Travel Review Analysis System with Big Data (TRAS)

Chakkrit Snae Namahoot[1,2(✉)], Sopon Pinijkitcharoenkul[3], and Michael Brückner[4]

[1] Department of Computer Science and Information Technology, Faculty of Science, Naresuan University, Phitsanulok, Thailand
chakkrits@nu.ac.th
[2] Center of Excellence in Nonlinear Analysis and Optimization, Faculty of Science, Naresuan University, Phitsanulok, Thailand
[3] Information Technology Center, Pibulsongkram Rajabhat University, Phitsanulok, Thailand
sopon_b@psru.ac.th
[4] Department of Educational Technology and Communication, Faculty of Education, Naresuan University, Phitsanulok, Thailand
michaelb@nu.ac.th

Abstract. This paper introduces a process for online travel review analysis in Thai language employed in a recommender system supporting travelers (TRAS). The process covers three main categories: attractions, accommodation, and gastronomy. The filtering and queuing results gained with MapReduce build the input for three main steps: (1) the analysis process for element scores, (2) the analysis process for the total scores of the reviews, and (3) the travel guidance system based on users' selections. The extensive tests revealed that the system operates properly regarding functional and non-functional requirements. We employed 60,000 travel reviews containing all categories to test the analysis process for steps (1) and (2). We found that the number of adjectives and modifiers in each review affects the time used for analysis. In contrast to previous recommender systems, TRAS applies a more diverse and transparent rating and ranking approach. Travelers can select the features they are interested in and get personalized results, so that a given location might achieve different rankings for different travelers.

Keywords: Big data · Data analysis · MapReduce · Decision support system

1 Introduction

At present, the tourism industry in Thailand is receiving increasing attention. General travel information and specific reviews are important in planning and deciding where to visit. Potential travelers find tourist information on many Websites. Visitors can find information about tourism to make decisions and plan for tourism. However, with so much information available on the Internet, they need to spend a lot of time searching to make sure they get all information they need to make informed decisions. The information retrieved from the Internet covers such items as attractions,

M. Qiu (Ed.): SmartCom 2018, LNCS 11344, pp. 18–28, 2018.
https://doi.org/10.1007/978-3-030-05755-8_3

accommodation, hotel and restaurants but cannot be acknowledged whether the tourist information is good or not. Travelers mostly read reviews or comments from other travelers who have been to that place before. The reviews contain the pros and cons and can be used as identifying alternatives or for planning decisions. Travel information online reviews, particularly high-score reviews from tourists who have been there, can be used to support decision making on travel planning, for example, for choosing tourist places, accommodations, or restaurants. Due to the large number of reviews that users often encounter, it takes a lot of time reading those reviews and making informed decisions based on the results.

Regarding the tool, the large amount of data impedes analysis and processing of useful information within a reasonable time [1, 2]. To analyze these large data sets in data warehouse-like fashion such specialized pieces of software as Hadoop and Hive have been employed. The software library Hadoop enables the processing of large amounts of data, whereas Hive builds an SQL-like interface for extracting, manipulating and managing the data.

Attacking the problem mentioned above, we have developed a system for extracting and analyzing data from online travel reviews in Thai language with Hadoop and Hive. TRAS uses a novel approach with both qualitative and quantitative ranking analysis of user selected components found in traveler reviews on accommodation, attractions and gastronomy.

Another point that needs attention is the recognition of out-of-date reviews, which contents may be of no use at the time of decision making. The Traveler Review Analysis System (TRAS) analyses the posting dates of reviews and uses only those reviews that have been posted during the past 18 months of system run, whereby both Christian Era (AD) and Buddhist Era (BE) can be handled. Reviews without date are dismissed.

Previously published recommender systems do not take into account such features as date or reviews, key contents of facilities and food offered, among others. Instead they focus on overall ratings as the number of stars (hotels) and the number of positive and negative words used in the reviews divided by the total number of those words.

2 Related Work

The extraction of knowledge from only online hotel reviews using fuzzy logic [3] is a research that extract online reviews from travelers who stay at the hotel by building a knowledge base ontology of tourism. The ontology contains a glossary of terms that define level of scores from 1 (very poor) to 5 (very good) and the context-free grammars construction. Also the knowledge extraction task is to calculate the level of scores of the review sentences which is divided into two levels: (1) specific context/subject scores such as hotel features, and (2) overall scores using specific scores from (1) and a fuzzy calculation system. This research uses only online reviews in English and does not support Thai language. The emphasis is on data extraction, and the overview analysis uses a simple fuzzy rule based on hotel facilities.

A similar research covered the development of an ontology knowledge base for automated online news analysis [4] by considering the importance of words. Using the

term frequency-inverse document frequency (TF-IDF) in the news content to gather appropriate words for the development of the ontological knowledge base and weighting of key words in the news content. Only the number of important words influences the analysis but not the position of the words in the texts, although this may be important for understanding the meaning of message.

Data mining and customer feedback summary [5] is a research that analyze customer reviews of online merchants and summarize a positive and negative opinion. There are three parts in the process: (1) use of data mining techniques and natural language processing techniques to mine the characteristics of the product from customer feedback, (2) analysis of each comment to indicate whether the comment is positive or negative, and (3) the conclusion of the above steps. This provides very helpful information for other customers and also the producer/owner of the product who can continue to adjust the products, yet again there is only a summary of positive and negative opinions without further details of the characteristics.

Jian et al. [6] conducted a research to analyze big data relating hotel customer responses based on cloud data. Hadoop was deployed on a cloud server for data collection and analysis together with WebCrawler to collect the feedback. Techniques of Neural networking and unsupervised learning were used in the analysis of the comments to understand their meaning. Hadoop and the K-means algorithm were used to arrange and update the data in the database.

3 System Architecture Design

3.1 Big Data

Big data is often labeled as the 3Vs: (1) Volume, i.e. the size of the data, (2) the Variety of data: structured, semi-structured and unstructured, and (3) the Velocity of data processing [7]. The problem of large data is data processing, and the software we use for processing large data comprises:

- Hadoop: an open source software developed by Java which is designed to accommodate a wide variety of structured and unstructured data and can process large data. The first version of Hadoop has two main components, the Hadoop Distributed File System (HDFS) and MapReduce. HDFS serves as a storage area. This stores large data classified into large subdirectories at a large number of Data Nodes. There is a Master Node that specifies the location of the data stored in the Data nodes. MapReduce works as the data processor to analyze data in the form of Map and Reduce function so that the system can distribute them in parallel to many Hadoop machines.
- Hive: a data warehouse system for Hadoop that easily facilitates data summary. It is a query tool that stores information in HDFS with SQL language instead of MapReduce programming. Hive is responsible for translating SQL statements into MapReduce. Large-scale data analysis using Hadoop Pig and Hive [8] is a research that presents large-scale data analysis using Hive, which runs in SQL format. There is also a research on Hive software that works on Hadoop [9] such as Facebook's

Data Infrastructure team presented loading and adding data with Hive, which looks at the data in tabular form.

3.2 System Process

In the process of analyzing online travel reviews, three main subjects (categories) of tourist information are covered: tourist attractions, accommodation and restaurants. An overview of the system process is presented in Fig. 1.

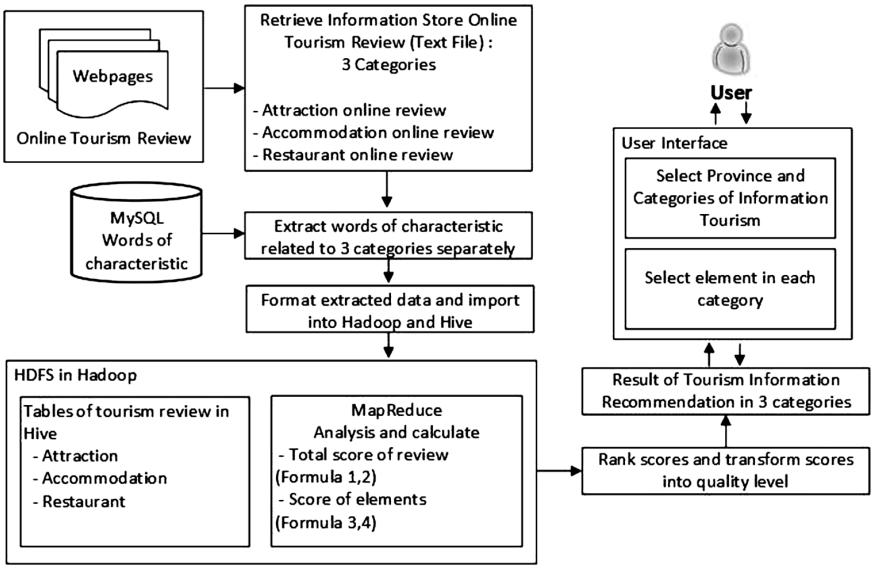

Fig. 1. System process

User: selects province (e.g., Phitsanulok Province) and categories of tourism information (attraction, accommodation and restaurant) they are interested in.

Select element in each category, for example a user selects accommodation category and refines by elements of service, cleanliness of a place and price.

The process used by the system is as follows (see formulas in the next section):

- Store online travel reviews related to attractions, accommodations and restaurants to a text file, then extract feature data, such as words of characteristics (adjectives) describing attractions, hotels and restaurants.
- Format and import the data into tables in Hive.
- Analyze and calculate the score of three types of tourist information review: the total score of reviews (Formula 1 and 2), score of all elements and score of the selected elements (Formula 3 and 4), rank the scores and transform them into quality levels.
- Display the results with ranking scores and quality levels.

3.3 Extraction of Review Characteristics from Big Data

A database is created to collect the words of the characteristics that describe important features of tourism information as used in the extraction process. The following types of words have been categorized:

- A word that indicates the quality, appearance, and characteristic of tourist attractions, accommodation and restaurants, such as stunning, beautiful, cheap, far, poor, bad, like, favor etc. 3 values of the words of characteristic: 1 represents positive, 0 represent fair and −1 represent negative words of review.
- Words that extend the words of characteristic as prefix (freaking, bloody (โคตร), very, a lot, (มาก)), suffix (a lot, so much (มาก)) etc.
- Words that indicating opposite, such as not (ไม่), for example not good, not ok, not so good, not bad, or not far.

All categorized words have a value level that is used in the analysis: the value of −2, −1, 0, 1, and 2 to show the levels of bad, fair and good (can be calculated from Table 1). The symbols +, − and 0 are "status symbols". The concepts of the item characteristics are extracted from our tourism ontology [10, 11].

Table 1. An example of words (characteristic) value calculation from review

Review: not clean, very nice location, cheap compared to quality, I like the room a lot even it smells bad, service is ok					
Prefix modifier	Words of characteristic	Suffix modifier	Opposite modifier	Word value calculation	Sum of word value
-	Clean (+1)	-	Not (-)	−1	−1
Very(+1)	Nice (+1)	-	-	1 + 1	2
-	Cheap (+1)	-	-	1	1
-	Like (+1)	a lot (+1)	-	1 + 1	2
-	Smell (−1)	Bad (−1)	-	−1 + −1	−2
-	Ok (0)	-	-	0	0

Information obtained by the extraction process is then formatted and stored on Hive.

3.4 Review Analysis Process

In this process, the score of the adjectives extracted from Sect. 3.3 is calculated in three steps: (1) the total score of review, (2) the score of all elements, and (3) the score of selected elements by users, see the following algorithm.

Pre-process:

1. Input a review text and check the date of the review and exclude those without dates or with a date older than 18 months.

2. Classify elements and words identify them in each category as follows: Elements of attraction (convenience, service, atmosphere, facility, good place, and others), accommodation (convenience, service, clean place, food, atmosphere, price, and others) and restaurant (service, atmosphere, taste, price, music, fast delivery).
3. In each category (attraction, accommodation or restaurant), list all names (e.g. name of attraction, hotel, restaurant) to be analyzed.

Process:

Initialization: Review (ri) = 1, Words of characteristic = 0
Count all elements of each category and return number of all element (a)
Count words of characteristic in each review and return number of words (n)
    Return position of each word
If position of words of characteristic is next to each other (no extended words) then
    Return the words with value (compared to the words and value in database)
If not (there is extended word) check types of extended words and calculate values of words of characteristic (Table 1)
    Return number of words (n) of characteristic in each review
Compute score of each review using formula 1
    Return score of each review (pr)
Do until review (i) = m (all review)
Compute score of all review using formula 2.
    Return total score of all review (pr_{sum})
Compute score of all elements (pe_{all}) in each category using formula 3
    Return score of all elements (pe_{all})
Transform score into a quality level (very good, good, fair, bad, very bad, Table 2)
    Return a quality level for user Interface (Tourism Recommendation System with an Online Tourist In-formation Reviews Analysis):
User selects province, category with elements
    Return number of selected elements (s) according to the interface
Compute score of selected elements (pe_{select}) in each category using formula 4
Display elements found, recommend names in each selected category with total score of review (pr_{sum}), the quality level
Terminate after the last review to be processed

The score of each review pr is computed as

$$pr = \frac{\sum_{i=1}^{n} Wc_i}{n} \tag{1}$$

where Wc_i is the value of each word of characteristic i to n, n is a number of words characteristic found in each review.

The score of all reviews pr_{sum} is computed as

$$pr_{sum} = \frac{\sum_{i=1}^{m} Pr_i}{m} \tag{2}$$

where pr_i the score of each review i to m, and m is the number of review found.

The score of all elements pe_{all} is computed as

$$pe_{all} = \frac{\sum_{i=1}^{a} Ca_i}{a} \tag{3}$$

where Ca_i is the score of each element i to a, and a is the number of all elements found in each category.

The score of user selected elements is

$$pe_{select} = \frac{\sum_{i=1}^{s} C_i}{s} \tag{4}$$

where C_i the score of each element i, and s is the number of selected elements in each category.

All scores (1)–(4) are then transformed into quality levels which are as follows:

From Table 1, word values can be divided into 5 levels (2 (very good), 1(good), 0 (fair), −1 (bad), −2 (very bad)) which is a level of measurement of class interval data. Thus, the class width can be calculated as (maximum score − minimum score)/number of class which is (2−(−2))/5 = 0.8 and score interval of quality level is as follows: Score ranges from 1.21 to 2.00 indicates a very good quality level, 0.41 to 1.20 indicates a good quality level, −0.40 to 0.40, indicates a fair quality level, −1.20 to −0.41 indicates a poor quality level and −2.00 to −1.21 indicates a very poor quality level.

Table 2. Quality level and score of all reviews (pr_{sum}), score of each element and score of all elements (pe_{all}) relating accommodation category

Element of accommodation	Number of reviews (m)	($\sum Pr_i$)	$\sum C_i$	Quality level of elements
Service	4	(−1) + (1) + (2) + (2) = 4	1	Good
Atmosphere	1	0	0	Fair
Convenience	2	(1) + (−1) = 0	0	Fair
Food	6	(1) + (−1) + (1) + (−2) + (−1) + (1) = −1	−0.17	Fair
Price	4	(−1) + (−1) + (−1) + (1) = −2	−0.5	Bad
Clean place	3	(−1) + (−1) + (−1) = −3	−1	Bad
Others	1	(−1) = −1	−1	Bad
Sum	21	−3	$\sum$ Cai = −1.67	Fair (−0.24)
pr_{sum} = −3/21 = −0.14 (quality level = fair)				
pe_{all} = −1.67/7 = −0.24 (quality level = fair)				

In Table 2 examples of calculations and scores of all reviews (pr_{sum}), scores of each element and scores of all elements (pe_{all}) of the accommodation category are presented.

4 Result and Discussion

To test the development of the tourism recommendation system, three main categories of approximately 60,000 online tourist information reviews have been used: attractions, accommodations, and gastronomy. The results are as follows: Fig. 2 shows the total score of all ten reviews regarding Phitsanulok United Guest House, which is 0.97 (the

Phitsanulok United Guest House

Review no.	Number of words characteristic (n)	Wc_i	Pr	Quality of Review
1	3	3	1.00	Good
2	15	12	0.80	Good
3	4	1	0.25	Fair
4	7	8	1.14	Good
5	5	5	1.00	Good
6	2	2	1.00	Good
7	2	3	1.50	Very Good
8	3	3	1.00	Good
9	4	4	1.00	Good
10	5	5	1.00	Good
		Total Score of Reviews		0.97
		Quality level		Good

Fig. 2. Total score based on all reviews regarding Phitsanulok United Guest House

Phitsanulok United Guest House

element (a)	No of word (m)	Ca_i	Pe_{all}	Quality of level of element
Convenience	1	1	1.00	Good
Service	5	5	1.00	Good
Clean place	24	22	0.92	Good
Food	1	0	0.00	Fair
Atmosphere	2	0	0.00	Fair
Price	2	2	1.00	Good
Other	15	16	1.07	Good
		Total Score of Elements		0.71
		Quality level		Good

Fig. 3. Sample score calculation results from all elements.

sum of score of each review is divided by the number of all reviews (9.69/10 = 0.97). Then the score of 0.97 is transformed into a quality level, which means that the Phitsanulok United Guest House has a good quality level based on the ten reviews.

Figure 3 shows the results of the total score analysis of all selected elements for Phitsanulok United Guest House which is 0.71 (total score of all elements is divided by the total number of all elements (4.99/7 = 0.71). The result of the analysis shows that Phitsanulok United Guest House has good quality based on selected elements.

Figure 4 displays the screen when the user selects Phitsanulok Province, all categories, and two, four and four elements from attraction, accommodation and restaurant, respectively. The system calculates the scores as described and the results are sorted from highest to lowest total score (Fig. 5).

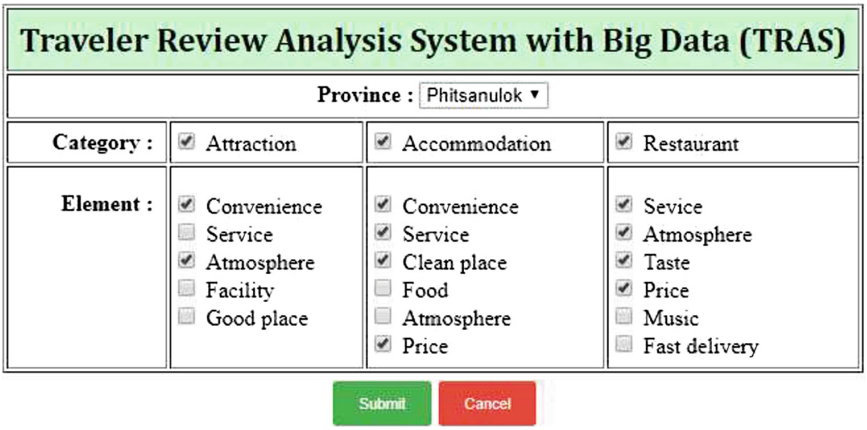

Fig. 4. User interface of TRAS

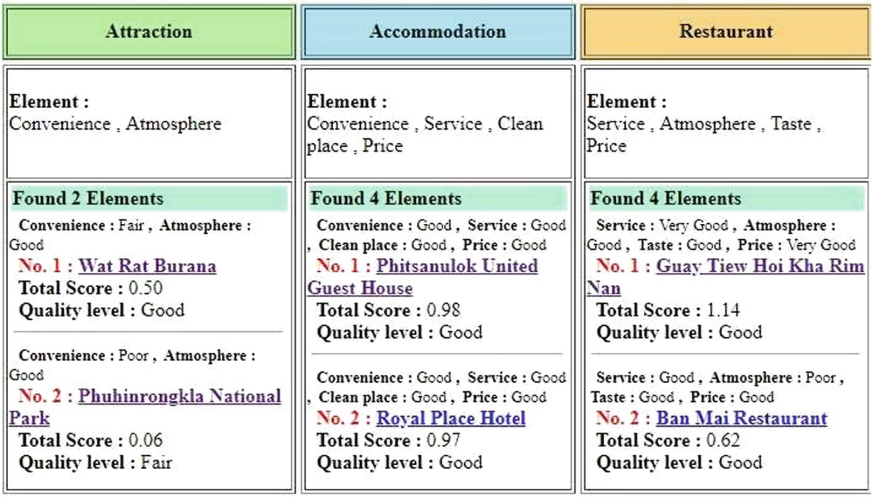

Fig. 5. Personalized TRAS recommendations for Phitsanulok Province

Table 3 shows a comparison of features of similar systems developed previously and TRAS.

Table 3. Comparative results with the [3] and [6]

System features	[3]	[6]	TRAS
Hotel review	✓	✓	✓
Tourism recommendation	×	×	✓
Ranking overview	✓	✓	✓
Ranking tourism features	×	×	✓
Mobile application	×	×	✓
Analysis with total score quality level	✓	×	✓
Selected scores of tourism information	×	×	✓
Analysis with big data analytic tools	×	✓	✓

5 Conclusion and Further Work

In this research, the Traveler Review Analysis System (TRAS) based on an online tourist information reviews analysis process using big data has been developed. TRAS analyzes travel information in Thai language and covers three main categories: attractions, accommodations, and gastronomy. Review data are extracted, transformed and stored using Hadoop and Hive applications. Individual and overall scores are calculated considering adjectives and modifiers. The results are displayed as recommendations via the user interface. Users can select province, travel elements and travel categories, and then the system ranks the items individually selected by the users.

In contrast to previously developed recommender systems, TRAS takes into account requirements selected by users and ranks the locations individually. This may lead to different ranking results for different travelers. Additionally, TRAS is available as a smartphone application and can therefore be accessed on the spot.

The results of approximately 60,000 test cases show that the system operates properly. In terms of timing, we employed approximately 7,000 reviews of travel information from the three categories to test the analysis process. The number of adjectives and modifiers affects the time used for analysis. This is because the system checks all adjectives, pre-modifiers, including post-modifiers of the adjectives before performing the analysis.

A limitation of TRAS is inherent to the concept of traveler reviews: fake reviews posted by parties close to the item of review cannot be spotted easily and subsequently excluded. For the next version of TRAS, we plan to employ only reviews posted on high quality Websites that check the plausibility of reviews and ask reviewers to unveil their hotel reservations, tickets and so forth. It remains to be seen, though, how much information will be left after using this strategy.

References

1. Snae, C., Pawarawat, N.: A study of internet user behaviour using techniques of data mining and temporal ontology. In: The Third Naresuan Research Conference, Phitsanulok, Thailand, 28–29 July 2007 (2007). (in Thailand)
2. Snae, C., Brückner, M.: Data cleaning and clustering of internet log files based on a temporal ontology. In: The Third Mahasarakham International Workshop on AI (MIWAI 2009) (2009)
3. Kitwattanathaworn, P., Angsakul, T., Angsakul, C.: Knowledge extraction system from online hotel review using fuzzy logic. J. KMUTNB **23**(2), 363–377 (2013)
4. Chotirat, W., Boonrawd, P., Wichian, S.N.: Developing an ontology knowledge based for automatic online news analysis. Inf. Technol. J. **7**(14), 13–18 (2011)
5. Hu, M., Liu, B.: Mining and summarizing customer reviews. In: Proceedings of the 2004 ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 168–177 (2004)
6. Jian, M.S., Fang, Y.C., Wang, Y.K., Cheng, C.: Big data analysis in hotel customer response and evaluation based on cloud. In: 19th International Conference on Advanced Communication Technology (ICACT) (2017). https://ieeexplore.ieee.org/document/7890201/. Accessed 21 Aug 2018
7. Bhosale, H.S., Gadekar, D.P.: A review paper on big data and hadoop. Int. J. Sci. Res. Publ. **4**(10), 1–7 (2014)
8. Dhawan, S., Rathee, S.: Big data analytics using hadoop components like pig and hive. Am. Int. J. Res. Sci., Technol., Eng. Math. (AIJRSTEM) **2**, 88–93 (2013)
9. Thusoo, A., et al.: Hive - a warehousing solution over a map-reduce framework (2009). https://research.facebook.com/publications/hive-a-warehousing-solution-over-a-map-reduce-framework/. Accessed 18 July 2018
10. Namahoot, C.S., Panawong, N., Brückner, M.: A tourism recommendation system for thailand using semantic web rule language and K-NN algorithm. INFORMATION **19**(7), 3017–3024 (2016)
11. Namahoot, C.S., Brückner, M., Panawong, N.: Context-aware tourism recommender system using temporal ontology and Naïve Bayes. In: Unger, H., Meesad, P., Boonkrong, S. (eds.) Recent Advances in Information and Communication Technology 2015. AISC, vol. 361, pp. 183–194. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-19024-2_19