Context-Aware Tourism Recommender System Using Temporal Ontology and Naïve Bayes

Chakkrit Snae Namahoot[1,*], Michael Brückner[2], and Naruepon Panawong[3]

[1] Department of Computer Science and Information Technology, Faculty of Science, Naresuan University, Phitsanulok, Thailand
chakkrits@nu.ac.th

[2] Department of Educational Technology and Communication, Faculty of Education, Naresuan University, Phitsanulok,Thailand
michaelb@nu.ac.th

[3] Department of Applied Science, Faculty of Science and Technology, Nakhon Sawan Rajabhat University, Nakhon Sawan, Thailand
jnaruepon.p@gmail.com

Abstract. In this paper, we present a Context Aware Thai Tourism Recommender System (CAT-TOURS) that applies a complex Naïve Bayes Model with boundary values, tourism ontology for Thailand and a temporal ontology to support decision making in tourism. Promising results are presented in the form of precision, recall and F measure for Websites related to Thailand's tourism industry. We compare the results with those gained with Latent Semantic Indexing (LSI).

This research was guided by the following aims: (1) find a simple method to classify Thai tourism Web documents that contain information on more than one topic, and (2) take into account time constraints in the process of making recommendations.

Keywords: Recommender System, Temporal ontology, Tourism, Naïve Bayes, LSI.

1 Introduction and Related Work

Currently, a vast number of Web sites provide services for finding travel information. However, tourists need to know more about touristic places and areas, e.g. relating attractions, hotels, dining, One Tambon One Product (OTOP, a government program to stimulate markets for local products) shops, and events. Most tourists search for interesting areas with a search engine, but the set of search results is often difficult to consume and confusing to understand because of the overload of information, mixed unwanted information, and uncategorized incoherent presentation. This leads to waste of time when extracting all relevant information and leads to inconvenient information gathering, even from a single information source.

---

[*] Corresponding author.

H. Unger et al. (eds.), *Recent Advances in Information and Communication Technology 2015,*
Advances in Intelligent Systems and Computing 361, DOI: 10.1007/978-3-319-19024-2_19

Automatic document clustering has gained increasing interest in the research community, especially in the course of the ever accelerating increase in the amount of information presented on Web pages in the World Wide Web. A review of machine learning for online document clustering of online news, blogs, e-mail, and digital library [1] refers to Rocchio's Algorithm [2], K-Nearest Neighbor [3], Naïve Bayes Algorithm (see, for example, [4]), and Support Vector Machine and states that Naïve Bayes is the best for e-mail, numeric and document clustering. Naïve Bayes also needs less data for the learning step and is not too complex. An additional advantage is that Naïve Bayes is convenient to develop when compared to other algorithms and is very fast [5].

When tourists seek travel recommendations, appropriate solutions have to take into account such contexts as time. An ontology-based recommender system should, therefore, support a temporal ontology. There are many representations of temporal ontologies, and some are listed in Table 1.

Table 1. Temporal ontology representations

Name	Specification	Source
SWRL Temporal Ontology	OWL GUI by Protégé	http://protege.cim3.net/cgi-bin/wiki.pl?SWRLTemporalOntology
W3C Temporal Ontology	OWL	http://www.w3.org/TR/owl-time/
RETR	Based on Allen's notation of events	[6]
CHRONOS	Handling of temporal ontologies with Protégé	[7]

However, not so many temporal ontologies deal with terms and appellations of non-Western calendars used in countries of Southeast Asia, among other regions in the world. There are some approaches based on the work of Eade [8], e.g. by Brückner [9], which can be applied not only to historic documents but also to contemporary events in the tourism market.

The tourism domain is not only well represented on the World Wide Web but has also gained increasing attention by semantic modeling experts. A number of ontologies have been created and published with different approaches and content (see [7], for a recent overview of tourism ontologies and their ranking following an Analytical Hierarchy Process). Some of these ontologies can be used to classify Web documents. This research was guided by the following aims: (1) to find a simple method to classify Thai tourism Web documents that contain information on more than one topic, and (2) to take into account time constraints in the process of making recommendations.

In this paper, we reduce the error in Web page clustering and apply sets of words, thereby introducing an improvement of the Naïve Bayes algorithm - not only by adding synonyms and related words into the clustering technique but also by allowing the categorization of a Web page into more than one group by adjusted boundary values of the Naïve Bayes algorithm. As an example, a Web page may contain (1) เที่ยว (*travel*), ภู (*hill*), เขา (*mountain*) categorized as Attraction; (2) อาหาร (*food*), ร้านอาหาร (*restaurant*), ทาน(*dine*), and กิน (*eat*) categorized as Dining; and (3) ที่พัก (*accommodation*), รีสอร์ท (*resort*) , นอน (*sleep*) categorized as Accommodation. Consequently, this Web page should be categorized into three categories: Attraction, Dining and Accommodation. In the following, we lay out the methodology of this research, report on the testing and results, and finally draw conclusions and indicate some directions of future work.

2 Methodology

The Context-Aware Thai Tourism Recommender System (CAT-TOURS) is divided into four parts: the novel (1) Naïve Bayes algorithm with Boundary values (NBB), (2) a tourism ontology focusing on Thailand, (3) a temporal ontology, and (4) the Tourism Recommender System. These parts will be elaborated on in the following.

2.1 Naïve Bayes with Boundary Values (NBB)

A Naïve Bayes classifier is a simple probabilistic classifier based on applying Bayes' Theorem with strong independence assumptions. These assumptions make the computation of Bayesian classification approach more efficient but this assumption severely limits its applicability. Depending on the precise nature of the probability model, the Naïve Bayes classifiers can be trained very efficiently by requiring a relatively small amount of training data to estimate the parameters that are necessary for classification [10].

In this paper, we propose an improvement of the Naïve Bayes algorithm called Naïve Bayes with Boundary values (NBB) for the classification of Web documents. We concentrated on touristic Web sites linked from Truehits, a DMOZ-like Web site in Thai language, and implemented the system as follows:

1. Apply Naïve Bayes algorithm for learning with 1,048 tourism Web pages from Truehits and six categories including Attraction (233 Web pages), Accommodation (200 Web pages), Dining (318 Web pages), Souvenir (54 Web pages), OTOP (88 Web pages), and Events (155 Web pages).
2. Improve the Naïve Bayes algorithm for a Thai tourism Web classification with increased efficiency.
3. Test of the Web document classification by applying 500 Web pages from Google search results with the NBB algorithm and measure the efficiency using F-Measure.
4. Apply NBB to CAT-TOURS.

The NBB algorithm is based on Eq. (1). Since the results of the calculation of the NB were negative in all categories and the frequency of queries to the group with the most negative results of calculations, we used the absolute value (positive values) to determine the Web category classification in Eq. (2).

$$C_{map} = \operatorname*{argmin}_{c \in C} \left(P(c) \sum_{k=1}^{n} \log \left(P(t_k | c)\right)\right) \tag{1}$$

$$C_{map} = \operatorname*{argmax}_{c \in C} \left(P(c) \sum_{k=1}^{n} \log \left(P(t_k | c)\right)\right) \tag{2}$$

$$C_{iMax} <= \text{threshold} >= C_{iMin},$$
$$C_{iMax} = \max(Cmap) \text{ and } C_{iMin} = \min(Cmap), \text{with } 1 <= i => 6 \tag{3}$$

where Cmap is the Web classification result and means the probability of the multiplied result by the probability P (c) with P (t_k | c) or briefly called probability of Naïve Bayes (P(NB)), Ci is probability of Attraction (At), Accommodation (Ac), Dining (D), Souvenir (S), OTOP (O) and Events (E). t_k is the word frequency, C is a category, P(c) is the probability of each of category and i is category counter 1 to 6 (i.e., Attraction, Accommodation, Dining, Souvenir, OTOP and Event). The improved Naïve Bayes algorithm with Boundary values (NBB) can be described as follows:

1. Initialize algorithm parameters
2. Used Web pages from test data set 1,048 Web pages from Truehits Web directory (by removing HTML Tags and choosing only web contents) and divide into six categories (Attractions, Accommodation, Dining, Souvenir, OTOP, and Events).
3. Calculate P (NB) of each category with Naïve Bayes algorithm (equation 1) using terms from tourism ontology([11], [12]) and many words from variation word and dictionary for tourism Web classification.
4. Define minimum (CiMin) and maximum (CiMax) of P (NB) of each category as threshold values using equation 2 and for Web classification automatically (Table 2).
5. Test 500 websites for classification by calculating P(NB)and compare P(NB) with CiMin and CiMax in Table 2 of every category using following condition:
   If P(NB) is between CiMin and CiMax of each category, then categorize; if not, then do not categorize
6. Repeat for all categories

Terminate after 500 websites and summarize Web classification results.

The NBB algorithm for Web classification can classify a Web page into more than one category, which is more in line with the diverse content on tourism information Web documents. This leads to an improvement of efficiency, see Table 3 and Table 4.

Table 2. Minimum and maximum boundary values for each category using Naïve Bayes (1,048 web documents)

Category	Minimum Boundary Value (Cmin)	Maximum Boundary Value (Cmax)
Attraction	0.08	20.06
Accommodation	0.08	4.35
Dining	0.13	14.97
Souvenir	0.02	0.47
OTOP	0.03	1.17
Events	0.05	4.21

2.2 Thai Tourism Ontology

Touristic spots have many different characteristics and attributes that have to be taken into account for the process of tourism recommendation system. For the tourist there are such criteria as cost, accommodation, attractions, restaurant, souvenir, traveling time or season and facilities that might be important for the purchasing recommendation. The data on tourism ontology comprise 10 classes including Province, Amphoe, Tambon, Event, Accommodation, Attraction, Restaurant, Souvenir, OTOP, and Transportation (Figure 1).

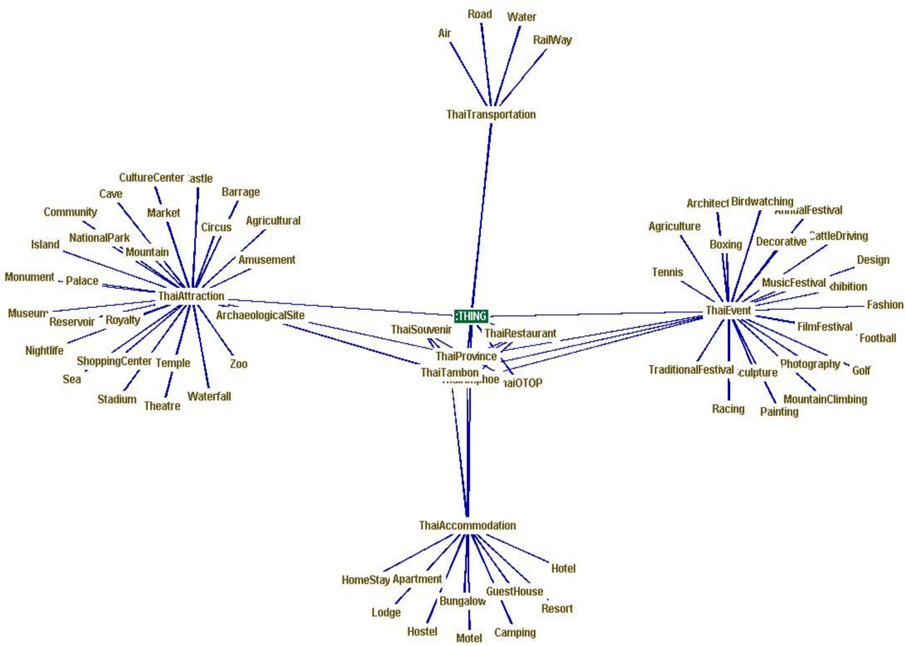

Fig. 1. Thai tourism ontology

The accommodation subclasses for Thailand that tourists can search for include Motel, Lodge, Resort, Guesthouse, Hotel, Hostel, Apartment, Camping, Home stay, and Bungalow. The event classes cover architecture, music festival, mountain climbing, sculpture, bird watching, decorative, racing, fashion, exhibition, photography, football, traditional festival, golf, design, annual festival, painting, film festival, tennis, boxing, agriculture and bull riding and trekking as sub classes. The attraction classes include mountain, waterfall, museum, cave, national park, circus, shopping center, barrage, theatre, reservoir, amusement, zoo, castle, nightlife, archaeological site, island, stadium, sea, community, royalty, temple, market, palace, culture center, monument and monastery (in Thailand mainly Buddhist temples).

2.3 Temporal Ontology

Information for seasons and time as well as important traditional Thai dates is something that tourists should consider, because they will not miss important events regarding their region of interest. Figure 2 shows the design of the Temporal Ontology that links the Attraction and Event (festivals) classes with the three main seasons in Thailand: summer, rainy season, and winter. When tourists input a travel time, the system checks traditional Thai events and attractions that tourists are recommended to visit during that time.

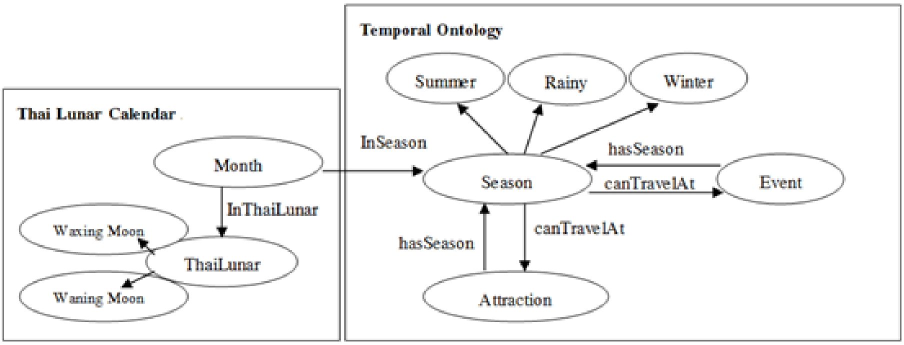

Fig. 2. Temporal ontology overview

There are three types of Thai lunar years:

1. Pokatimas Pokatiwan year (Pokatimas) which has 354 days
2. Pokatimas Atikawan year (Atikawan) which has 355days
3. Atikamas Pokatiwan year (Atikamas) which has 384days

Every year the Thai traditional dates are different from previous years. Rules for the calculation of the lunar calendar in the yearbook are based on Ok-Phansa day (End of The Buddhist Lent) which is a full moon day and is the 15th day of the 11th waxing moon of a Thai lunar calendar. If we know Ok-Phansa day of each year, then we can calculate other Buddhist days of each year as well. Therefore, we have implemented the Thai Lunar Calendar Algorithm to calculate Thai Lunar (Waxing moon and Waning moon) dates automatically to get exactly the Thai Buddhist Days described in the following.

The Thai Lunar Calendar algorithm:

1. Input: Tourist travel date e.g. 25/11/2015
2. take the year from 1 (2015 in this case)
3. identify types of the Thai lunar year above (Pokatimas, Atikawan and Atikamas) by calculating the number of days in the input year (from tourist) using following formula [14].

$$days = \sum_{i=1}^{12} dm_i + d_{in} - d_{pre} \quad (4)$$

where days is number of days in the year of input year, dm_i is number of days in each month of the input year, i is the number of month. (1-12),d_{in} is the number of the Ok Phansa (òk pan-săa) day in the input year, and d_{pre} is the number of the Ok Phansa (òk pan-săa) day in the previous year of the input year. For example, the Ok Phansa day (15th day of the 11th waxing moon) for input year and previous year are 27th October 2015 and 8th October 2014 respectively. Thus, the type of Thai lunar year is 365+27-8 = 384 which can be identified as Atikamas year using the following conditions

   3.1. If days equal to 354, then identify as the Pokatimas year
   3.2. If days equal to 355, then identify as the Atikawan year
   3.3. If days equal to 384, then identify as the Atikamas year

4. Find other Buddhist days from Ok-Phansa day using following conditions
   4.1. increase number of month of Ok-Phansa day (October, i=10) by 1 (i of Ok-Phansa day +1 which is)until i is equal to 12
   4.2. decrease number of month of Ok-Phansa day by 1 (i-1) until i equal to 1
   4.3. find the 15th day of the waxing moon and the 1st day of the waning moon
      4.3.1. increase i by 1, we find Loy Krathong day (15th day of the 12th waxing moon)which is 25th November 2015
      4.3.2. decrease i by 3 we find Asalha Puja day (15th day of the 8th waxing moon) which is 30th July 2015 and Buddhist lent day (1st day of the 8th waning moon), which is 31th July 2015.
      4.3.3. decrease i by 4 we find Visakha Puja day (15th day of the 7th waxing moon), which is 1st June 2015.
      4.3.4. decrease i by 6 we find Makha Puja Day (15th day of the 3th waxing moon) which is 4th March 2015
5. Display all Thai lunar dates according to Buddhist days.
6. Compare results from 5 and if the input day and month are exact day and month or even close to Buddhist days then recommend tourist for traveling
7. Display results and terminate.

3 Tourism Recommender System Architecture

Figure 3 shows the architecture of CAT-TOURS (Context-Aware Tourism Recommendation System), which can be described as follows:

The Temporal Ontology helps find the Thai festivals during the time of the year. NBB enables to find more information regarding six categories as accommodation, restaurant, attraction, etc. The system can help save time by finding various pieces of information and dismissing unwanted data.

1. Tourists choose travelling dates, and the system checks with temporal ontology whether or not the dates match with important Thai lunar dates
2. Users provide words or a sentence, and the system checks whether they are related to a province or other tourist spot, in which case the system crawls relevant Web documents from Google related to the tourist information of the province/spot and stores them in the database to prepare the Web document classification.
3. Classify those websites using NBB and show the results of classification of travel sites in six categories: Attractions, Accommodation, Dining, Souvenir, OTOP, and Events.
4. From 2, if it is not a province the system checks the input data and search for more detail of the region concerned for identifying a name of the province related to input data and then shows fundamental information of this province using tourism ontology pattern that are stored in OWL.
5. The system recommends tourist information based on the time and season applying the temporal ontology.

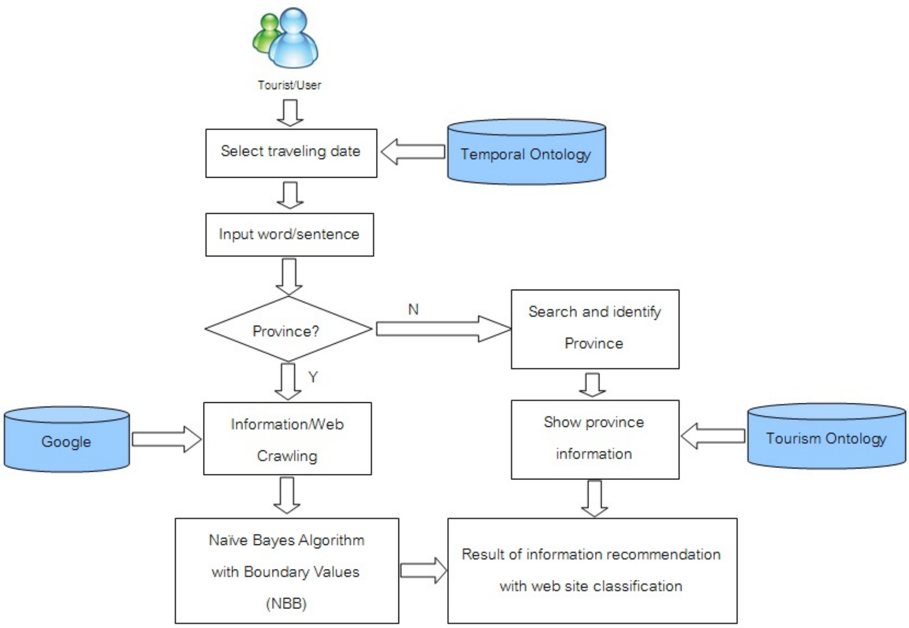

Fig. 3. CAT-TOURS system architecture

4 Testing and Results

4.1 Classification Results

The results relating CAT-TOURS are divided into two parts:

- Tourist websites classification results, and
- The interface of the system showing results suggesting travel sites based on the classification algorithm of NBB, result for travel information by time or season (Thai lunar calendar).

We used 500 tourism Web documents for classification and calculated the F-Measure for measuring the efficiency of NBB by Equation 5, for which results are shown in Table 3.

$$F = 2 * \left(\frac{P * R}{P + R} \right) \tag{5}$$

P is True Positive/(True Positive + False Positive)
R is True Positive/(True Positive + False Negative)
True Positive is web site in category and system was in that category
False Positive is web site not in category and system was in that category
False Negative is web site in category and system was not in that category

Another rather mature approach for automated document classification is the Latent Semantic Indexing (LSI), which uses a strictly mathematical method and is, therefore, independent of the natural languages involved.

Table 3 shows a comparison of tourism Web classification using LSI, Naïve Bayes (NB) and Naïve Bayes with boundary values (NBB). Precision P, recall R and F-Measure F have been calculated. LSI and NB algorithms result in 74.37% and 78.25% for precision, 62.88% and 70.04% for recall, 67.01% and 72.51% for F-Measure, respectively, because LSI and NB algorithms classify the Web documents into one category. The LSI algorithm produces some incorrect site classifications, such as those in the category Events. (41.83% accuracy), but classifies in the category of tourist or some other category, due to the content of those pages not being consistent with the frequency of the actual content. NB algorithm only uses P (NB) value and not so much the frequency of the search term in this category, which results the low efficiency of classification, especially in categories of Souvenir (Sou) and Event (Even) with accuracy 55.59% and 64.21%, respectively. This may be due to websites with rich content on the various categories such as a web site about the attraction but including other contents or categories as hotels and restaurants. However, the frequency of words appearing in the section of the restaurant is higher than attraction section. As a result, the classification does not correspond to the attraction website content rather it is classified in the restaurant category. NBB algorithm leads to 100% for precision, 97% approximately for F-Measure, and Web documents can correctly be in more than one category (Table 3). Moreover, the F-Measure data show that the NBB algorithm is more efficient in classification Web documents for Thailand tourism information.

Table 3. Performance of the classification algorithms

Category	LSI			NB			NBB		
	P	R	F	P	R	F	P	R	F
Att	81.82	57.91	67.82	70.42	88.57	78.46	100	95.71	97.81
Acc	87.76	54.34	67.12	94.76	63.21	75.83	100	100	100
Res	71.87	96.18	82.27	83.27	91.25	87.08	100	100	100
Souv	82.58	66.72	73.81	55.62	55.57	55.59	100	82.01	90.12
OTOP	80.34	63.23	70.77	100	58.61	73.90	100	95.34	97.61
Even	41.83	38.88	40.30	65.45	63.01	64.21	100	93.54	96.66
Average	74.37	62.88	67.01	78.25	70.04	72.51	100	94.43	97.03

4.2 System Testing

Figure 4 shows the interface of Tourism Recommendation System for Thailand in English. A user types in “travel in Phitsanulok” in English (or in Thai for the Thai Interface) and enters a travel date 03-11-2557 (3th November, 2014).The system shows basic information about Phitsanulok Province on the left screen. On the right screen, six tourist information categories ranked by popularity with URL websites are displayed. On the bottom of the screen, traveling recommendation based on Thai lunar calendar is shown, which mentions correctly the Pakthong Chai tradition, the Loy Krathong Festival and Buddha Day on 5th, 6th and 14th November 2014, respectively.

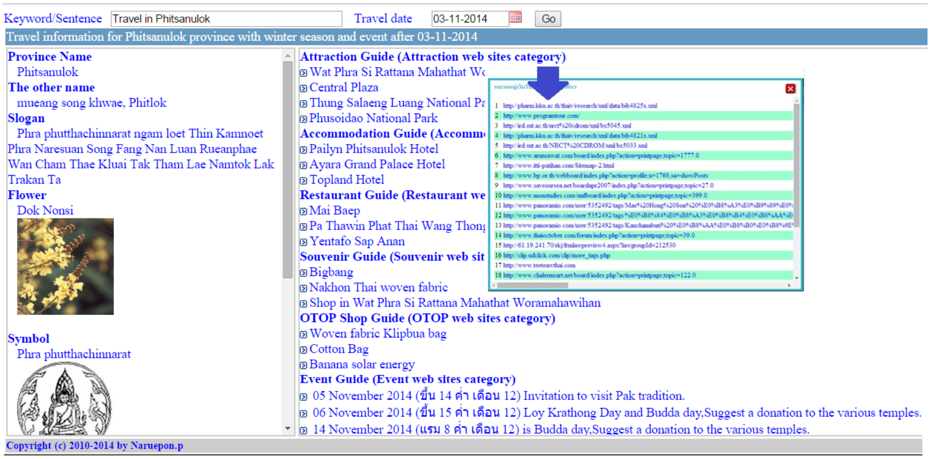

Fig. 4. Result example of Tourism Recommendation System in Thailand English version

5 Conclusion and Further Work

This paper introduces the Context Aware Thai Tourism Recommendation System (CAT-TOURS) with a Temporal Ontology and a novel Naïve-Bayes Algorithm with Boundary values (NBB). The NBB exhibits increased efficiency, which has been tested for Thailand tourism Website classification. We used tourism Web pages as a learning data set and defined thresholds based on minimum and maximum boundary values for probabilities for each of the six categories used in the classification testing with LSI, NB, and NBB algorithms. The result shows that NBB algorithm is the most efficient resulting in 100% for precision, 94.43% for recall and 97.03% for F-Measure. We have implemented the Temporal Ontology which gave additional details for attractions, events and festivals that tourists should visit during the time of their stay. Tourists specify only the province and the travelling dates, which results in the system suggesting where and when special and important traditional events are located near the destination.

This research also has some limitations. Due to the location-unaware setting via conventional Internet access, we did not consider the user's current location as a parameter for recommendations. This is planned for one of the next rollouts of CAT-TOURS, which will be based on a similar road segment ontology as described in [14]. Moreover, we will allow users to choose types of tourism, accommodation and foods by taking into account gender, age, and the number and duration of travels. Then, the system will be able to use the data and compare them with the rules of the tourism ontology based on SWRL, which is also stored in the data model. Finally, the system calculates the travel recommendation to meet the needs of users with the help of data mining techniques.

Another limitation is the rudimentary usability test of CAT-TOURS. However, we plan to set up a similar test suite as has been described by Zins et al. [15].

References

1. Ceci, M.: Classifying Web documents in a hierarchy of categories: a comprehensive study. Journal of Intelligent Information Systems 28, 37–78 (2007)
2. Rocchio, J.: Relevance feedback in information retrieval. In: The SMART Retrieval System: Experiments in Automatic Document Processing, pp. 313–323. Prentice Hall, Englewood Cliffs (1971)
3. Groza, A., Dragoste, I., Sincai, I., Jimborean, I., Moraru, V.: An ontology selec-tion and ranking system based on the Analytical Hierarchy Process. In: The 16th International Symposium on Symbolic and Numerical Algorithms for Scientific Computing, Timisoara, Romania (2014)
4. McCallum, A., Nigam, K.: A Comparison of Event Models for Naïve Bayes Text Classification. Journal of Machine Learning Research 3, 1265–1287 (2003)
5. Beineke, P., Hastie, T., Vaithyanathan, T.: The sentimental factor: improving review classification via human-provided information. In: Proceedings of the 42nd Annual Meeting of Association for Computational Linguistics (ACL 2004), Barcelona, Spain. Association for Computational Linguistics (July 2004) (2014), doi:10.3115/1218955.1218989

6. Hemalatha, M., Uma, V., Aghila, G.: Time ontology with Reference Event based Temporal Relations (RETR). International Journal of Web and Semantic Technology 3(1), 23–31 (2012)
7. Perventis, A.P., Polyxeni Marki, E., Petrakis, G.M., Sotirios, B.: CHRONOS: A Tool for Handling Temporal Ontologies in Protégé. In: IEEE 24th International Conference on Tools with Artificial Intelligence, pp. 460–467 (2012)
8. Eade, J.C.: The Calendrical Systems of South-East Asia. Leiden, Brill (1995) ISBN 9789004104372
9. Brückner, M.: Multilingual semantic modeling for cultural heritage organizations in ASEAN. ACM Journal of Computing and Cultural Heritage (in press, 2015)
10. Khan, A., Baharudin, B., Lee, L.H., Khan, K.: A review of machine learning algorithms for text-documents classification. Journal of Advances in Information Technology 1(1), 4–20 (2010)
11. Panawong, N., Snae Namahoot, C.: Performance Analysis of an Ontology-Based Tourism Information System with ISG Algorithm and Name Variation Matching. Naresuan University Science Journal 9(2), 47–64 (2013) (in Thai)
12. Panawong, N., Snae, C.: Search System for Attractions in Thailand with Ontolo-gy and Name Matching. Journal of Information Science and Technology 1(2), 60–69 (2010) (in Thai)
13. Siri, C.: Auspicious Thai calendar. Bangkok, Kurusapa Printing Ladphrao (2006)
14. Sadeghi Niaraki, A., Kim, K.: Ontology based personalized route planning system using a multi-criteria decision making approach. Expert Systems with Applications 36, 2250–2259 (2007)
15. Zins, A.H., Bauernfeind, U., Messier, F.D., Venturini, A., Rumertshofer, H.: An experimental usability test for different destination recommender systems. In: Information and Communication Technologies in Tourism (ENTER), Cairo, Egypt, pp. 228–238. Springer (2004)