ICIC Express Letters
Part B: Applications
Volume **15**, Number **5**, May **2024**

ICIC International ©2024 ISSN 2185-2766
pp. **455–465**

DEVELOPING AN AUTOMATIC OPEN QUESTION AND ANSWERING OF ASSOCIATION QUESTION DATA FOR COMMUNITY TOURISM

Chatklaw Jareanpon[1], Umaporn Chaisoong[2], Thummarat Boonrod[3]
and Khanabhorn Kawattikul[4,*]

[1]POLAR Lab, Department of Computer Science
[2]Department of Information Technology
Faculty of Informatics
Mahasarakham University
Khamriang Sub-District, Kantarawichai District, Maha Sarakham 44150, Thailand
chatklaw.j@msu.ac.th; umaporn.ch@rmuti.ac.th

[3]Department of Digital Technology
Faculty of Administrative Science
Kalasin University
62/1, Kasetsomboon Road, Muang District, Kalasin 46000, Thailand
thummarat.bo@ksu.ac.th

[4]Department of Information Technology
Faculty of Social Technology
Rajamangala University of Technology Tawan-ok
131 Moo 10, Phluang, Khao Khitchakut, Chanthaburi 22210, Thailand
*Corresponding author: khanabhorn_ka@rmutto.ac.th

Received September 2023; accepted November 2023

Abstract. *The research aims to develop automatic questions and answers using Open-QA, resolving the problem of tourists' spatial information access from the association question data. The difficulty of Thai language processing is that the language has no space in a sentence, without punctuation or word separation. Correctly tokenizing or separating words affects precision and accuracy. The efficient data access of the tourists is a challenge of this research. The similarity method, the cosine similarity technique, is based on the Vector Space Model (VSM) using TF-IDF weighting and Bag of Words (BoWs). It is efficient by bringing the outstanding points of text vectorization to calculate and acquire the crucial features for being the document representative efficiently. The result of the Bows stage is 19,501 terms from 1,237 documents. The evaluation of model effectiveness has an Accuracy value of 99%, which best indicates the ability to describe the answer and efficacy of the model.*

Keywords: Information retrieval, Thai word segmentation, TF-IDF, Cosine similarity, Natural language processing, Tourism

1. **Introduction.** Sustainable tourism development [1] is one of the goals applied to the global master plan of AGENDA 21 [2] and corresponds to the National Economic and Social Development Plan. Community-Based Tourism (CBT) [3] describes the trend of emphasizing communication, innovation, and technology that provides unique tourist experiences, including accessing information rapidly and meeting requirements, such as Recommendation System (RS) or Question Answering System (QA). However, the main problems of the tourism industry still need to be resolved precisely. For example, the public relations efforts do not align with the target audience, and the improved technologies still cannot meet the problem of spatial news and information access, including quick

DOI: 10.24507/icicelb.15.05.455

information preparation. Furthermore, a large amount of information derived from the information query by a search engine is limited to access: 1) The data retrieval is irrelevant, 2) the information is excessive, and 3) it is not easy to investigate its resources and facts. Hence, to develop the search engine, it is necessary to have specific qualifications for retrieving accurate documents efficiently from the outstanding features of tourism information [4]. The QA can roughly be divided into textual QA and Knowledge Base (BK)-QA. Textual QA mines answer from unstructured text documents, while KB-QA is from a predefined structured KB that is often manually constructed. The Open-domain (OpenQA), categorized as textual QA, tries to answer a given question without any specified context. The first OpenQA was defined as extracting the top 5 probable snippets containing the correct answer from a collection of news articles in the QA [5]. A well-known QA system called AskMSR [6] translates the user's question into queries relying on predefined rewriting rules to gather relevant documents from search engines. It adopts a series of n-grams based on algorithms to mine, filter, and select the best answer. Nowadays, QA encounters challenges when indexing documents of considerable volume and diverse formats [7], resulting in inefficient, laborious, and time-consuming information retrieval [8]. Different from any other association dataset from our previous research [9], we cluster the data from the online question of tourism in Thailand using the Exploratory Factor Analysis (EFA) and Principle Component Analysis (PCA). The output of this research reduces the non-statistical correlation by Kaiser-Meyer-Olkin Measure of Sampling Adequacy and Bartlett's Test. After that, Chaisoong et al. [9] find the relationship of those factors using Apriori. The example output is shown in Figure 1, and the examples of the Question Pattern of Cluster N are

1) Rule 1: The community has various tourist attractions | Local identity | Outstanding identity | Conservation of local culture History, culture, way of life and traditions?
2) Rule 2: The community has tourist attractions unique to the local area | Outstanding identity | Preserve culture, local history culture. Is it a way of life and traditions?
3) Rule 3: The community has various tourist attractions | Local identity | Preserves local culture and cultural history. Is it a way of life and traditions?

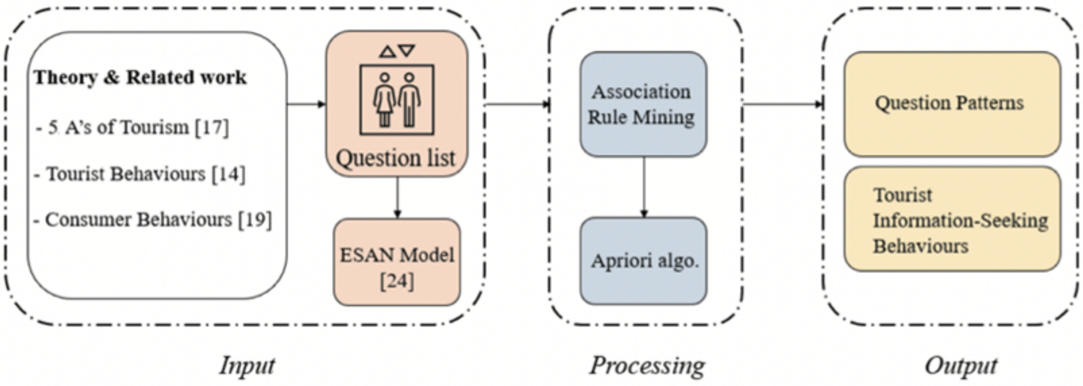


Figure 1. The framework of [9,10] (Image from [10])

The challenge of this research is free text question that is automatically generated from the rule and various document sources and types for querying the output. The research aims to develop a framework of automatic open questions and answers for tourists with an OpenQA model that gets the results rapidly, accurately, and precisely. It can build tourist opportunities, stimulate tourism, and build the economic foundation.

2. **Background.** This section introduces background research; the details are as follows.

1) Question Answering (QA) [11,12]. QA based on Information Retrieval (IR) aims to provide an answer that is precise for the user's questions. The QA system is based on keyword searches with specific questions in natural language. QA can be divided into three sub-tasks: question analysis, document retrieval, and answer extraction [4], as shown in Figure 2. Additionally, this QA will extract from the text in unstructured that will use the background below to solve this problem. Moreover, the OpenQA tries to answer a given question without any specified context.

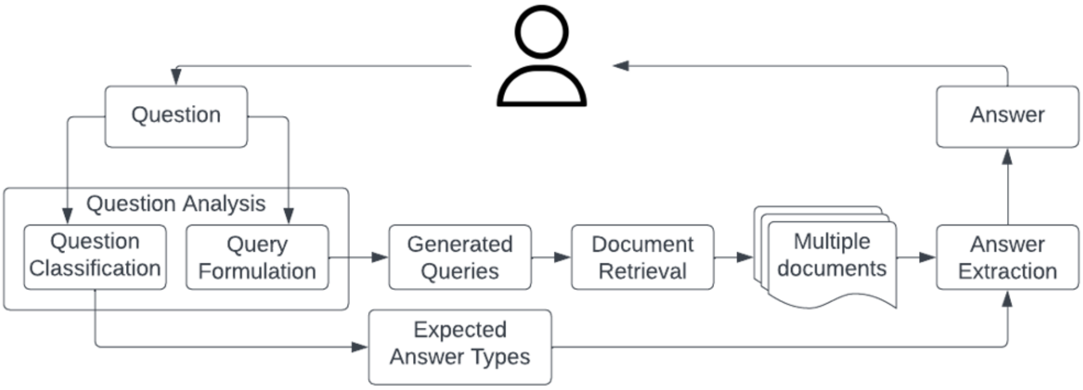


Figure 2. Architecture of the OpenQA system

2) Vector Space Model (VSM) [13]. It describes the Term-Document Matrix (TDM) and creates the index for retrieval. Each unit in the matrix is the weight value relying upon the term frequency found in two documents for comparison.

3) Thai word segmentation. Unlike the English language, the difficulty of Thai language processing is that the language has no space in a sentence, without punctuation or word separation. Correctly tokenizing or separating words affects the precise and accurate search. Many techniques can support the tokenization of Thai words, such as longest matching, probabilistic model, feature-based approach, maximal matching (newmm) [14], and the greedy algorithm. PyThaiNLP provides the tool used for the word tokenization. It consists of significant functions such as word_tokenize, thai_stopwords, and wordnet that work quickly and accurately using the Dictionary-based (DCB) principle and the WordNet.

4) Term weighting. It is a method to determine the weight values of keywords or document representation for indicating the significance of the terms and the index of such documents. The terms' weighting will be considered by the terms' frequency appearing in those documents and the number of all documents appearing with such terms. A well-known method for the term weighting is called term frequency-inverse dense frequency [15].

Term Frequency-Inverse Dense Frequency (TF-IDF) [15]. It is a technique used to analyze the messages working on a based matrix by extracting the crucial data from the text in the document. It is the components of words considered within the sentence of documents or retrieval selected by the attracting contents appearing in such documents to be the agent of several ones. There are two parts of the main components as the following: 1) Term Frequency (TF), and 2) Inverse Document Frequency (IDF). Hence, TF-IDF will find out the weight value of the word's frequency appearing in the document and the weight value of the inverse in the document frequency calculated from all documents, as shown in Equation (1)

$$TfIdf_{ik} = Tf_{ik} \times \log\left(\frac{N}{n_i}\right), \text{ and } TfIdf_{ik} = 1 + \log\left(Tf_{ik}\right) \times \log\left(\frac{N}{n_i}\right) \tag{1}$$

where Tf_{ik} is the term frequency, and N is the number of the words.

Bag of Word model is a simplifying representation. It is commonly used in document classification methods, where each word's (frequency of) occurrence is used as a feature for a classifier.

5) Similarity measurement and model evaluation, as the following details:

Similarity [16]. It is a similarity indicator by comparing methods and measuring the distance between the two documents. The result will have a value between 0 and 1. '0' means the minimum similarity, and '1' refers to the maximum similarity. There are numerous ways to calculate the similarity by measuring the distance. For the reasons of our dataset, this research selected the cosine similarity because these methods bring the vectors of any two data in the 'i' dimensional space to calculate the dot product to find out the angle value at such two vectors. The two vectors will get a similarity if the angle gets a minimum value. Additionally, even if the two similar data objects are far apart by the Euclidean distance because of their size, they could still have a smaller angle between them. One advantage of cosine similarity is its low complexity, especially for sparse vectors: only the non-zero coordinates must be considered.

Information retrieval model evaluation [17]. A good retrieval model should precisely extract the information relevant to the inquiries and sequence as significant to satisfy the users. Hence, evaluating the model competence is considered a crucial factor for describing the efficiency of the created model. For the model evaluation, the researcher selected four values: Precision, Recall, Accuracy, and F1, as in Equation (2), and all of the variables are shown in Table 1.

$$F1 = \frac{2 * Precision * Recall}{(Precision + Recall)} \tag{2}$$

where $Precision = \frac{TP_j}{TP_j+FP_j}$, $Recall = \frac{TP_j}{TP_j+FN_j}$ and $Accuracy = \frac{TP_j+TN_j}{TP_j+TN_j+FP_j+FN_j}$.

Table 1. Accuracy, Precision, Recall, and F1 variables

	Actual class	
Predicted class	TP_j (True Positive) (Correct result)	FP_j (False Positive) (Unexpected result)
	FN_j (False Negative) (Missing result)	TN_j (True Negative) (Correct absence of result)

3. Methodology.

For the conceptual framework, this research realizes the problems affecting tourism, including the perspectives on the opportunity to promote the images and travel decisions of the tourists in the South Northeastern Region for extending and using academic benefits based on the previously publicized research [18,19] in four aspects as shown in Figure 3, and more details of each step are as follows.

3.1. Query collections.

The input data are obtained from the experiment in the previous research [18,19]. There are three steps: Step 1: Analyzing the inquiry lists with Exploratory Factor Analysis using the Principal Component Analysis (PCA) method in a type of Varimax Orthogonal Rotation. A result is a group of relevant question lists, then determined by the new group name, the ESAN model. It consists of four classes: 1) E: Excellent Service, 2) S: Standard Facilities, 3) A: Accuracy of Information, and 4) N: Noble Culture [18]; Step 2: finding the association rule mining with the Apriori Algorithm from the group of question lists got from the first step by considering two parts: 1) confidence value and support value up from 98%, and 2) rules with empirical and outstanding features. It is concordant with the concept of Maslow's Hierarchy of Needs by removing redundant rules. It could create 14 best rules for Field [19]; Step 3: Data pre-processing.

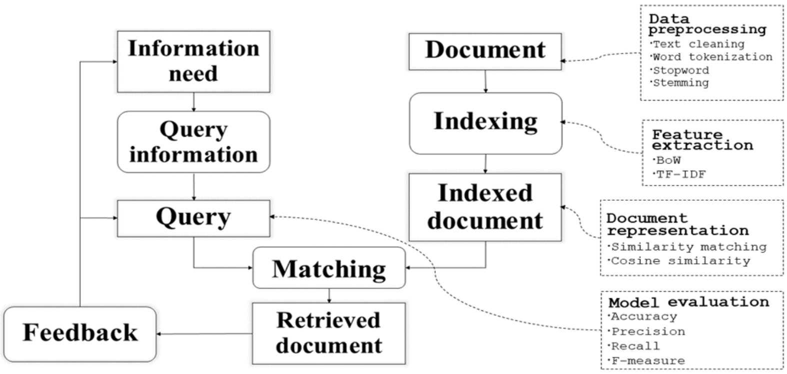


Figure 3. Proposed conceptual framework

Table 2. Data pre-processing in query collections

Class	Thai words segmentation (Tokenization)
E	“แหล่งท่องเที่ยว” ,“ห้องน้ำ” ,“สะอาด” ,“มาตรฐาน” ,“นักท่องเที่ยว” ,“ร้านค้า” ,“สินค้า” ,“บริการ” ,“ความหลากหลาย” , “เจ้าหน้าที่” ,“ดูแล” ,“ความปลอดภัย” ,“ราคา” ,“ร้านอาหาร” ,“ได้มาตรฐาน” ,“อาหาร” ,“เอกลักษณ์” ,“ชุมชน”
S	“กิจกรรม” ,“การท่องเที่ยว” ,“เอกลักษณ์” ,“วัฒนธรรม” ,“วิถีชีวิต” ,“ภูมิปัญญาท้องถิ่น” ,“ที่พัก” ,“ความปลอดภัย” ,“ความรู้” , “ความเข้าใจ” ,“ชุมชน”
A	“แหล่งท่องเที่ยว” ,“ระเบียบ” ,“ข้อปฏิบัติ” ,“นักท่องเที่ยว” ,“สารสนเทศ” ,“เว็บไซต์” ,“ข้อมูล” ,“ประชาสัมพันธ์” ,“กิจกรรม” , “การท่องเที่ยว” ,“คู่มือ” ,“เอกสาร” ,“เผยแพร่” ,“รายละเอียด” ,“เจ้าหน้าที่” ,“คำแนะนำ” ,“บริการ” ,“ดี” ,“คุณภาพ” ,“การสื่อสาร”
N	“ชุมชน” ,“แหล่งท่องเที่ยว” ,“หลากหลาย” ,“อัตลักษณ์” ,“เฉพาะท้องถิ่น” ,“เอกลักษณ์” ,“โดดเด่น” ,“การอนุรักษ์” ,“วัฒนธรรม” , “ท้องถิ่น” ,“เข้มแข็ง” ,“ประวัติศาสตร์” ,“สะท้อน” ,“วิถีชีวิต” ,“ประเพณี”

There are four steps, which include 1) text cleaning by spacing and punctuation removal, 2) word segmentation or tokenization, 3) stop word removal, and 4) stemming, as shown in Table 2.

3.2. **Answer collections.** It is about the data collection of the South Northeastern Region, Thailand, and tourism in four provinces, a total of 1,237 documents. The data were from primary and secondary information, such as search engines, paper, brochures, and digital files, from the involved organizations, and they are in a document in MS Word format, as shown in Table 3. Then, all data were put into data pre-processing and feature extraction. The result included Term, BoW, and Term Vector as answers to the users' queries.

Table 3. List of classes in the corpus

Class/Label	Corpus		
	Documents	(%)	Storage (Mb)
E	524	42.36	8.54
S	267	21.58	4.3
A	201	16.25	3.23
N	245	19.81	3.78
Total	1,237	100.00	19.85

Text cleaning and pre-processing: This stage is a part of information preparation, including the text cleaning process, such as specific character removal, stop words, word stemming, space removal, number removal, and repeated terms. The researcher uses the tool provided by Google Colaboratory with Python [20] language program and

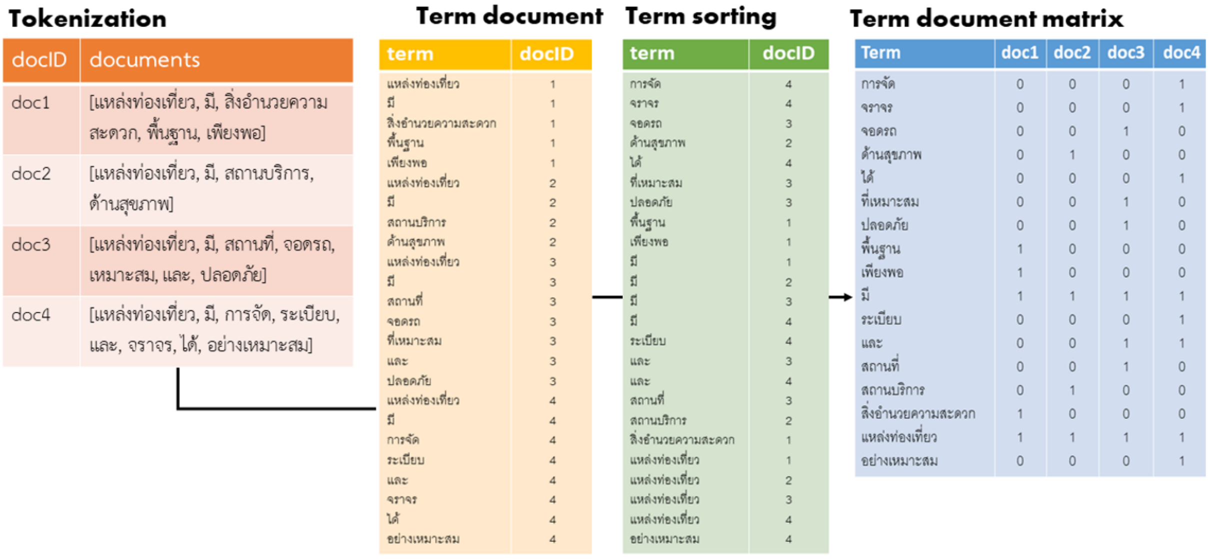
PyThaiNLP, for data analysis in four main steps: 1) data pre-processing, 2) feature extraction, 3) document representation, including feature selection, and 4) similarity matching for textual information retrieval [21], and the details are as follows.

Tokenization: It organizes the document in the string form or separation of texts into sub-units to find the scope of the terms, the smallest language unit. It can be divided into three main groups: 1) rule-based, 2) corpus-based, and 3) Dictionary-based (DCB). The reference method from the dictionary of WordNet [22] gathers the synset or set of terms with a similar meaning in the form of terms network, interpreted based on the newmm [12]. For the mentioned reason, this research used the newmm method.

Stop words: It finds an enormous number of unnecessary words in BoW without changing the meanings of the document. The superfluous words in the Thai language are often found, such as “โดย (by)”, “คือ (is)”, “เพื่อ (for)”, and “ซึ่ง (then)”. This process results in the reduction of unnecessary words from the feature extraction stage.

Stemming words: It finds the stemming words by transferring the different forms of words to be similar. For example, “State (รัฐ)” and “City (เมือง)” to let the explanation be in the same direction. Finding the stemming words is conducted before indexing to increase the retrieval effectiveness using the Porter Algorithm [11].

Feature extraction: The feature extraction stage consists of two methods: 1) BoW and 2) TF-IDF. **BoW:** It is the crucial feature extraction process for the information in a document type to be the agent of document features. It will be stored in a vector form substituted by the numbers 0 and 1 and substituted by Term Frequency. The prominent point is the unique ID, but the disadvantage is that it ignores the grammar and word order [23]. It can be solved by the TF-IDF method, as shown in Figure 4. **TF-IDF:** It is about indexing document representatives. The technique reduces the document’s complexity and helps filter commonly used terms. The vector numbers will substitute the documents to record the frequency statistics and consider the order of data results. The main words will be ordered following the letters with the frequency of the terms found in the document. Then, all document lists will be represented as document numbers. TF and IDF are divided into two parts, using the VSM, which substitutes the text with a single word [21].



Tokenization

docID	documents
doc1	[แหล่งท่องเที่ยว, มี, สิ่งอำนวยความสะดวก, พื้นฐาน, เพียงพอ]
doc2	[แหล่งท่องเที่ยว, มี, สถานบริการ, ด้านสุขภาพ]
doc3	[แหล่งท่องเที่ยว, มี, สถานที่, จอดรถ, เหมาะสม, และ, ปลอดภัย]
doc4	[แหล่งท่องเที่ยว, มี, การจัด, ระเบียบ, และ, จราจร, ได้, อย่างเหมาะสม]

Term document

term	docID
แหล่งท่องเที่ยว	1
มี	1
สิ่งอำนวยความสะดวก	1
พื้นฐาน	1
เพียงพอ	1
แหล่งท่องเที่ยว	2
มี	2
สถานบริการ	2
ด้านสุขภาพ	2
แหล่งท่องเที่ยว	3
มี	3
สถานที่	3
จอดรถ	3
ที่เหมาะสม	3
และ	3
ปลอดภัย	3
แหล่งท่องเที่ยว	4
มี	4
การจัด	4
ระเบียบ	4
และ	4
จราจร	4
ได้	4
อย่างเหมาะสม	4

Term sorting

term	docID
การจัด	4
จราจร	4
จอดรถ	3
ด้านสุขภาพ	2
ได้	4
ที่เหมาะสม	3
ปลอดภัย	3
พื้นฐาน	1
เพียงพอ	1
มี	1
มี	2
มี	3
มี	4
ระเบียบ	4
และ	3
และ	4
สถานที่	3
สถานบริการ	2
สิ่งอำนวยความสะดวก	1
แหล่งท่องเที่ยว	1
แหล่งท่องเที่ยว	2
แหล่งท่องเที่ยว	3
แหล่งท่องเที่ยว	4
อย่างเหมาะสม	4

Term document matrix

Term	doc1	doc2	doc3	doc4
การจัด	0	0	0	1
จราจร	0	0	0	1
จอดรถ	0	0	1	0
ด้านสุขภาพ	0	1	0	0
ได้	0	0	0	1
ที่เหมาะสม	0	0	1	0
ปลอดภัย	0	0	1	0
พื้นฐาน	1	0	0	0
เพียงพอ	1	0	0	0
มี	1	1	1	1
ระเบียบ	0	0	0	1
และ	0	0	1	1
สถานที่	0	0	1	0
สถานบริการ	0	1	0	0
สิ่งอำนวยความสะดวก	1	0	0	0
แหล่งท่องเที่ยว	1	1	1	1
อย่างเหมาะสม	0	0	0	1

Figure 4. An example of feature extraction using BoW model

Inverted indexing: It creates the documents representative in a weight vector form, which is the calculation of finding the terms weight. The research uses inverted indexing through a Single Word type for evaluating the significance of words or that feature [14]. Term's weight value can be found by comparing the query with BoW and counting the number of keywords in the inquiries to calculate and find the terms weight, as shown in Table 3.

The result is kept in the Corpus with the inverted indexing system for faster retrieval. It consists of three parts: 1) dictionary, 2) posting in the main memory unit, and 3) pointer to point out various announcements kept in the second main memory unit. Keywords will be arranged according to the letters with the frequency of the terms found in all documents. Similar words will be counted in the frequency altogether, called Dictionary, as shown in Table 4.

Table 4. An example of term vector

Class	IDF score	TF-IDF score
E	'แหล่งท่องเที่ยว': 1.14568, 'ห้องน้ำ': 1.74027, 'สะอาด': 1.65780, …'ชุมชน': 1.07559	'แหล่งท่องเที่ยว': 0.00703, 'ห้องน้ำ': 0.01068, 'สะอาด': 0.010178, …'ชุมชน': 0.00660
S	'กิจกรรม': 1.51923, 'การท่องเที่ยว': 1.0, 'เอกลักษณ์': 1.12745, ...'ชุมชน': 1.07559	'กิจกรรม': 0.01356, 'การท่องเที่ยว': 0.00893, 'เอกลักษณ์': 0.01007, …'ชุมชน': 0.00960
A	'แหล่งท่องเที่ยว': 1.14568, 'ระเบียบ': 1.0, 'ข้อปฏิบัติ': 2.77270, … 'การสื่อสาร': 1.0	'แหล่งท่องเที่ยว': 0.00606, 'ระเบียบ': 0.00529, 'ข้อปฏิบัติ': 0.01467, …'การสื่อสาร': 0.00529
N	'ชุมชน': 1.07559, 'แหล่งท่องเที่ยว': 1.14568, 'หลากหลาย': 1.86930, … 'ประเพณี': 2.09194	'ชุมชน': 0.00717, 'แหล่งท่องเที่ยว': 0.00764, 'หลากหลาย': 0.01246, …'ประเพณี': 0.01395

Cosine Similarity [21]. It is a statistical method to measure the similarity by angle or vector distance between the two documents. It is an indicator of cosine similarity. The method brings the vectors of any two data in the 'i' dimensional space to calculate the dot product and determine the angle value at such two vectors. The value of cosine between vector angles is between 0 and 1. The two vectors will get a similarity if the angle gets a minimum value. The cosine value of two vectors can be calculated by Equation (3), as shown in Table 5.

$$cossim(d_1, d_2) = \frac{d_1 \cdot d_2}{\|d_1\| \|d_2\|} = \frac{\sum_i (d_{i1} * d_{i2})}{\sqrt{d_{i1}^2} * \sqrt{d_{i2}^2}} \tag{3}$$

where d_{i1} and d_{i2} are the ith components of vectors d_1 and d_2, respectively.

Table 5. The result of the top cosine similarity of each rule in 4 classes

Rules	Class E		Class S		Class A		Class N	
	doc_id	cosine	doc_id	cosine	doc_id	cosine	doc_id	cosine
1	325	0.892	64	0.872	81	0.867	22	0.745
2	181	0.863	133	0.820	81	0.840	79	0.713
3	452	0.874	165	0.943	173	0.771	122	0.805
4	416	0.867	135	0.958	–	–	–	–

4. **Experiment and Result.**

We develop and demonstrate the framework of an Open-QA automatic question and answering for community tourism association data, a case study of the South Northeastern Region, Thailand emphasizes creating an information retrieval model in a text form to be the guideline for problem-solving regarding the data access of the tourists and the involved efficiently using Google colab by Python programming language. It performs based on various models by selecting the outstanding points of each model for a balanced application. The procedures consist of five steps, as shown in Figure 3, and the final step is to generate confidence before the authentic use. It can separate the documents relevant to the query from the irrelevant ones. The result will be in order. This model effectiveness evaluation is performed in 1) expert and 2) system. The result was in the same direction and considered a term in the query that is rare in the document collection.

For the crucial objectives in evaluating the competency of the model to find the relationship between automatic questions and answering for community tourism by the concept of textual information retrieval, this research assesses the model's effectiveness in two parts: 1) three experts' evaluation, all in 3 aspects consist of tourism, data mining, and linguistics. The result of this part is concordant with the system evaluation, and 2) the system evaluation includes P, R, $Acc.$, and $F1$, to benefit the readers. The researcher represents the experimental result, the effectiveness of the cosine similarity measure, and the top cosine similarity of each rule in 4 classes.

Table 5 shows the result of the top 5 cosine similarity measures in 4 classes; the experimental result found the documents that had the relationship according to the users' retrieval. The model could measure the similarity of the queries with the documents. Therefore, Figure 5 indicates the ability to describe the answers and effectiveness of the top cosine similarity of each rule in 4 classes. The models randomly selected the documents from 4 classes: 1) E (524 docs), 2) S (267 docs), 3) A (201 docs), and 4) N (245 docs). The results between the query and the answer were compared and computed using statistical methods into four classes.

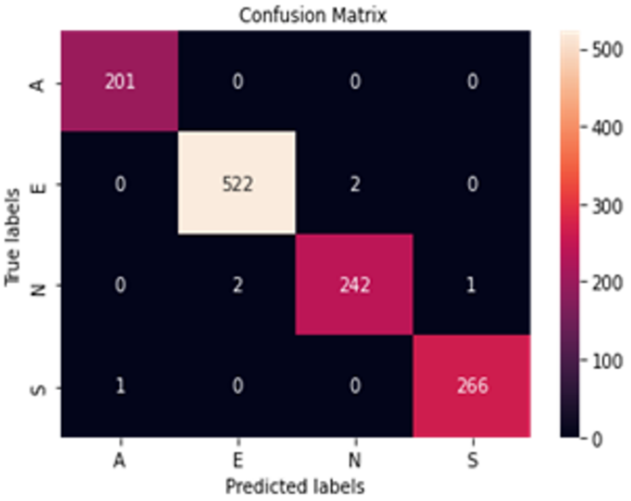
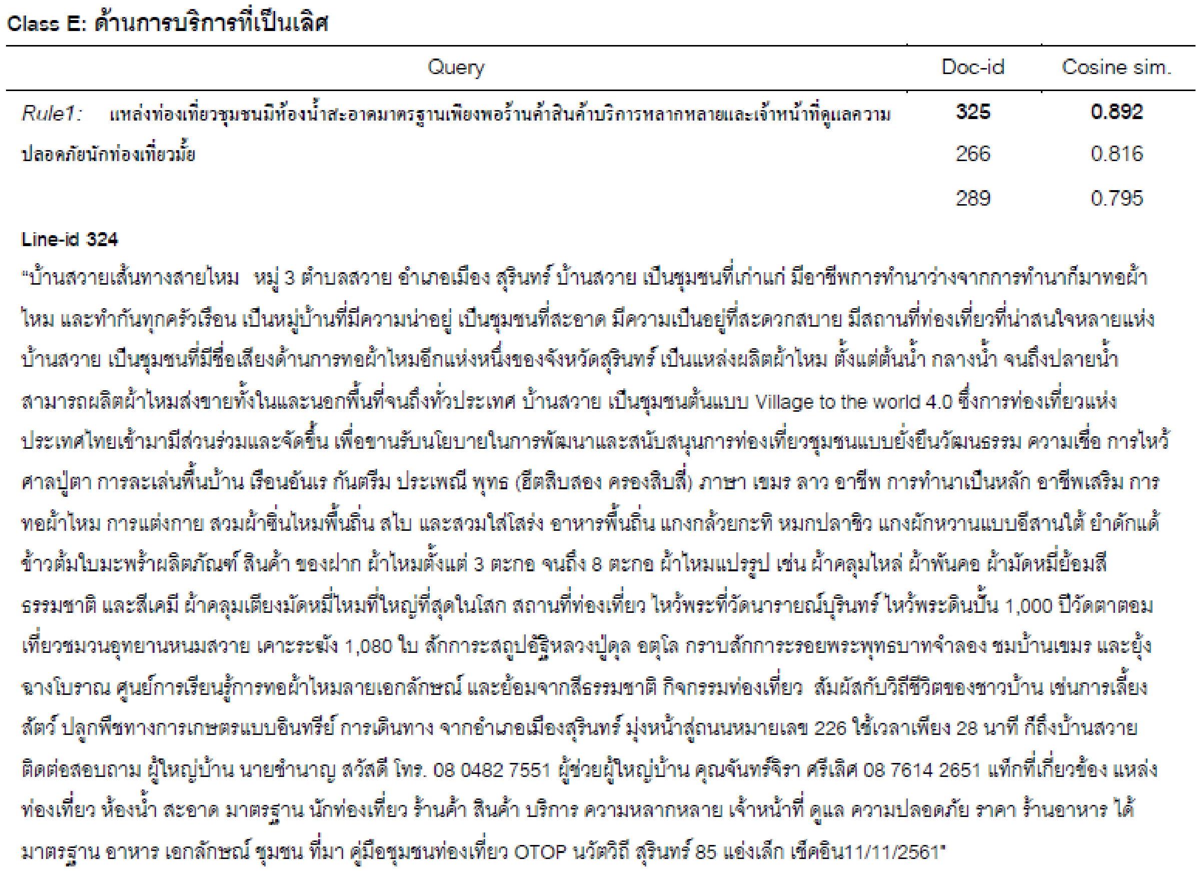


FIGURE 5. Confusion matrix

The result revealed that 1) Class A correctly predicted all 201 documents. 2) Class N predicted the most inaccurately because the query does not match the answer and only features are close to the question; the result will be considered a term in the query that is rare in the document collection, for example: query; "การท่องเที่ยว (karn thong theiw)", feature; "เที่ยว (theiw)", "ท่องเที่ยว (thong theiw)", "สถานที่ท่องเที่ยว (satharn thi thong theiw)", and "แหล่งท่องเที่ยว (hlaeng thong theiw)". For example, OpenQA is as follows.



Class E: ด้านการบริการที่เป็นเลิศ

Query	Doc-id	Cosine sim.
Rule1: แหล่งท่องเที่ยวชุมชนมีห้องน้ำสะอาดมาตรฐานเพียงพอร้านค้าสินค้าบริการหลากหลายและเจ้าหน้าที่ดูแลความปลอดภัยนักท่องเที่ยวมั้ย	**325**	**0.892**
	266	0.816
	289	0.795

Line-id 324

"บ้านสวายเส้นทางสายไหม หมู่ 3 ตำบลสวาย อำเภอเมือง สุรินทร์ บ้านสวาย เป็นชุมชนที่เก่าแก่ มีอาชีพการทำนาว่างจากการทำนาก็มาทอผ้าไหม และทำกันทุกครัวเรือน เป็นหมู่บ้านที่มีความน่าอยู่ เป็นชุมชนที่สะอาด มีความเป็นอยู่ที่สะดวกสบาย มีสถานที่ท่องเที่ยวที่น่าสนใจหลายแห่ง บ้านสวาย เป็นชุมชนที่มีชื่อเสียงด้านการทอผ้าไหมอีกแห่งหนึ่งของจังหวัดสุรินทร์ เป็นแหล่งผลิตผ้าไหม ตั้งแต่ต้นน้ำ กลางน้ำ จนถึงปลายน้ำ สามารถผลิตผ้าไหมส่งขายทั้งในและนอกพื้นที่จนถึงทั่วประเทศ บ้านสวาย เป็นชุมชนต้นแบบ Village to the world 4.0 ซึ่งการท่องเที่ยวแห่งประเทศไทยเข้ามามีส่วนร่วมและจัดขึ้น เพื่อขานรับนโยบายในการพัฒนาและสนับสนุนการท่องเที่ยวชุมชนแบบยั่งยืนวัฒนธรรม ความเชื่อ การไหว้ศาลปู่ตา การละเล่นพื้นบ้าน เรือนอันเร กันตรึม ประเพณี พุทธ (ฮีตสิบสอง ครองสิบสี่) ภาษา เขมร ลาว อาชีพ การทำนาเป็นหลัก อาชีพเสริม การทอผ้าไหม การแต่งกาย สวมผ้าซิ่นไหมพื้นถิ่น สไบ และสวมใส่โสร่ง อาหารพื้นถิ่น แกงกล้วยกะทิ หมกปลาซิว แกงผักหวานแบบอีสานใต้ ยำดักแด้ ข้าวต้มใบมะพร้าวผลิตภัณฑ์ สินค้า ของฝาก ผ้าไหมตั้งแต่ 3 ตะกอ จนถึง 8 ตะกอ ผ้าไหมแปรรูป เช่น ผ้าคลุมไหล่ ผ้าพันคอ ผ้ามัดหมี่ย้อมสีธรรมชาติ และสีเคมี ผ้าคลุมเตียงมัดหมี่ไหมที่ใหญ่ที่สุดในโลก สถานที่ท่องเที่ยว ไหว้พระที่วัดนารายณ์บุรินทร์ ไหว้พระดินปั้น 1,000 ปีวัดตาตอม เที่ยวชมวนอุทยานหนมสวาย เคาะระฆัง 1,080 ใบ สักการะสถูปอัฐิหลวงปู่ดุล อตุโล กราบสักการะรอยพระพุทธบาทจำลอง ชมบ้านเขมร และยุ้งฉางโบราณ ศูนย์การเรียนรู้การทอผ้าไหมลายเอกลักษณ์ และย้อมจากสีธรรมชาติ กิจกรรมท่องเที่ยว สัมผัสกับวิถีชีวิตของชาวบ้าน เช่นการเลี้ยงสัตว์ ปลูกพืชทางการเกษตรแบบอินทรีย์ การเดินทาง จากอำเภอเมืองสุรินทร์ มุ่งหน้าสู่ถนนหมายเลข 226 ใช้เวลาเพียง 28 นาที ก็ถึงบ้านสวาย ติดต่อสอบถาม ผู้ใหญ่บ้าน นายชำนาญ สวัสดี โทร. 08 0482 7551 ผู้ช่วยผู้ใหญ่บ้าน คุณจันทร์จิรา ศรีเลิศ 08 7614 2651 แท็กที่เกี่ยวข้อง แหล่งท่องเที่ยว ห้องน้ำ สะอาด มาตรฐาน นักท่องเที่ยว ร้านค้า สินค้า บริการ ความหลากหลาย เจ้าหน้าที่ ดูแล ความปลอดภัย ราคา ร้านอาหาร ได้มาตรฐาน อาหาร เอกลักษณ์ ชุมชน ที่มา คู่มือชุมชนท่องเที่ยว OTOP นวัตวิถี สุรินทร์ 85 แฮ่งเล็ก เช็คอิน11/11/2561"

Table 6 shows the summary results of model evaluation in 4 classes. The target is retrieving tourism information from four classes: 1) E, 2) S, 3) A, and 4) N. Considering all, it was found that two classes, consisting of E and A, could describe the best answers and effectiveness of the created model in every class. Hence, it is for describing the Term-Document Matrix (TDM) and creating the index for retrieval. Each unit in the matrix is the weight value relying upon the term frequency found in two documents for comparison. The results of inquiries and answering papers are sequenced from words and similarities. It gave the Accuracy at 99%. Furthermore, it increased the effectiveness in three crucial values: P, R, and $F1$. It started from the class with the highest value of three classes consisting of E, S, and A; the value all at one and class N; the value was 0.99, 0.99, and 0.99, respectively.

Table 6. The result of the confusion matrix and model evaluation is in 4 classes

Class	Confusion matrix				Model evaluation		
	E	S	A	N	P	R	$F1$
E	522	2	0	0	1	1	1
S	0	266	1	0	1	1	1
A	0	0	201	0	1	1	1
N	2	1	0	242	0.999	0.99	0.99
Accuracy				0.99			
Macro avg			0.99	1			
Weighted avg		1	1	1			

5. **Conclusion.** The efficient data access of the tourists and the involved is a challenge of this research. The methods to find the similarity by the cosine similarity technique based on the VSM bring the outstanding points of the text vectorization to calculate to acquire the crucial features for being the document representatives efficiently [23]. The

result in the BoWs stage is that the Bag of Words is 19,501 terms in 1,237 documents. The evaluation of model effectiveness has an Accuracy value of 99%, which best indicates the ability to describe the answer and efficacy of the model. Each of them would get the various features and information and be independent from each other [24]. Therefore, it is appropriate with the unstructured data [25] created by the research team and analysis. It could describe the answers and evaluate the generated retrieval model effectively. Unlike other research, our data comes from association form, which can predict the following question and answer. Our challenge is Thai word segmentation and how to select the word using selected weighting and BoWs. Compared with the QAs in Thai language, this paper has an Accuracy of 99% for 4 classes. Hence, the limitation of BoWs models encodes every word in the vocabulary as a one-hot-encoded vector. As vocabulary may run into millions, bag-of-word models face scalability challenges. It can be solved with factor analysis.

Acknowledgment. This research project was financially supported by Mahasarakham University.

REFERENCES

[1] B. Bramwell and B. Lane, Sustainable tourism, progress, challenges and opportunities: An introduction, *Journal of Cleaner Production*, vol.111, pp.285-294, DOI: 10.1016/j.jclepro.2015.10.027, 2015.

[2] O. S. Kolbasov, UN conference on environment and development, *Izv. – Akad. Nauk. Seriya Geogr*, vol.6, no.6, pp.47-54, 1992.

[3] C. Jareanpon and K. Kawattikul, Book cover and content similarity retrieval using computer vision and NLP techniques *Multi-Disciplinary Trends in Artificial Intelligence*, pp.34-44, 2021.

[4] V. López, V. S. Uren, M. Sabou and E. Motta, Is question answering fit for the semantic web?: A survey, *Semantic Web*, vol.2, pp.125-155, 2011.

[5] E. M. Voorhees, *The Trec-8 Question Answering Track Report*, NIST, Tech. Rep., 1999.

[6] E. Brill, S. Dumis and M. Banko, An analysis of the AskMSR question-answering system, *Proc. of the 2002 Conference in Empirical Methods in Natural Language Processing (EMNLP 2002)*, Association for Computational Linguistics, pp.257-264, 2002.

[7] C. D. Manning, *An Introduction to Information Retrieval*, Online Edition, Cambridge University, 2009.

[8] O. Kolomiyets and M. F. Moens, A survey on question answering technology from an information retrieval perspective, *Inf. Sci. (Ny).*, vol.181, no.24, pp.5412-5434, 2011.

[9] U. Chaisoong, S. Tirakoat and C. Jareanpon, Tourist information-seeking behaviours using association rule mining, *ICIC Express Letters*, vol.15, no.9, pp.915-923, 2021.

[10] U. Chaisoong and S. Tirakoat, The clustering of questions affect to tourist's decision making for chatbot design, *The 17th Int. Conf. Electr. Eng. Comput. Telecommun. Inf. Technol. (ECTI-CON 2020)*, pp.784-787, 2020.

[11] C. Haruechaiyasak, S. Kongyoung and M. N. Dailey, A comparative study on Thai word segmentation approaches, *2008 5th International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology*, 2008.

[12] T. Horsuwan, K. Kanwatchara, P. Vateekul and B. Kijsirikul, A comparative study of pretrained language models on Thai social text categorization, *Lecture Notes in Computer Science*, vol.12033 LNAI, pp.63-75, 2020.

[13] P. Saipech and P. Seresangtakul, Automatic Thai subjective examination using cosine similarity, *The 5th Int. Conf. Adv. Informatics Concepts Theory Appl. (ICAICTA 2018)*, pp.214-218, 2018.

[14] C. Sirichanya and K. Kraisak, Semantic data mining in the information age: A systematic review, *Int. J. Intell. Syst.*, vol.36, no.8, pp.3880-3916, 2021.

[15] M. Eklund, Comparing feature extraction methods and effects of pre-processing methods for multi-label classification of textual data, *Degree Proj. Comput. Sci. Eng.*, 2018.

[16] R. K. Mishra, S. Urolagin and A. A. Jothi J, A sentiment analysis-based hotel recommendation using TF-IDF approach, *2019 International Conference on Computational Intelligence and Knowledge Economy (ICCIKE)*, pp.811-815, 2019.

[17] N. Khamphakdee and P. Seresangtakul, A framework for constructing Thai sentiment corpus using the cosine similarity technique, *2021 13th International Conference on Knowledge and Smart Technology (KST 2021)*, pp.202-207, 2021.

[18] W. Phatthiyaphaibun, K. Chaovavanich, C. Polpanumas, A. Suriyawongkul, L. Lowphansirikul and A. Chormai, PyThaiNLP: Thai natural language processing in python, *PyThaiNLP Documentation*, 2016, https://pythainlp.github.io/docs/2.3/, Accessed on Mar. 18, 2022.

[19] K. Kesorn, *Information Retrieval System: Concepts and Future Directions*, 1st Edition, Department of Computer Science and Information Technology, Naresuan University, Phitsanulok, 2015.

[20] J. Intasorn, S. Gertphol and U. Sammapun, Thai sentiment lexicon construction, *2021 13th International Conference on Knowledge and Smart Technology (KST 2021)*, pp.123-128, 2021.

[21] K. Kowsari, K. J. Meimandi, M. Heidarysafa, S. Mendu, L. Barnes and D. Brown, Text classification algorithms: A survey, *Inf.*, vol.10, no.4, pp.1-68, 2019.

[22] P. Netisopakul and K. Thong-iad, Thai sentiment resource using Thai WordNet, in *Complex, Intelligent, and Software Intensive Systems. CISIS 2018. Advances in Intelligent Systems and Computing*, vol.772, L. Barolli, N. Javaid, M. Ikeda and M. Takizawa (eds.), Cham, Springer, DOI: 10.1007/978-3-319-93659-8_29, 2019.

[23] W. Chen, Z. Xu, X. Zheng, Q. Yu and Y. Luo, Research on sentiment classification of online travel review text, *Appl. Sci.*, vol.10, no.15, 5275, 2020.

[24] P. Meesad, Thai fake news detection based on information retrieval, *Natural Language Processing and Machine Learning*, vol.2, 425, 2021.

[25] O. Shahmirzadi, A. Lugowski and K. Younge, Text similarity in vector space models: A comparative study, *Proc. of the 18th IEEE Int. Conf. Mach. Learn. Appl. (ICMLA 2019)*, pp.659-666, DOI: 10.1109/ICMLA.2019.00120, 2019.