Event Based Multiple Tourism Themes' Determination From Texts For Alternative Tourism Recommendations

Nattapong Savavibool and Chaveevan Pechsiri

Dept. of Information Technology, Dhurakij Pundit University
Bangkok, Thailand.

ABSTRACT – This research aims to determine the multiple tourism themes based on attractiveness events expressed by the action verbs from the tourism web documents of the selected region. These several themes can be used for recommending tourists with alternative tourism theme choices. The problems of tourism themes' acquisition from the web blog texts are to determine the touristic themes and to identify a touristic event base on a simple sentence or EDU (Elementary Discourse Unit). This research proposes using the k-means clustering technique based on events expressed by verb phrases to determine the multiple tourism themes for a group of provinces within a region. Each cluster represents its own events while some of these events can be determined as the tourism themes by using verb-noun co-occurrences with the tourism event concepts. The result of the event-based tourism theme determination is evaluated by comparing to the answer set of the tourism highlight of each province provided by Tourism Authority of Thailand (http://thai.tourismthailand.org/), and our proposed methodology shows successfully results

KEYWORDS -- Multiple Tourism Theme; Clustering; Word Co-Occurrence

1. Introduction

The Event Based Tourism themes (named EBT themes) are tourism knowledge that is important for recommending the attractive tourism events expressed by the action verbs (i.e. hike, boat, swim, roam, cook, ride a horse, view, and etc,) of a certain region to people for preparing their trips with the alternative choices of themes. The EBT themes are also useful for guiding the tourism company to set up the alternative program tours with attractiveness to the certain area or province. How to automatically determine multiple EBT themes of the region as the destination from the web documents is a challenge task. According to [1] and WordNet (http://wordnet.princeton.edu/), theme is "a unifying idea that is a recurrent element in a literary or artistic work". To obtain the theme from text, especially the tourism theme, it is necessary to read information/ knowledge on documents from several resources of a destination for the concept unification, which is time consuming. Hence, the automatic system is required to determine multiple EBT themes from documents within the selected region. Furthermore, there are many types of tourism according to the tourism categorization, such as Culinary Tourism, Cultural Tourism (including Historical Tourism and Heritage Tourism), Ecotourism, Agritourism, Extreme Tourism (involving travel to dangerous places as mountains, jungles, deserts, caves, etc.), Geotourism, Medical Tourism, Nautical Tourism, Adventure Tourism, Academic Tourism, Wildlife Tourism, and etc. (http://en.wikipedia.org/wiki/tourism) However, our research will focus on Cultural Tourism, Ecotourism and Adventure Tourism, which follows the Tourism Authority of Thailand (http://thai.tourismthailand.org). The documents used in our research are based on the tourism web blogs on which most people express their opinions (about where to go, their tourist objective, the enjoyable activities, and so on) being necessary for the EBT theme determination. The EBT themes are beneficial for the tourism business to manage the tour programs. Moreover, each EBT expression on texts is based on EDU (Elementary Discourse Unit) which is a simple sentence/a clause defined by [2].

Past researches have worked on theme determination from textual data ([3] [4] and [5]). In 2004, Stanley Loh et al. [3] applied the association text mining technique of to discover a theme of interesting areas in the messages from Web chat. Qiaozhu Mei et al. [4] used a probabilistic approach to the text mining model on web blogs for spatiotemporal theme pattern determination. And, Qiang Hao et al. [5] worked on mining the knowledge by using the statistical based approach to determine the global topics from travelogues for representing common themes shared by various locations. Most of the previous researches determine several themes (expressed by all words of nouns, verbs and adjectives) of each destination of their certain location from web documents whereas our research concerns on determining multiple EBT themes (expressed by verb phrases) of one region as the destination from web blogs for recommending the alternative interesting tour programs. However, there are some problems in our research as the multiple-EBT-themes determination problem, the EBT-EDU identification problem, and the sub-region (equivalent to the province in this research) names' ellipsis problem, especially the web blog containing several provinces. According to (http://home.dei.polimi.it/matteucc/Clustering/tutorial_html/), clustering is defined as "the process of organizing objects into groups whose members are similar in some way". Therefore, we propose using the clustering technique to cluster the events occurred on the tourism web blogs along with the Natural Language Processing techniques to achieve the determination of multiple EBT themes.

Our research will be separated into 5 sections. In section II, related work is summarized. Problems of tourism theme determination from Thai documents will be described in section III and in section IV our framework for multiple tourism theme determination. In section V, we evaluate and conclude our proposed model.

2. Related Work

There are many research works on the theme determination but only a few researched on the tourism themes ([3] [4] [5] and [6]).

In 2004, Stanley Loh et al 2004, In 2004, Stanley Loh et al. [3] proposed using the TextMiningSuite software along with the tourism ontology ries to discover options of interesting areas from Web chat to recommend travel agents' customers who did not know where to go. The tourism ontology used in their research has been manually constructed with inserting the tourism theme/ category attributes inside. Their system was focused only in the problem of finding explicit cities and attractions for the customer. Their text mining module is used to identify the themes related to each term/word item. One item may be associated to many themes causing to determine the degree of relation by using fuzzy reasoning process to reduce the association rules. However, their relationships or the association rules between terms and themes are based on the explicit city whereas there are some implicit province's names (equivalent to city's names).

Qiaozhu Mei et al. (2006) [4] stated that themes are subtopics highly associated with a broad event (or topic). Then, they constructed each data set by collecting blog entries that are relevant to a given topic. They proposed a novel probabilistic approach to determine the subtopic themes and spatiotemporal theme patterns simultaneously. The proposed model discovers spatiotemporal theme patterns by extracting common themes/subtopic from weblogs by ; 1) generating theme life cycles for each given location; and 2) generating theme snapshots for each given time period. In the Hurricane Rita data set, 1403 documents out of 1754 have location labels. As in Hurricane Katrina, they choose a state as the smallest granularity of locations. Compare with the themes extracted from the Hurricane Katrina data set from 7 118 documents out of 9377 have location information. Two data sets shared several similar themes excepting the "Personal Life" theme. However, they did not show how to evaluate their model.

In 2006, [6] mined association rules of activities or events and tourism attractions in touristic destinations were mined by Apriori algorithm from a 10,000 Japanese blog. Their feature includes nouns and activity verbs, except verbs indicating movements. An average of 0.58 and 0.575 were achieved for precision and recall respectively.

Qiang Hao et al. (2010) [5] mined location-representative knowledge from a large collection of travelogues, They proposed a probabilistic topic model, named as Location-Topic model, which consisted of local topics and global topics. The local topics characterized locations whereas the global topics represented other common themes shared by various locations. The representation of locations in the local topic space encoded both location-representative knowledge and similarities between locations. The research decomposed travelogue documents into local topics and global topics. A travelogue collection could be represented by a Term-Document matrix decomposed into multiple matrices, including Term-LocalTopic, Term-GlobalTopic, and LocalTopic-Location matrices. GlobalTopic-Document matrix and Location-Document matrix were also learned for gaining some information used for the location extraction. According to the decomposition of likelihood of a term/word (w) in a document (d) as p(w|d) , each word in a document was assumed to be a binary decision between two paths: a location with a local topic and a global topic. As the decomposed matrix, a travelogue collection preserved its location-representative knowledge in LocalTopic-Location matrix, and topics in Term-LocalTopic and Term-GlobalTopic matrices. According to the location representation by metrix in the local topic space, the symmetric similarity between two locations l1and l2 was measured by the distance between their corresponding multinomial distributions over local topics from each location along with the Jensen-Shannon (JS) divergence and the term vocabulary based on a probability distribution over the learnt local topics Their framework shows promising results of effectively recommend destinations for flexible queries and summarize destinations from the large corpus of approximately 100,000 travelogues written in English with relating to tourist destinations in the United States.

The previous researches [3] [4] [5] and [6] worked on the themes based on all words of nouns, verbs, and adjectives on documents/web-documents whereas our research emphasized on determining multiple tourism themes based on events expressed by verb phrases from the small size corpus of approximately 1000 Thai tourism web blogs for recommending the alternative interesting tour programs.

3. Problems of Touristic Destination's Theme Determination

There are three problems: the multiple-EBT-themes determination problem, the EBT-EDU identification problem, and the sub-region/province names' ellipsis problem.

3.1 Multiple-EBT-Themes Determination Problem

Several people have several ideas of their own activities on their touristic traveling at the several destinations as shown in the following.

blog1:

EDU1 "ไปพักผ่อนที่สวนแม่ฟ้าหลวง/Take a rest at Mae Fah Laung Park." (***Ecotourism***)

EDU2 "ไปทานอาหารยูนนานที่ดอยแม่สลอง/Have You Nan Food at Doi Mae Slong." (***Culinary Tourism***)

EDU3 "ชมงานแสดงพื้นเมือง/Watch local entertainment." (***Cultural Tourism***)

EDU4 "เช้าวันต่อมาออกเดินทางไปที่เชียงใหม่/Next morning, go to Chiang Mai."

EDU5 "แวะเที่ยวชมวัดร่องขุ่น/Visit the Wat Long Khun temple" (***Historical Tourism***)

EDU6 "ไปทานข้าวซอยรสเด็ด/Have the delicious Khou Soi meal." (***Culinary Tourism***)

EDU7 "เดินช็อปปิ้งกันจนมืด/Go shopping until dark."

blog2:

EDU1 "มุ่งสู่วัดร่องขุ่นที่เชียงราย/Direct to the Wat Long Khun temple at Chiang Rai." (***Cultural Tourism***)

EDU2 "ตอนบ่ายแวะชมมหาวิทยาลัยแม่ฟ้าหลวง/In the afternoon, visit Mae Fah Laung University." (***Academic Tourism***)

EDU3 "รุ่งเช้าไปที่สวนแม่ฟ้าหลวง/Next morning, go to Mae Fah Laung Park." (***Ecotourism***)

EDU4 "ซื้อของพื้นเมือง/Buy some souvenirs."

blog3:

EDU1 "มาถึงเชียงใหม่ตอนเช้า/Arrive at Chiang Mai in the Morning."

EDU2 "ไปไหว้พระที่ดอยสุเทพ/Go to worship at the Doi Suthep mountain." (***Cultural Tourism***)

EDU3 "ตอนบ่ายเล่นน้ำสงกรานต์/In the afternoon, throw water in the Songkran Festival." (***Cultural Tourism***)

EDU4 "รุ่งขึ้นไปดูหลินปิง/Next day, go to see Lin Ping (Go to the zoo to see a panda name Lin Ping.)." (***Ecotourism***)

EDU5 "ตอนบ่ายเล่นน้ำสงกรานต์/In the afternoon, throw water in the Songkran Festival." (***Cultural Tourism***)

From blog1 to blog3, there are several tourism events within one region (consisting of several provinces) where to determine multiple tourism themes is challenge. Therefore, we propose applying the clustering technique with the verb and noun features gaining from the verb phrases of EDU event expression. The EDU event is clustered along with the province names of the selected region for determining the tourism themes.

3.2 EBT-EDU Identification Problem

How to identify the event based tourism EDU from each cluster is one of the major problems, as shown in the following example.

EDU1 "เดือนกันยายนพวกเราอยู่ที่ปราสาทพนมรุ้ง/ In September, we stayed in Phanom Rung Stone Castle."

EDU2 "ดูพระอาทิตย์ขึ้น/ [We] saw the sunrise."

EDU3 "เห็นแสงลอดประตูทั้ง15 บาน/ [We] saw the sun beam passing through 15 doors."

Where the […] symbol means ellipsis the content inside the square bracket. EDU2 and EDU3 are the events expressed by the action verbs. How to determine the EBT-EDU concept is using the verb-noun co-occurrence with the tourism event concept from the verb phrase annotation corpus.

3.3 Tourism Location Identification Problem

The tourism location identification is another major problem as shown in the following example.

blog title:เที่ยวจังหวัดเชียงราย

EDU1 "เช้ามืดสู่วนอุทยาน/Early in the morning, Direct to the park."

EDU2 "ทุกคนเดินขึ้นภูอย่างสนุกสนาน/Everyone hikes joyfully to the mountain." (Ecotourism)

EDU3 "สวยมากที่ยอดภู/It is very beautiful at the mountain submit."

From the example, there is no province name occurrence in EDU2, or EDU3 which is the problem of clustering. This problem can be solved by using the explicit title name.

4. A Framework for Touristic Destination's Theme Extraction

There are five steps in our framework. The first, corpus preparation step, includes NLP (Natural Language Processing) technique. The second is Verb-Noun Co-Occurrence Annotation with touristic concept. The third is the event EDU extractions that are expressed by the action verb. The fourth is the clustering step followed by the EBT theme determination step (see Figure 1).

4.1 Corpus Preparation

The documents used in our research are from the Thai tourism blog websites, (http://1081009.tourismthailand.org). The other source is from the tourism authority of Thailand, TAT, (http://thai.tourismthailand.org), which is used as the answer set for our system evaluation. These documents are related to 13 provinces of Thailand (Chiang Mai, Chiang Rai, Lampang, Lamphun, Mae Hong Son, Loei, Nan, Phra Nakhon Si Ayutthaya, Phayao, Phetchabun, Phitsanulok, Phrea, Sukhothai).

The corpus preparation involves using Thai word segmentation tools [7] (that include the name entity [8]). After the word segmentation is achieved, EDU segmentation is then dealt with [9]. These annotated EDUs are kept as an EDU corpus. The corpus contains 2000 EDUs separated into 2 equal parts; one part from TAT is used for manually annotating verb-noun co-occurrences with the EBT concept according to their explicit categories' occurrence on text. The other part from tourism blog is used for clustering and named this part as "Clustering Corpus".

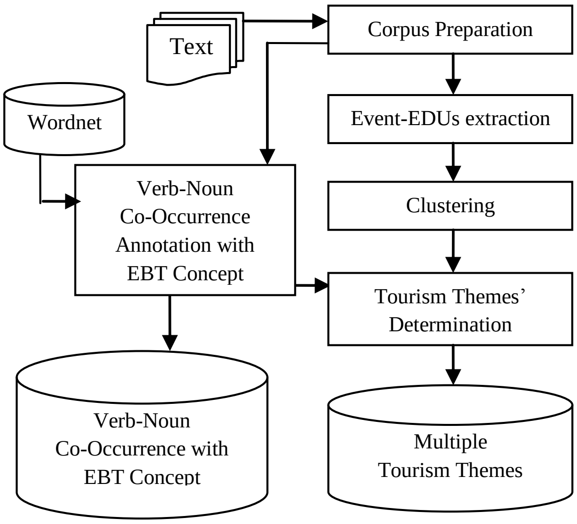
Text
Corpus Preparation
Wordnet
Event-EDUs extraction
Verb-Noun Co-Occurrence Annotation with EBT Concept
Clustering
Tourism Themes' Determination
Verb-Noun Co-Occurrence with EBT Concept
Multiple Tourism Themes

Figure 1. System Overview.

4.2 Verb-Noun Co-occurrence Annotation

All verb phrases of the EDU corpus from the previous step is used for manually annotating the EDU main verb and the noun right after the main verb as the verb-noun co-occurrences with the EBT concepts gaining from Wordnet and the Thai tourism encyclopedia, as shown in Figure 2. Then, these verb-noun co-occurrences can be summarized into Table1.

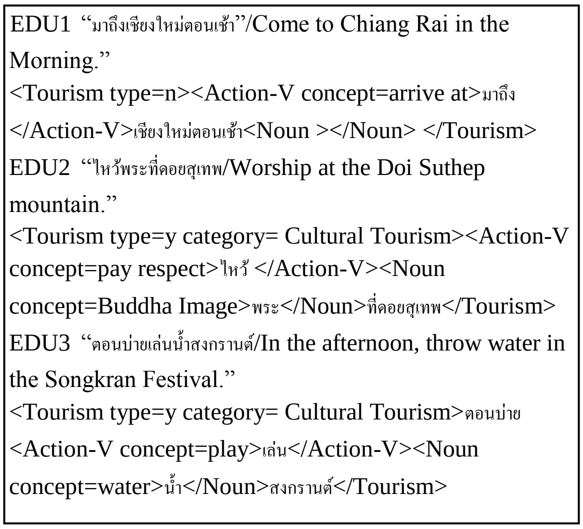
EDU1 "มาถึงเชียงใหม่ตอนเช้า"/Come to Chiang Rai in the Morning."
<Tourism type=n><Action-V concept=arrive at>มาถึง </Action-V>เชียงใหม่ตอนเช้า<Noun ></Noun> </Tourism>
EDU2 ""ไหว้พระที่ดอยสุเทพ/Worship at the Doi Suthep mountain."
<Tourism type=y category= Cultural Tourism><Action-V concept=pay respect>ไหว้ </Action-V><Noun concept=Buddha Image>พระ</Noun>ที่ดอยสุเทพ</Tourism>
EDU3 "ตอนบ่ายเล่นน้ำสงกรานต์/In the afternoon, throw water in the Songkran Festival."
<Tourism type=y category= Cultural Tourism>ตอนบ่าย <Action-V concept=play>เล่น</Action-V><Noun concept=water>น้ำ</Noun>สงกรานต์</Tourism>

Figure 2. Verb-Noun Co-occurrence Annotation with EBT Concept.

Table 1. Verb-Noun Co-Occurrences with tourism category.

Verb-Noun co-occurrence	Category
ชม/see-วัด/temple, … ไหว้/pay Respect-พระ/buddha, ชม/see-ศิลปะ/art, ...	Cultural
ชม/see-วิว/scenery, …	Ecotourism
ล่อง/float-แก่ง/islet, …	Adventure
รับประทาน/consume-อาหารพื้นเมือง/traditional food, …	Culinary
. . .	. . .

4.3 Event-EDUs Extraction

The objective of this step is to recognize and extract EDU events based on the action verb set, V_a (http://examples-help.org.uk/parts-of-peech/action-verbs.htm), from clustering corpus.

V_a = *{'ชม/see', 'ล่อง/float', 'รับประทาน/consume', 'ไหว้/pay Respect', 'ปีน/hike', ...}*

The province names' ellipsis has been solved by using the topic name containing the explicit province name.

```
Assume that each EDU is represented by (NP VP).
L is a list of EDU.
V_a is the tourism action verb concept set ($v_a \in V_a$)
N_t is the tourism noun concept set ($n_t \in N_t$)
EVENT_BASED _TOURISM__EXTRACTION (L, V_a, N_t)

   k ← 1, TOURACT_EDU← ∅
   while k ≤ length[L] do
     {   If ($v_{ak} \in V_a$) and ($n_{tk} \in N_t$)
         {        //identify the tourism activity EDU
           TOURACT_EDUS ← TOURACT_EDUS ∪ EDU_k;
           k=k+1;   }}
   return
```

Figure 3. Tourism Activity EDUs Extraction Algorithm.

4.4 Clustering

This step is to organize "provinces" instances into groups whose features are similar. To achieve the determination of multiple EBT themes, we propose using k-mean clustering technique [10] which is one of the simplest unsupervised learning algorithms through a certain number of clusters fixed a priori.

According to an EDU or a sentence consists of a noun phrase (NP) and a verb phrase (VP) as shown in the following regular expression.

$$\begin{aligned} &\text{EDU} \rightarrow \text{NP VP} \\ &\text{VP} \rightarrow V_a\ \text{NP1} \\ &\text{NP1} \rightarrow N_t \end{aligned}$$

Where V_a is the action verb concept set and N_t is the tourism noun concept set.

Given a set of province instances P={P_1, P_2, ⋯, P_z}, where each P is a real vector of v_a frequency, fv_a, and n_t frequency, fn_t, (where $v_a \in V_a$ and $n_t \in N_t$) as shown in the following.

P_j= {fv_{a1}, fv_{a2},..., fv_{ar}, fn_{t1}, fn_{t2},.. fn_{ts}} where j=1,2,…,z. r>1, s>1.

To normalize these features values into range of [0, 1], we use the equation as follow:

$$\delta = \frac{d - d_{\min}}{d_{\max} - d_{\min}} \tag{1}$$

Suppose the features values is in the range of [$d_{\min}$,$d_{\max}$] Where δ is the normalized value and d is the original value. The provinces matrix after normalization is shown in Table 2.

Table 2. The normalized provinces matrix.

Province	v_{a1}	v_{a2}	v_{a3}	…	n_{t1}	n_{t2}	n_{t3}	n_{t4}	...
Chiang Mai	0.80	0.60	0.50	…	0.70	0.40	0.70	0.80	...
Chiang Rai	0.83	0.75	0.58	…	0.83	0.58	0.58	0.75	...
Mae Hong Son	0.71	0.50	0.50	…	0.71	0.57	0.57	0.64	...
Nan	0.75	0.25	0.38	...	0.25	0.13	0.50	0.13	...
Phetchabun	0.78	0.22	0.22	...	0.44	0.33	0.11	0.11	...
Phrae	0.80	0.20	0.20	...	0.20	0.10	0.40	0.10	...
Sukhothai	0.62	0.38	0.77	...	0.08	0.08	1.00	0.23	...
…	...	...	...	...	...	...	...	...	...

v_{a1} = ชม/see
v_{a2} = แวะ/visit
v_{a3} = ไหว้/pay Respect

n_{t1} = วิว/scenery
n_{t2} = น้ำตก/waterfall
n_{t3} = วัด/temple
n_{t4} = เมือง/town

When performing k-means, the number of clusters is an input parameter. We use rule of thumb [11] as shown in (2) to determine number of clusters (k).

$$k \approx \left(\frac{z}{2}\right)^{\frac{1}{2}} \tag{2}$$

k-means clustering aims to partition the n provinces instances into k clusters, C = {C_1, C_2, ⋯, C_k}, so as to minimize the within-cluster sum of square (E) as below.

$$E = \sum_{i=1}^{k} \sum_{P_j \in C_i} \left\| P_j - m_i \right\|^2 \tag{3}$$

Where m_i is the mean of points in C_i. Given an initial set of k means $m_1^{(1)}$,…, $m_k^{(1)}$ as below, the algorithm proceeds by alternating between two steps. 1) Assign each P to the cluster with the closest mean as in (4). 2) Calculate the new means to be the centroid of cluster as in (5). Both steps 1) and 2) are iterated until the assignments do not change.

$$C_i^{(t)} = \{P_j : \left\| P_j - m_i^{(t)} \right\| \le \left\| P_j - m_{i^*}^{(t)} \right\| \text{ for all } i^* = 1,\ldots,k\} \quad (4)$$

$$m_i^{(t+1)} = \frac{1}{|C_i^{(t)}|} \sum_{P_j \in C_i^{(t)}} P_j \quad (5)$$

We use the software Weka (http://cs.waikato.ac.nz/ml/weka/) to perform clustering. The result is as follows (Table 3).

Table 3. Experiment results of 3 clusters.

Province	Cluster$_1$	Cluster$_2$	Cluster$_3$
Chiang Mai	1	0	0
Chiang Rai	1	0	0
Mae Hong Son	1	0	0
Nan	0	0	1
Phetchabun	0	0	1
Phrae	0	0	1
Sukhothai	0	1	0
…	…	…	…

4.5 Tourism Themes Determination

According to Table 2 and 3, each cluster centroid value based on the normalized value determined from feature frequencies in (1). Then, the high frequency feature yields the high cluster centroid feature value. Therefore, the high cluster centroid feature value (SV_a and SN_t), where $SV_a \subseteq V_a$ and $SN_t \subseteq N_t$, are used to determine the EBT theme of each cluster. In the experiment, we select sv_a and sn_t which value ≥ 0.5, as shown in Table 4.

Table 4. Features of cluster1 ranked by cluster centroid.

***sv*$_a$**	**Cluster Centroid**	***sn*$_t$**	**Cluster Centroid**
ชม/see	0.77	วิว/scenery	0.75
แวะ/visit	0.60	เมือง/town	0.75
ไหว้/pay Respect	0.54	วัด/temple	0.62
…	…	น้ำตก/wat□rfall	0.52
		…	…

The EBT theme of each cluster can be determined by matching the Cartesian products between SV_a and SN_t of each cluster to the verb-noun co-occurrences with EBT concepts, including the tourism categories, gained from the Verb-Noun Co-occurrence Annotation step, as shown in Table 5.

Table 5. Verb-Noun Co-occurrence as one EBT-Theme for one cluster.

Cluster Theme	Matched Verb-Noun Co-Occurrences with EBT Concepts	Category
1	ชม/see-วิว/Scenery, ชม/see-ดอกไม้/flower, …	Ecotourism
	ชม/see-เมือง/town, ชม/see-วัด/town ไหว้/pay Respect-พระ/buddha, …	Cultural
2	ชม/see โบราณสถาน/ancient monument, …	Cultural
3	ชม/see-ธรรมชาติ/nature, …	Ecotourism
	ชม/see-หมู่บ้าน/village, …	Cultural
	ล่อง/float-แก่ง/islet, …	Adventure

5. Evaluation and Conclusion

Each tourism theme determined from each cluster can be evaluated by matching its' tourism category result to the highlight/theme category of each province of the answer set, as shown in the following Table 6.

Table 6. Tourism theme evaluation results.

Answer Set			% Matches of Category
Cluster Theme	**Province Name**	**Highlight Category**	
1	Chiang Mai	Ecotourism, Cultural Tourism, Adventure Tourism	67%
	Chiang Rai	Cultural Tourism, Ecotourism	100%
	MaeHong Son	Cultural Tourism, Ecotourism	100%
2	Sukhothai	Cultural Tourism	100%
3	Nan	Adventure Tourism	33%
	Phrae	Ecotourism, Cultural Tourism	67%
	Phetchabun	Ecotourism, Adventure Tourism	67%

In conclusion, this research is able to successfully determine the multiple tourism themes based on the attractiveness event expressed by the action verbs for tourism destinations (cover several provinces) by applying clustering techniques along with verb-noun co-occurrences. Successful determination of multiple EBT themes can assist in GIS applications for recommending the alternative tourism theme choices. Furthermore, the results of this research can also conduct the knowledge of which province names that have the same tourism category.

References

[1] U.H. Graneheim and B. Lundman, "Qualitative content analysis in nursing research: concepts, procedures and measures to achieve trustworthiness," Nurse Education Today, vol. 24(2), pp. 105-112, 2004.

[2] L. Carlson, D. Marcu, and M.E. Okurowski, “Building a Discourse-Tagged Corpus in the Framework of Rhetorical Structure Theory,” In Current Directions in Discourse and Dialogue, pp. 85-112, 2003.

[3] S. Loh, F. Lorenz, R. Saldana and D. Licthnow, “A tourism recommender system based on collaboration and text analysis,” Journal of Information Technology & Tourism, vol. 6, pp. 157–165, 2004.

[4] Q. Mei, C. Zhai, “Discovering Evolutionary Theme Patterns from Text: an Exploration of Temporal Text Mining,” KDD '2005, 2005.

[5] Q. Hao, R. Cai, C. Wang, R. Xiao, J. Yang, Y. Pang, and L. Zhang, “Equip Tourists with Knowledge Mined from Travelogues,” WWW '2010, 2010.

[6] T. Kurashima, T. Tezuka. and Tanaka K, “Mining and Visualizing Local Experiences from Blog Entries,” DEXA’2006, 2006.

[7] S. Sudprasert and A. Kawtrakul, “Thai Word Segmentation based on Global and Local Unsupervised Learning,” NCSEC’2003, 2003.

[8] H. Chanlekha and A. Kawtrakul, “Thai Named Entity Extraction by incorporating Maximum Entropy Model with Simple Heuristic Information,” IJCNLP’2004, 2004.

[9] J. Chareonsuk, T. Sukvakree and A. Kawtrakul, “Elementary Discourse unit Segmentation for Thai using Discourse Cue and Syntactic Information,” NCSEC’2005, 2005.

[10] J. B. Macqueen, "Some Methods for classification and Analysis of Multivariate Observations," Proc. of 5th Berkeley Symposium on Mathematical Statistics and Probability, University of California Press, 1967.

[11] K. V. Mardia, J. T. Kent and J. M. Bibby, *Multivariate Analysis*, Academic Press, 1979.