVS and LDA on the data set of *nouns*. The results are shown in Figure 5.

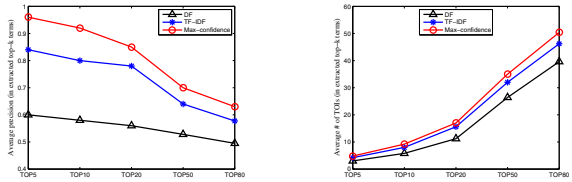


(a) Average precision. (b) Average $\#tp$.

Figure 5. Comparison results between TVS and LDA.

Notably, in extracting a small number of ToI (e.g., less than 50), the precision of TVS is superior to that of LDA (Fig 5(a)). This shows that the TVS method is good at extracting terms that linked very closely with the key term (*hot location*). However, the accuracy of TVS begins to decrease with the increasing number of top-k threshold (provide more ToI), whereas, LDA performs better. One possible explanation is that TVS must run with a relative lower threshold of *max-confidence* when it is required to provide more extraction contents. Obviously, this may introduce increasing number of noise into the ToI candidates.

In additional, TVS is a method basing on feature selection with the metric of *max-confidence*. So, we do some experiments to see the differences between the classic TF-IDF, DF and the *max-confidence* in ToI extraction. The data set used here are $CUT(b_i^H)$, where b_i^H $(i = 14)$ is one of the top-14 *hot locations* in Hongkong (see Fig. 2(b)). The averaged results are shown in Figure 6:



(a) Average precision. (b) Average $\#tp$.

Figure 6. Comparison results between metrics.

- For the results, *max-confidence* dominated all other metrics, and the DF metric shows the worst performance. *Max-confidence* is better than TF-IDF because TF-IDF does not take into account interesting word co-occurrences containing terms with low IDF [28].
- Along with the number of requested features becomes bigger, the number of extracted ToI by different metrics are all keep increasing (Fig 6(b)). This result makes sense that, a bigger threshold will result in a larger set of candidates for real ToI.

VIII. CONCLUSION

In this work, we propose a research methodology to summarize the popular information from massive tourism blog data. To this end, we firstly crawl blog contents from website and divide them into semantic *word vectors* as data source. Second, we collect the geographical data from all the blog vectors, and mine the hot tourism locations and their *frequent sequential relations* in it. The results of this part can be used to summarize the popular information about "where to go" (trip route) in a set of tourism blogs. Then, we propose a vector subdividing method to collect local features for each *hot location*, and introduce the *max-confidence* metric to identify the ToI for the corresponding *hot location*. The captured ToI for each *hot location* are account for the question about "what to play" at a specific tourism location. Notably, the significant result of this method is that the disturbances from high frequent irrelevant word (noise) are shielded. Finally, we illustrate the benefits of this approach by applying it to a Chinese online tourism blog data set.

The experiment results show that the proposed method can be used to explore the hot tourism locations (their sequences as well) and their corresponding ToI from massive blogs efficiently. Future work is about reducing the algorithm complexity and presenting an optimization method to extract more precise correlations for a hot term.

ACKNOWLEDGMENTS.

The work was partly supported by the National Natural Science Foundation of China (71271044/U1233118/ 71102055).

REFERENCES

[1] H. Werthner and F. Ricci, "E-commerce and tourism," *Communications of the ACM*, vol. 47, no. 12, pp. 101–105, 2004.

[2] Q. Cao, W. Duan, and Q. Gan, "Exploring determinants of voting for the "helpfulness" of online user reviews: A text mining approach," *Decision Support Systems*, vol. 50, no. 2, pp. 511–521, 2011.

[3] A. Ghose and P. G. Ipeirotis, "Estimating the helpfulness and economic impact of product reviews: Mining text and reviewer characteristics," *IEEE Transactions on Knowledge and Data Engineering*, vol. 23, no. 10, pp. 1498–1512, 2011.

[4] M. Hu and B. Liu, "Mining and summarizing customer reviews," in *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2004, pp. 168–177.

[5] B. Pang and L. Lee, "Opinion mining and sentiment analysis," *Foundations and Trends in Information Retrieval*, vol. 2, no. 1-2, pp. 1–135, 2008.

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

[7] N. Li and D. D. Wu, "Using text mining and sentiment analysis for online forums hotspot detection and forecast," *Decision Support Systems*, vol. 48, no. 2, pp. 354–368, 2010.

[8] N. Sharda and M. Ponnada, "Tourism blog visualizer for better tour planning," *Journal of Vacation Marketing*, vol. 14, no. 2, pp. 157–167, 2008.

[9] B. Liu, "Sentiment analysis and subjectivity," *Handbook of Natural Language Processing*, vol. 2nd ed, 2010.

[10] E. Cambria and A. Hussain, *Sentic Computing: Techniques, Tools, and Applications.* Netherlands: Springer, 2012.

[11] E. Cambria, B. Schuller, Y. Xia, and C. Havasi, "New avenues in opinion mining and sentiment analysis," *IEEE Intelligent Systems*, vol. 28, no. 2, pp. 15–21, 2013.

[12] S. Poria, E. Cambria, G. Winterstein, and G. Huang, "Sentic patterns: Dependency-based rules for concept-level sentiment analysis," *Knowl.-Based Syst.*, vol. 69, pp. 45–63, 2014.

[13] T. Hofmann, "Probabilistic latent semantic indexing," in *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 1999, pp. 50–57.

[14] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *Journal of Machine Learning Research*, vol. 3, pp. 993–1022, 2003.

[15] A. Banerjee and S. Basu, "Topic models over text streams: A study of batch and online unsupervised learning." in *SDM*. SIAM, 2007, pp. 431–436.

[16] S. Moghaddam and M. Ester, "On the design of LDA models for aspect-based opinion mining," in *Proceedings of the 21st ACM International Conference on Information and Knowledge Management*, 2012, pp. 803–812.

[17] H. D. Kim, D. H. Park, Y. Lu, and C. Zhai, "Enriching text representation with frequent pattern mining for probabilistic topic modeling," *Proceedings of the American Society for Information Science and Technology*, vol. 49, no. 1, pp. 1–10, 2012.

[18] M. Rokaya, E. Atlam, M. Fuketa, T. C. Dorji, and J.-i. Aoe, "Ranking of field association terms using co-word analysis," *Information Processing and Management*, vol. 44, no. 2, pp. 738–755, 2008.

[19] F. Figueiredo, L. Rocha, T. Couto, T. Salles, M. A. Gonçalves, and W. Meira Jr., "Word co-occurrence features for text classification," *Information Systems*, vol. 36, no. 5, pp. 843–858, 2011.

[20] T. Liu, S. Liu, Z. Chen, and W.-Y. Ma, "An evaluation on feature selection for text clustering." in *ICML*. AAAI, 2003, pp. 488–495.

[21] J. Gao, M. Li, C.-N. Huang, and A. Wu, "Chinese word segmentation and named entity recognition: A pragmatic approach." *Computational Linguistics*, vol. 31, pp. 531–574, 2005.

[22] R. Sproat, C. Shih, W. A. Gale, and N. Chang, "A stochastic finite-state word-segmentation algorithm for Chinese," *Computational Linguistics*, vol. 22, pp. 377–404, 1996.

[23] A. Stavrianou, P. Andritsos, and N. Nicoloyannis, "Overview and semantic issues of text mining," *SIGMOD Record*, vol. 36, no. 3, pp. 23–34, 2007.

[24] J. Tang, H. Li, Y. Cao, and Z. Tang, "Email data cleaning," in *Proceedings of the Eleventh ACM SIGKDD International Conference on Knowledge Discovery in Data Mining*, 2005, pp. 489–498.

[25] H. T. Ng and H. B. Lee, "Integrating multiple knowledge sources to disambiguate word sense: An exemplar-based approach," in *Proceedings of the 34th Annual Meeting on Association for Computational Linguistics*, 1996, pp. 40–47.

[26] D. Lin, "Using syntactic dependency as local context to resolve word sense ambiguity," in *Proceedings of the Eighth Conference on European Chapter of the Association for Computational Linguistics*, 1997, pp. 64–71.

[27] T. Wu, Y. Chen, and J. Han, "Re-examination of interestingness measures in pattern mining: A unified framework," *Data Mining and Knowledge Discovery*, vol. 21, no. 3, pp. 371–397, 2010.

[28] A. Pons-Porrata, R. Berlanga-Llavori, and J. Ruiz-Shulcloper, "Topic discovery based on text mining techniques," *Information Processing & Management*, vol. 43, no. 3, pp. 752–768, 2007.