Journal of Computers Vol. 30 No. 2, 2019, pp. 116-125
doi:10.3966/199115992019043002010

A Thailand Tourism Web Analysis and Clustering Tool Using a Word Weight Calculation Algorithm

Chakkrit Snae Namahoot[1*], Desmond Lobo[2]

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

[2] School of Information and Communications Technology, Faculty of Applied Science and Engineering Technology, Seneca College, Toronto, Ontario, Canada
desmond.lobo@senecacollege.ca

Received 15 December 2018; Revised 15 February 2019; Accepted 15 February 2019

Abstract. The result of searching for tourism material with a search engine typically ends up with an overload of information. These results are often presented in an uncategorized and incoherent manner since most travel sites are not classified. This causes difficulties in searching as well as a lot of time wasted extracting the relevant information. It also leads to inconvenient information gathering, even from a single information source. In this study, the researchers resolved these issues by developing a Word Weight Calculation Algorithm (WWCA). The algorithm calculates weights of words and are applied not only for the analysis of Thai travel websites, but also for their classification using four categories of tourist information: attractions, accommodation, restaurants and souvenir shops. Parts of the website HTML structure were extracted and used for the analysis and classification of 800 Thai tourist websites. The results of the WWCA were measured in terms of its efficiency using the F-measure statistic. The outcomes showed that (1) the content within the HTML body tag alone is sufficient to classify the sites and (2) the WWCA was a good indicator for the Thai travel websites classification.

Keywords: algorithms, metadata, text mining, web clustering, website analysis

1 Introduction

In this study, the researchers revised the text clustering technique for classifying Thai tourism websites that had been developed by Namahoot et al. in 2014 [1]. The authors had developed algorithms for the classification of travel sites with a Thai text analysis technique. There have been other researchers that have conducted similar research. Manikandan and Sivakumar [2] classified and clustered e-documents, online news, blogs, e-mails and digital libraries. They provided a review of the principles, advantages and applications of document classification, clustering and text mining.

Lee et al. [3] applied a latent factor model to generate clustered text features and used them for text classification. The main contribution of their research was to extend the Latent Dirichlet Allocation (LDA) by considering word weights in sampling and maintaining balances of topic distributions. A series of experiments were used to confirm that the topic distributions generated by the balance weighted topic modeling method added some discriminative power to feature generations for classification.

Luss and d'Aspremont [4] predicted intraday price movements of financial assets using text from news articles. They used multiple kernel learning to combine equity returns with text as predictive features to increase classification performance. They then developed an efficient method for analytic center cutting to solve the kernel learning problem efficiently. This produced significantly better performance than historical returns alone.

* Corresponding Author

Allahyari and his colleagues [5] focused on text mining and extracting meaningful data from text. Because the tremendous volume of mostly unstructured text cannot simply be processed by computers, they noted that efficient and effective techniques and algorithms are required to discover useful patterns. In their paper, they described several of the most fundamental text mining tasks and techniques including text pre-processing, classification and clustering, especially in the domains of biomedical and health care.

In 2017, Sammouda [6] focused on Arabic text classification with applications in text mining, web search, social media, security, and other fields. His research aim was to compare supervised learning methods on Arabic text classification. His results confirmed that the accuracy of the 10-fold cross validation test mode obtained by the Lagrangian support vector machine algorithm using the inverse document frequency transform technique was the most efficient compared to other supervised learning methods.

Nogueira el al. [7] compared the performance of the Naïve Bayes algorithm with one for grouping documents. They utilized the WEKA program and found that the Naïve Bayes algorithm yielded the best results. However, those results were not able to show the comparison efficiency and promptness. Sathapornvajana [8], on the other hand, presented a technique to classify tourists' interests in Thailand using data mining and ontology techniques. He focused on five domains: attractions, accommodation, travels, events, and festivals. Similarly, Chatcharaporn and his colleagues [9] used a group of five different domains from the Lonely Planet website that were categorized into groups: accommodation, night life, restaurants, shopping sites, and attractions. The efficiency of the decision tree, Naïve Bayes, and support vector machine techniques were compared and analyzed. The results showed that Naïve Bayes generated information with the highest accuracy, but only for the English language.

Chotirat et al. [10] designed and developed a knowledge-based ontology for automatically analyzing online news and for the classification of sites using ontology as a tool. Initially, they used 200 online news articles based on the Deep South as a subject and then grouped the key words that were related. They determined the frequency of words appearing in the news by analyzing the key words that could potentially affect the online news.

2 Methodology

There were several steps in this research process.

2.1 Tourism Word Collection

In the analysis of Thai tourism websites, related words can be classified into four categories: attractions, restaurants, accommodation, and souvenir shops. These words were derived from the tourism ontology developed by Namahoot et al. [11-12].

To cover all of the words in the content of tourism websites, some words that occur frequently in a sample of websites were also extracted and collected for the analysis. Words used in this research were divided into the following categories:

- Words in the attractions category include travel, tour, scenic, festival, holiday, cave, waterfalls, mountains, sea, culture, old town, rafting, island, show, bird watching, mall, photography, tradition, exhibition, etc.
- Words in the restaurants category include food, eat, taste, drink, delicious, salad, toast, grilled, fried, boiled, rice, curry, bakery, taste, steak, BBQ, suki, cake, wine, fast food, sandwich, vegetarian, etc.
- Words in the accommodation category include hotels, motels, bungalows, homestays, hostels, lodgings, rooms, sleep, accommodation, deluxe, townhouse, house, garden house, mansion, resort, spa, etc.
- Words in the souvenir shops category include shop, OTOP (One Tumbon One Product), fabric, community, souvenirs, bags, decoration, herbs, wisdom, district, product, shop, etc.

2.2 URL Collection and HTML Structure Data Extraction of Travel Websites

The next step consisted of the collection of 400 travel websites from the Tourism Authority of Thailand (TAT) and travel sites (travel.truehits.net). Only four categories were selected and each of these were further divided into 100 websites. These were to be used for learning to determine the appropriate boundaries or threshold values in each category.

To extract the HTML structure data from each URL, the contents of each of 200 web pages were retrieved and dissected into the web HTML structure of the site. This consisted of five parts: Title (T), Description (D), Keywords (K), Body (B), and Links (L). Next, all the structured data were stored into a MySQL database with the following format:

- ID is the sequence of website URLs.
- URL is the universal resource locator of the website.
- GRP is the category (1 for attractions, 2 for accommodation, 3 for restaurants, and 4 for souvenir shops)
- Title is the HTML structure of the website in the Title section.
- Description is the HTML structure of the website in the description section.
- Keywords is the HTML structure of the website in the keywords section.
- Body is the HTML structure of the website in the body.
- Links are the links which are connected to the other webpages. The structure information within the links can be stored in the four structures: T, D, K and B of the Sub Link.

2.3 Analysis and Classification Algorithm

In this step, the Word Weight Calculation Algorithm (WWCA) was developed and the process can be described as follows:

- Apply string matching for matching words in travel websites of the four categories to find the frequency of each word.
- Compute the word weight ($WW_{c,s}$) values using equation (1).
- Return the results of the word weight values of each structure in each category (Table 1).

Table 1. A sample of words weight values of each HTML structure from each of the four categories

Category	Word	T	K	D	B
Attractions	เทยว (travel)	0.69	0.92	0.84	1.00
	ภ (hill)	0.03	0.21	0.14	0.94
	เขา (mountain)	0.03	0.29	0.00	0.86
	ทะเล (sea)	0.08	0.45	0.10	0.82
	วด (temple)	0.00	0.08	0.05	0.82
Accommodation	พก (stay)	0.68	0.70	0.74	0.98
	ทพก (accommodation)	0.63	0.70	0.61	0.84
	รสอรท (resort)	0.32	0.53	0.42	0.82
	โรงแรม (hotel)	0.39	0.63	0.45	0.78
	หองพก (room)	0.08	0.10	0.13	0.78
Restaurants	อาหาร (food)	0.73	0.77	0.89	0.98
	รานอาหาร (restaurant)	0.73	0.77	0.82	0.96
	กน (eat)	0.59	0.41	0.45	0.82
	อรอย (delicious)	0.30	0.38	0.50	0.76
	ดม (drink)	0.30	0.13	0.18	0.72
Souvenir Shops	OTOP	0.51	0.25	0.21	0.72
	ผลตภณฑ (product)	0.03	0.00	0.11	0.72
	ขาย (sell)	0.05	0.08	0.32	0.66
	ราน (shop)	0.13	0.25	0.37	0.64
	ของฝาก (souvenir)	0.15	0.17	0.26	0.40

- Choose the HTML structure that gives the best result of analysis, which is R_b, according to the pre-process testing and Namahoot et al. [1].
- Find the optimal threshold in each category by learning and determining the threshold value of each category, based on the R_b values and the F-measure.
- Compute the rating of all HTML structures in each category (R_t, R_b, R_k, R_d and R_l) and R_{web}.

- Analyze and classify Thai tourism websites using R_{web}, with an optimal threshold in each category.
- Analyze which HTML tags of the structure in the body that the most common words related to tourism appeared in each category.

The WWCA algorithm uses the weights of words that separate each section of the website's HTML structure. Therefore, each structure of a site that is the same category will have a different weight of words based on the structure of the site. Equation (1) was used to calculate the rating based on WWCA.

$$WW_{c,s} = N_w / N. \quad \textbf{(1)}$$

In this equation, $WW_{c,s}$ represents the word weight value, c represents one of the four categories, s represents one of the five website HTML structures, N_w is the number of sample websites in each category containing tourism words related to the category, and N is the number of sample tourism websites in each category.

For example, to calculate the weight of the word "travel" of the site in the body structure in the category of attractions of 100 sites, N = 100. Because the body structure found the word "travel" in 100 websites, $N_w = 100$. $WW_{c,s}$ can be calculated using equation (1) as follows.

$WW_{c,s}$ = Word Weight (WW) values for HTML structure(s) of the body in the category (c) of attraction is 100/100 = 1.00.

Hence, based on the WWCA, the weight of words was considered according to the site structure. The weight of the word "travel" in the category of attractions of the body (B) structure is 1.00 (Table 1).

Table 1 shows an example of the rating of words R_{cs} which is calculated using equation (1) for each of the four categories and for each of the HTML structures of the website. As can be seen, R_{cs} from any of the categories in the body web structure has maximum values. This is because the content in the body web structure has many words related to tourism information.

The tourism web analysis and clustering process uses equation (1) as WW_i in equations (3)-(7) with the five parts of the web HTML structure and can be used to analyze tourism categories using equation (2).

$$R_{web} = \frac{R_t + R_b + R_d + R_l}{N_s}. \quad \textbf{(2)}$$

$$R_t = \sum_{i=1}^{p} \frac{tWW_i}{p}. \quad \textbf{(3)}$$

$$R_b = \sum_{i=1}^{q} \frac{bWW_i}{q}. \quad \textbf{(4)}$$

$$R_k = \sum_{i=1}^{c} \frac{kWW_i}{c}. \quad \textbf{(5)}$$

$$R_d = \sum_{i=1}^{r} \frac{dWW_i}{r}. \quad \textbf{(6)}$$

$$R_l = \sum_{i=1}^{l} xyuvw(lWW_i). \quad \textbf{(7)}$$

In equations (3)-(7),

- R_{web} represents the rating of travel websites in each of the four categories and is calculated using equation (2).
- R_t is the rating value of each category in the HTML title tag of the default web page.
- R_b represents the rating value of each category in the HTML body tag of the default web page.
- R_k represents the rating value of each category in the HTML meta-name keyword of the default web

page and is calculated using equation (5).
- R_d represents the rating value of each category in the HTML meta-name description the default web page and is calculated using equation (6).
- R_l represents the rating value of each category in the HTML link within the Web page to other pages and is calculated using equation (7).
- N_s represents the number of structures used in the calculation of R_{web}.
- tR_i represents the probability of each category of the word and is the i^{th} individual word found in the HTML title tag of the website.
- p represents the number of words that appeared in each category in the HTML title tag of the default web page.
- bWW_i represents the rating value of the words in each category and in each division of the i^{th} individual word found in the HTML body tag of the pages of the site.
- q represents the number of words that appeared in each category in the HTML body tag of the default web page.
- kWW_i represents the rating value of the words in each category and in each division of the i^{th} individual word found on the web pages: the HTML meta-name keyword at the top of the website.
- c represents the number of words that appeared in each category in the HTML meta-name keywords of the default web page.
- dWW_i represents the rating value of the words in each category and in each division of the i^{th} individual word found on web pages: the HTML meta-name description at the top of the website.
- r represents the number of words that appeared in each category in the HTML meta-name description of the default web page.
- $t'WW_i$ represents the rating value of each category of words in each of the i^{th} individual word found in the HTML title tag of the web page that links to other pages.
- x represents the number of words found in each category in the HTML title tag of HTML links within the web page to other pages.
- $b'WW_i$ represents the rating value of each category of words in each of the i^{th} individual word found in the HTML body tag of a web page that links to other pages.
- y represents the number of words appeared in each category in the HTML body tag of the HTML Links within the web page to other pages.
- $k'WW_i$ represents the rating value of each category of words in each of the i^{th} individual words found in the web page HTML meta name keyword side links from all HTML links in a Web page to get started.
- u represents the number of words appeared in each category in the HTML meta-name keyword of the HTML links within the web page to other pages.
- $d'WW_i$ represents the rating value of each category of words in each of the i^{th} individual words found on web pages: the HTML meta name description of all HTML links to other pages.
- v represents the number of words found in each category in the HTML meta-name description of the HTML links within the web page to other pages.
- z represents the number of web pages that can be linked from the default web page of the site through HTML links.

2.4 Optimal Threshold Determination Process

This section describes the process of determining the optimal threshold for categorizing the tourism website in each category using R_b. The reason that we use only R_b instead of R_{web} is because the time of the whole process of analysis testing can be reduced and this makes the analysis process faster. In addition, according to our aim and the previous research (Namahoot et al. [1]) it was noted that the highest performance value was the HTML body tag. Hence, the content within the body tag alone could be used to classify the sites. The following process describes how to determine the optimal threshold for website classification.
- Take 400 travel websites that have already been categorized into 4 categories, 100 each for learning to find the optimal boundary (threshold) in each category.

- Calculate the R_b value (equation 4) in each category to find the minimum values that can be categorized in each category correctly
- Calculate the R_b value to find the maximum value that can categorize all categories (400 websites across all categories) correctly using equation 8 ($R_{c1,b\text{-}all}$)
- Define coarse ranges from minimum to maximum values in each category of R_b and $R_{c1,b\text{-}all}$ above and then compute the accuracy based on each value of coarse ranges using the F-measure statistic
- Select and set up the optimal threshold values in each category based on the minimum values of coarse ranges that give the best accuracy (100% or nearest by F-measure)
- Use the optimal threshold values in each category to categorize 800 travel websites (website of 400 learning sites and the 400 new sites tested for categorization) and 31 classification patterns of HTML structures. For website classification, use the following classification rule: if the R_{web} of the site is greater than or equal to the optimal values of each category (maximum value of $R_{c1,\ b\text{-}all}$), then that site will fall into the category, otherwise no classification.
- Display the results of 31 classification patterns of HTML structures and website classification with F-measure (equation 9)

$$R_{c1,b-all} = \sum_{i=1}^{qc1} \frac{bWW_i}{qc1 + qc2 + qc3 + qc4}. \quad \textbf{(8)}$$

In equation (8), $R_{c1,b\text{-}all}$ is the rating value of category 1 (can be any category) in the HTML body (b) tag of the default Web page. qc1, qc2, qc3 and qc4 are the number of words that appeared in all categories (categories 1-4 can be any of these: attractions, accommodation, restaurants, and souvenir shops) in the HTML body tag of the default Web page. For example, from equation (8), if the rating value of attraction category ($R_{c1,b\text{-}all}$) is calculated, first count the number of words that appeared in category of attractions (qc1) and then find bWWi using word weight values in Table 1. After that, count the number of any words that appeared in categories 2, 3, and 4 and sum up with qc1 as a divider. Finally, return the value of $R_{c1,b\text{-}all}$ for further analysis.

$$F - measure = \frac{2 * precision * recall}{precision + recall}. \quad (\textbf{9})$$

$$Precision = True\ Positive\ /\ (True\ Positive + False\ Positive)$$
$$Recall = True\ Positive\ /\ (True\ Positive + False\ Negative)$$

True Positive refers to the number of websites in that category and the algorithm analyzes in the correct category.

False Positive refers to the number of websites that are not in that category but the algorithm analyzes those website in that category. False Negative refers to the number of websites in that category but the algorithm analyzes those websites not in that category.

Table 2. Shows an example of the optimal boundary value from learning in each category

Category	Min of R_{web}	Max of $R_{c1,b\text{-}all}$	Course Range	Web Analysis
				F-measure
Attractions	0.59	0.45	0.55	100%
			0.50	100%
			0.45	100%
Accommodations	0.64	0.57	0.63	100%
			0.61	100%
			0.59	100%
			0.57	100%
Restaurants	0.55	0.54	0.55	100%
			0.54	100%
Souvenir Shops	0.46	0.46	0.46	100%

From Table 2 the optimal threshold of category attractions, accommodation, restaurants and souvenir shops is 0.45, 0.57, 0.54 and 0.46, respectively. These threshold values are used (1) to analyze which

pattern of HTML structure using WWCA to provide the best performance and (2) to categorizing Thai tourist websites. The results are shown in Section 3.

2.5 HTML Structure of the Body Analysis

This section presents the analysis of the most common tourist terms/words found in the HTML structure (HTML tags) of the body. The results of the previous research and pre-testing process found that the classification of the HTML structure of body provides the best performance. A further research has been investigated into which parts (HTML Tag) of the HTML structure of the body that the words related to travel site can be frequently found the most (results can be found in Table 5). The HTML tags are divided and examined as follows.

- UL (Unordered List) is an unordered list.
- OL (Ordered list) is ranked.
- H1-H6 (Heading1-Heading6) is the topic where H1 is the most important topic and H6 is the least important topic.
- P (Paragraph) is a new paragraph statement.
- TD (Table Data) is the text in the table in each field
- Strong is to emphasize messages that are more important than other messages.
- B (Bold) is a bold text.
- I (Italic) is an italicized text.

3 Results and Discussion

In this section, we analyzed the rating of travel websites in each of the four categories. Five parts of the website HTML structure were systematically combined, and this resulted in the 31 patterns displayed in Table 3.

Table 3. The performance of the entire site structure using all 31 different combinations

R_{web}	Travel (0.45)	Accommodation (0.57)	Restaurants (0.54)	Souvenir Shops (0.46)	Average
B	1.00	1.00	1.00	1.00	1.00
B+L	1.00	1.00	0.98	0.98	0.99
B+K+L	0.99	1.00	0.97	0.95	0.98
B+D+L	1.00	0.99	0.99	0.91	0.97
B+K	1.00	1.00	0.96	0.89	0.96
T+B+L	0.98	0.97	0.98	0.89	0.95
B+K+D+L	0.97	0.98	0.98	0.88	0.95
T+B+K+L	0.97	0.96	0.98	0.85	0.94
B+D	0.98	1.00	0.97	0.80	0.94
T+B	0.95	0.98	0.97	0.84	0.93
T+B+K	0.96	0.98	0.96	0.80	0.93
B+K+D	0.95	1.00	0.95	0.79	0.92
T+B+D+L	0.97	0.92	0.96	0.81	0.92
T+B+K+D+L	0.95	0.91	0.97	0.78	0.90
T+B+K+D	0.91	0.98	0.98	0.73	0.90
T+B+D	0.92	0.94	0.94	0.74	0.88
T+K+L	0.88	0.85	0.93	0.77	0.85
K+L	0.82	0.82	0.91	0.84	0.85
T+L	0.84	0.78	0.92	0.77	0.83
D+L	0.89	0.75	0.90	0.77	0.83
K+D+L	0.89	0.89	0.78	0.68	0.81
L	0.92	0.55	0.88	0.88	0.81
T+D+L	0.84	0.78	0.90	0.68	0.80
T+K+D+L	0.82	0.78	0.91	0.65	0.79
K	0.77	0.51	0.63	1.00	0.73
D	0.63	0.28	0.77	1.00	0.67

Table 3. The performance of the entire site structure using all 31 different combinations (continue)

R_{web}	Travel (0.45)	Accommodation (0.57)	Restaurants (0.54)	Souvenir Shops (0.46)	Average
D+K+T	0.70	0.53	0.82	0.53	0.64
K+D	0.68	0.47	0.75	0.00	0.64
T+K	0.65	0.59	0.77	0.49	0.62
T+D	0.67	0.44	0.81	0.48	0.60
T	0.53	0.41	0.74	0.53	0.55

The travel website classification used WWCA with R_b and the optimal threshold values to categorize the Thai website with the performance of analysis technique using F-measure (Table 4). In addition, the most common tourist terms/words found in the HTML structure (HTML tags) of the body were analyzed and the results can be found in Table 5.

Table 4. Shows an example of the optimal boundary value from learning in each category

Category	Optimal Threshold	WWCA		
		Precision	Recall	F-measure
Attractions	0.59	0.45	0.55	100%
Accommodations	0.64	0.57	0.57	100%
Restaurants	0.55	0.54	0.54	100%
Souvenir Shops	0.46	0.46	0.46	100%

Table 5. The most common words related to traveling that were found in the HTML structure of the body

Category	HTML Tags in the Body Structure								
	UL	H1	P	TD	Strong	H3	OL	B	H2
Attractions	3.24	1.28	2.18	1.74	0.70	0.22	0.20	0.20	0.38
Accommodations	1.68	0.84	0.70	0.16	0.26	0.36	0.02	0.52	0.28
Restaurants	1.02	1.28	0.24	0.00	0.60	0.48	0.86	0.00	0.12
Souvenir Shops	0.12	0.56	0.16	0.34	0.16	0.28	0.14	0.3	0.22
TOTALS	6.06	3.96	3.28	2.24	1.72	1.34	1.22	1.02	1.00

Table 3 shows the performance of the entire site structure using all 31 different combinations with four categories. The F-measure values were calculated by using the threshold values from and the learning values from Table 2.

From Table 3, it was evident that the structure of the site in the body tag generates the highest performance at an average of 1.0 (100%) in all categories. This was followed by the Body+Link, Body+Keyword+Link, Body+Description+Link, and Body+Keyword, with F-measure values of 0.99, 0.98, 0.97, and 0.96 respectively.

For the structure of the site using Title+Keyword, Title+Description, and Title, the performance results were the least effective (0.62, 0.60, and 0.55, respectively). Therefore, it was concluded that the structure of the site that should be used to analyze the classification of travel sites is the body tag alone. The structure of Body+Link, Body+Keyword+Link, or Body+Description+Link can be also used for website classification. However, the structure of the link may be complicated and time-consuming. Hence, another best good option would be Body+Keyword in terms of a faster process and higher efficiency.

From Table 4, WWCA was based on site structure in the categories of travel, accommodation and restaurant. The best performance is at a minimum of 97%, except for the souvenir shops. This is likely because there is less information and words related to the souvenir shops. In fact, for some websites, there was very little material presented regarding gift shops, food, and so on.

From Table 5, it can be noted that in one of the estimated travel sites can be found in the structure of the body of the UL (Unordered List). The most common categories were attraction 3.24 words, accommodation 1.68 words, restaurants 1.02 words, and souvenir shops 0.12 words (total 6.06 words in all categories). The second most frequent number of travel words was found in the structure of the body H1 (Heading 1) and P (Paragraph), respectively. The structure of H2 (Heading 2) and B (Bold text) contain the least number of travel words.

In terms of analysis of the most common terms in the HTML structure of the body, the structure of the website in the body of UL, H1 and P should be used to analyze the classification of Thai tourist sites.

3 Conclusion and Future Work

In analyzing and categorizing the Thai travel websites, the Word Weight Calculation Algorithm (WWCA) based on the structures of a website was developed using the rating analysis of travel websites. Five parts of the website HTML structure (the title tag, the body tag, the meta-name description, the meta-name keywords, and the links to other pages) were extracted and used for the analysis and classification of 800 Thai tourist websites. These websites were further divided into four categories of tourist information: attractions, accommodation, restaurants and souvenir shops. The optimal threshold values in each category were used to analyze the classification of tourist sites with 31 different HTML structures. The optimal threshold values were gained from the R_b algorithm, which is based on the learning of 200 travel websites and divided into four categories of 50 each, and then measured using the F-measure statistic to select the best threshold value for each category.

The analysis was focused on text mining only. The results of the WWCA was measured regarding the efficiency using the F-measure statistic. The results showed the following:

1. The WWCA gave a good indicator for the Thai travel websites classification.
2. The content within the HTML body tag alone is sufficient to classify the sites.
3. The UL (Unordered List) and H1 (Heading 1) in the HTML structure of the body is best use for a faster classification process, since the most common words related to travel websites are located within these two structures.

The finding and technique we discovered in this research can also be applied to other website categories such as criminals, sports, etc. For our future work, the accuracy and time of the structure of the website in the body of the UL, H1 and P should be analyzed and measured for a faster process for the classification of Thai tourist sites. Furthermore, we will improve the WWCA using the technique of co-occurrence (coincidentally) and cut off the words with less weight in order to make the algorithm and the process faster.

References

[1] C.S. Namahoot, D. Lobo, S. Kabbua, Enhancement of a text clustering technique for the classification of Thai Tourism Websites, in: Proc. International Computer Science and Engineering Conference (ICSEC), 2014.

[2] R. Manikandan, R. Sivakumar, Machine learning algorithms for text-documents classification: a review, International Journal of Academic Research and Development 3(2)(2018) 384-389.

[3] S. Lee, J. Kim, S. Myaeng, An extension of topic models for text classification: a term weighting approach, in: Proc. 2015 International Conference on Big Data and Smart Computing (BIGCOMP), 2015.

[4] R. Luss, A. d'Aspremont, Predicting abnormal returns from news using text classification, Quantitative Finance, 15(6)(2015) 999-1012.

[5] M. Allahyari, S. Pouriyeh, M. Assefi, S. Safaei, E.D. Trippe, J.B. Gutierrez, K. Kochut. A brief survey of text mining: Classification, clustering and extraction techniques. <http://arxiv.org/abs/1707.021919.pdf>, 2017.

[6] R. Sammouda, A comparative study of effective supervised learning methods on arabic text classification, International Journal of Computer Science and Network Security 17(12)(2017) 130-133.

[7] T.M. Nogueira, S.O. Rezende, H.A. Camargo, On the use of fuzzy rules to text document classification, in: Proc. Proceedings of the International Conference on Hybrid Intelligent Systems, 2010.

[8] N. Sathapornvajana, The use of text mining and ontology in classifying tourist interest of Thailand tourism, [dissertation] North Bangkok: King Mongkut's University of Technology, 2010.

[9] K. Chatcharaporn, T. Angskun, J. Angskun, Tourist attraction categorization models using machine learning techniques, Suranaree Journal of Social Science 6(2)(2012) 35-58.

[10] W. Chotirat, P. Boonrawd, S.N. Wichian, Developing an ontology knowledge based for automatic online news analysis, Information Technology Journal 7(14)(2011) 13-18.

[11] C.S. Namahoot, M. Bruckner, N. Panawong, Context-aware tourism recommender system using temporal ontology and Naïve Bayes, recent advances in information and communication technology, Advances in Intelligent Systems and Computing 361(2015) 183-194.

[12] C.S. Namahoot, M. Bruckner, N. Panawong, A tourism recommendation system for Thailand using semantic web rule language and K-NN algorithm, Information 19(7)(2016) 3017-3024.