ICIC Express Letters
Volume **12**, Number **2**, February **2018**
 ISSN 1881-803X
pp. **159–165**

THAI NATURAL LANGUAGE BASED CULTURAL TOURISM ONTOLOGY

Anongporn Salaiwarakul

Department of Computer Science and Information Technology
Naresuan University
Muang, Phitsanulok 65000, Thailand
anongporns@nu.ac.th

Received August 2017; accepted November 2017

Abstract. *We propose a cultural tourism ontology for Thailand which enables users to access tourism information using natural language queries in the Thai language. This research shows the value of using an ontology of core knowledge of a subject domain, able to be semantically searched by natural language queries. Given the importance of the tourism industry's contribution to the Thai economy, this is seen as a valuable contribution. Coupled with the lexical difficulty of converting Thai language text into a useable query, the availability of such an ontology and the natural language querying capability overcomes the significant limitations of keyword-based searches. This paper shows the ontology design which is developed using the Hozo editor. Evaluating the ontology and knowledge base was done by precision, recall and F-measure methods. The result from the evaluation demonstrated that our cultural tourism ontology, knowledge base and the proposed technique used to retrieve information to satisfy the user's natural language query are both accurate and efficient.*
Keywords: Cultural tourism ontology, Information retrieval, Thai natural language processing, Semantic web

1. **Introduction.** Tourism is a significant industry in most countries as a major contributor to the economy. Thailand, like many other countries with a long history and cultural heritage, utilizes its historical assets to earn revenue from tourism. Cultural tourism assets can be divided into three major categories, based on United Nation World Tourism Organization (UNWTO) [1] definition: historical tourism, cultural and traditional tourism, and rural and village tourism. Historical tourism includes destinations with archaeological themes and historical information, while cultural and traditional tourism shows the tourist local ceremonies and rites, providing them with new experience and understanding of different beliefs and religious practices. Rural and village tourism shows local folk wisdom, social practices and mores and local intellectual customs and understanding. Tourist attractions are diverse, usually varying significantly from place to place.

Huge amounts of information are available on the Internet, and users have easy access to this information through powerful and sophisticated search facilities. However, the amount of information can be overwhelming and access to the information can actually be severely restricted by users' inability to form queries that can reach the precise information required due to a number of factors, particularly including the redundancy of the information and the imprecision of the user's query text. To overcome these limitations, queries in natural language against a specific subject domain are essential. The availability of an ontology containing the specific domain information together with semantic searching based on natural language queries is one answer to the problem. The specific subject domain addressed in our research was cultural tourism, and the specific natural language selected is Thai. Natural language processing of Thai must overcome the significant obstacle inherent in written Thai text which is the characteristic continuity of the

text without indications of individual words, such as spaces between words as is the case in European languages.

Our paper is structured as a review of the related work in Section 2, which is followed by a discussion of our research methodology in Section 3, in which we show examples of various tourism ontologies together with the ontology which we designed and developed, together with the processes in our system for natural language processing of user queries. Section 4 comprises an evaluation of our ontology and demonstrates our approach to evaluating our ontology and the results achieved. A discussion on our conclusions completes the article.

2. **Literature Review.** A review of recently published research on related issues of natural language processing, semantic searching, ontology development, and discussions on matters related to tourism and the tourist industry, is illustrated in this section.

The study of the literature identified various ontologies relevant to the tourism industry which had previously been developed [2-4], and different tourism ontologies are discussed in [5]. From our perspective, each ontology focuses on a particular knowledge domain with the general knowledge base of tourism, such as cultural tourism, health and medical tourism, nature and environmental tourism or archaeological tourism, to name a few. Each domain has its own particular knowledge base. One specific, focused ontology is a cultural tourism ontology for a specific area of Thailand, the Dusit district of Bangkok, which consists of the classes 'temples', 'historical buildings', 'palaces' and so forth [6]. However, we consider that this ontology does not have a sufficiently comprehensive knowledge base of required information that serves the definition of cultural tourism attractions.

A cultural tourism ontology which stored a substantially more comprehensive set of information appropriate to Thailand was proposed and discussed in [7]. However, the limitations of analyzing and responding meaningfully to user queries were evident, particularly the 'keyword search' approach to satisfying user queries. Our solution to this problem was to develop the capability of natural language processing of user's queries which we consider to be crucial for responding to user queries and providing meaningful responses by way of semantic searching but with an easy to use, user-friendly interface.

Our position is that natural language processing is the key to providing ease of use and for searching information from the Internet. However, natural language processing of written Thai language text is the most problematic in that Thai script does not use word separators, such as spaces, but is a continuous stream of characters. This makes the identification of individual words difficult [8,9].

There are a number of approaches extant to natural language processing, including the rule base approach, the dictionary based approach, or a Corpus based approach. The rule based approach uses a set of defined rules to extract words from a given sentence [10,11]. This approach is good for extracting syllables but the ability to extract whole words is low. The dictionary approach can overcome the problem of multi-syllable words and the ambiguity of words in Thai that have different meanings depending on textual context and method of extraction, as has been discussed in several research papers [12,13]. However, the detection of words correctly and accurately requires access to a sufficiently comprehensive dictionary. The corpus based approach uses statistics on the frequency of word usage in a particular corpus as the knowledge base for word segmentation [14,15]. However, the level of correctness in word segmentation depends on the size of the corpus.

Two main aspects of tourism information in Thailand were identified in the literature which should be considered for the purpose of building tourism income generating potential and gaining further benefits for the country. To achieve this, the novelty of the research area must be extended. The first aspect is that cultural tourism information available in Thai language is formed in such a semantic way that it is inadequate to

fulfill the information requirements of Thai tourist. The second aspect is that the semantic searching capability of this information base, which should be done in a natural manner, with information queries being expressed in natural language, has significant limitations imposed due to the lexical and textual nature of the Thai language, particularly the usual style of Thai text which does not include word differentiation tokens, such as the space which is used in English. This makes applying the existing natural language techniques, especially now those being applied to publishing tourism information, problematic. Tourism 'attraction words' are difficult to identify and thus render keyword searching difficult and not especially successful. To the best of our knowledge, little has been done to acknowledge and overcome the problems presented by both of these aspects of tourism information, particularly that which is available in Thai language.

3. **Research Methodology.** This section shows our research methodology that was used to generate a cultural tourism ontology. The ontology that we developed is designed to provide all information that a tourist may need to decide on potential destinations, travel information, accommodation and so on. The ontology is stored as Thai text specifically.

3.1. **Knowledge base categorization.** We first categorized the tourism attractions based on the UNWTO definitions (as described briefly in Section 1) which provide a standard on which the structure and content of our tourism ontology are based, following the categorizations of tourism in that standard. Particularly, the classification of the category of cultural tourism and its definition and scope of attractions are in accordance with the UNWTO definition. As shown in Figure 1, the tourism information is collected, analyzed and categorized as a knowledge base for tourism ontology.

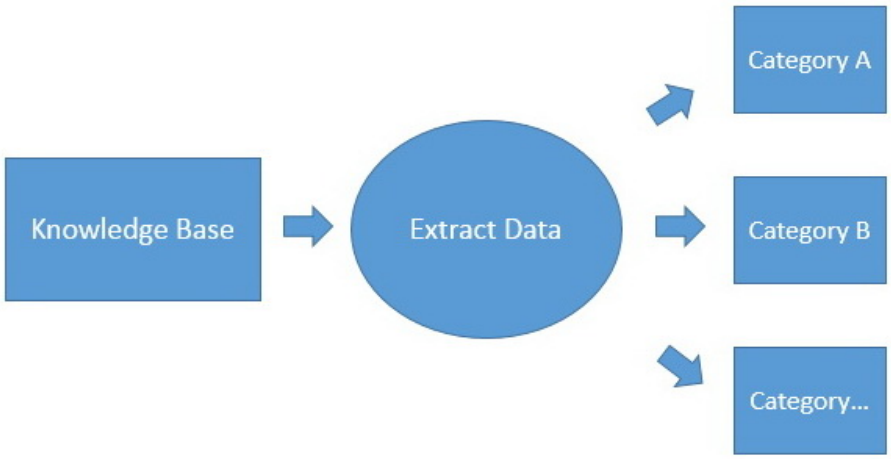

FIGURE 1. The tourism knowledge base is categorized.

The currently available tourism ontologies were studied and analyzed for comparison, and used to establish guidelines for the development of our ontology to ensure our ability to communicate with other ontologies and knowledge bases, and to enable the exchange of information between other ontologies and knowledge bases. This necessarily requires standardization of semantic meanings to be able to respond to user queries. Of particular interest here are a number of currently available tourism/travel ontologies which we analyzed, described in [16,17].

As indicated previously, we designed our cultural tourism ontology based on the definitions, classes and sub-classes included in these various currently available ontologies. Figure 2 shows the particular subclass of Attraction_type in which the major types of cultural tourism attractions are included. The diagram was developed with the Hozo-ontology editor.

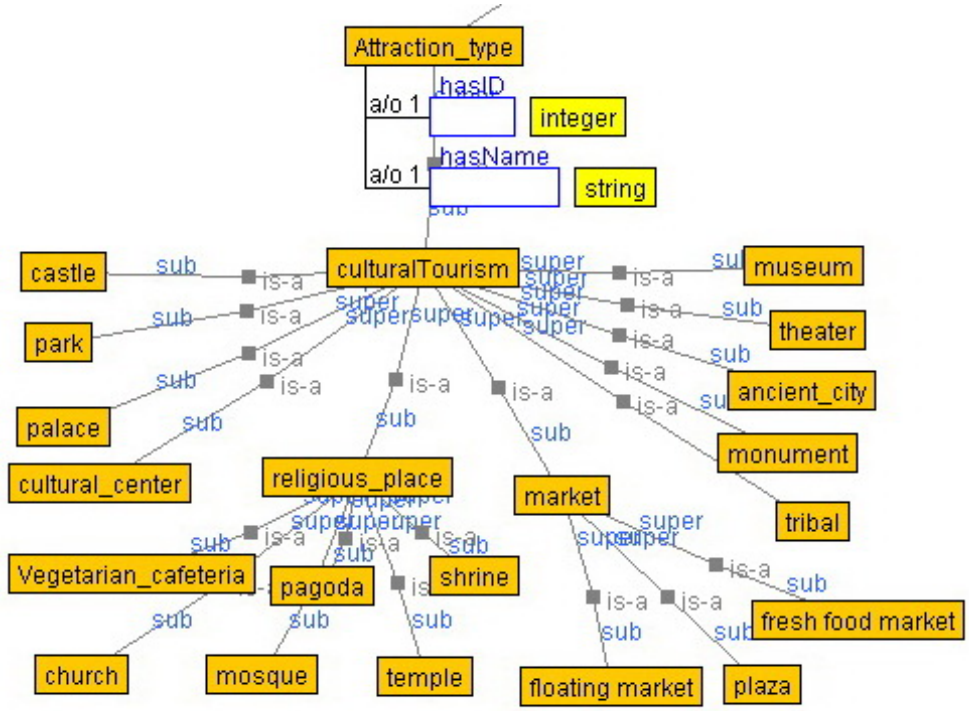

FIGURE 2. Subclass of the cultural tourism attractions

3.2. **Processing natural language to query language.** In order to retrieve information from the ontology, the SPARQL is employed. The SPARQL is a language for querying the stored data in the database semantically, in such a way as to enable the retrieval of correct information that satisfies the query. The user's query is analyzed and word segments are identified which are converted into a query in SPARQL. We used a graph technique to retrieve information stored in the ontology appropriate to the query. We constructed a tree-like graph which has a root which is a class that stores the answer to the user query. The tree-like graph is conducted as

$$(R\{E_1E_2E_3\ldots E_N\}) \tag{1}$$

where R is the root and E is the class property or literal.

For example, the query such as, "โรงแรมทอปแลนด์เบอร์โทรอะไร", (In English, "What is the telephone number of Topland Hotel?"). The Thai text can be segmented into its component 'word segments' which mean, in English 'Topland Hotel', 'telephone number', and 'what?' ('โรงแรมทอปแลนด์', 'เบอร์โทร' and 'อะไร'). Word segment 1 can be directly mapped into the exact same literal in the ontology (*literal: โรงแรมทอปแลนด์*, meaning "Topland Hotel") and Word segment 2 (เบอร์โทร) can be mapped with the *Property: hasTelephoneNo*. Word segment 3 (what?) is interpreted as the interrogative.

The SPARQL language from the above user's query is constructed as shown in Table 1.

The tree-like graph which illustrates this example is shown in Figure 3.

3.3. **Web ontology language.** The tourism information which is stored in the database will be wrapped using OWL (Web Ontology Language), based on the ontology schema design. The example of OWL which is transformed from our ontology is shown in Table 2.

4. **The Ontology Evaluation.** The effectiveness of the information retrieval from the ontology and knowledge base is evaluated using the F-measure method, the values from which are calculated using precision and recall values. The precision value refers to the

TABLE 1. SPARQL language example

Example of query language by SPARQL

PREFIX myonto = "http://www.myontology.com/tourism"
SELECT ?X
WHERE {
?Accommodation myonto:hasTelephoneNo ?X
FILTER (?name = "โรงแรมท็อปแลนด์")
}

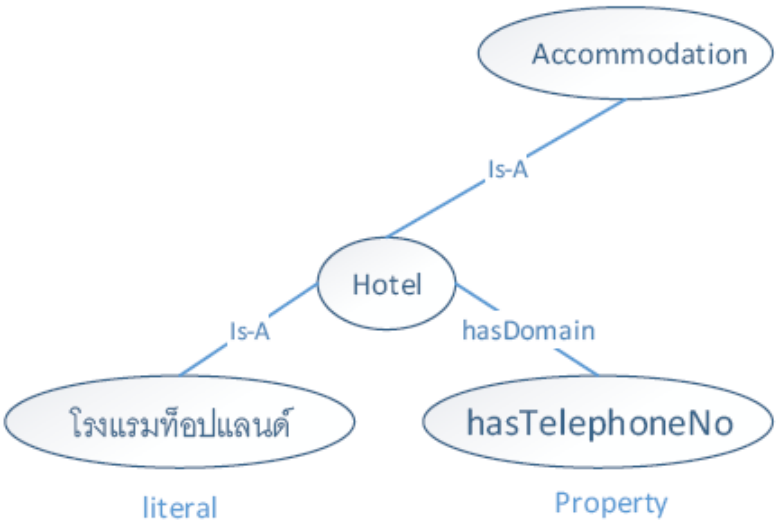

FIGURE 3. The tree-like mapping graph

TABLE 2. OWL example

```
<owl:Class rdf:about="http://thtourism.com/schema#attraction"/>
<owl:Class rdf:about="http://thtourism.com/schema#accomodation"/>
<owl:ObjectProperty rdf:about="http://thtourism.com/schema#nearby">
        <rdfs:domain rdf:resource="http://thtourism.com/schema#attraction"/>
        <rdfs:range rdf:resource="http://thtourism.com/schema#accomodation"/>
</owl:ObjectProperty>
<owl:DatatypeProperty rdf:about="http://thtourism.com/schema#name">
        <rdfs:domain rdf:resource="http://thtourism.com/schema#attraction"/>
        <rdfs:domain rdf:resource="http://thtourism.com/schema#accomodation"/>
        <rdfs:range rdf:resource="&xsd;string"/>
</owl:DatatypeProperty>
```

ratio of the number of retrieved elements that precisely match elements in the user's query text and the total number of retrieved elements in the query text. The recall is the ratio of the number of correct elements retrieved and the total number of correct elements in the knowledge base. The equations of the precision and recall are calculated in Equations (2) and (3).

$$Precision = \frac{tp}{tp + fp} \tag{2}$$

$$Recall = \frac{tp}{tp + fn} \tag{3}$$

where tp is the total number of elements retrieved correctly, relevant to the query. fp is the total number of elements retrieved but not relevant to the query. fn is the total number of elements relevant to the query but not retrieved.

According to the above equations, precision indicates the ability of the system to eliminate information which is irrelevant to the query, while the recall value indicates the ability of the system to retrieve information relevant to the query. Table 3 gives examples of query sentences and the returned results from the experiment, relevant to measuring precision and recall.

TABLE 3. Example of the experiment result from one individual complex sentence

Examples of the complex sentences	**Precision**	**Recall**
Accommodation in type ***hotel*** near "Ja-Ta-Wee" museum	1	1
Attraction in type ***temple*** which is located in Amphoe Muang, Phitsanulok	1	0.9
Type ***tent accommodation*** in Phitsanulok province	1	0.85
Attractions in type ***religious place*** or ***temple*** in Phitsanulok	1	1

In Table 3 it can be seen that the system cannot retrieve all temples located in Amphoe Muang, Phitsanulok stored in the knowledge base, but it found 90% of the information on temples located in that location. Nonetheless, the system did correctly identify that information on temples located in Amphoe Muang, Phitsanulok was required.

We evaluated the accuracy of the ontology against the user's query by testing with 50 complex sentences. The result of a precision value of 1 in each example sentence shows that there are no false results in the retrieved information while the recall value of 0.89 (from 50 complex sentences).

We can represent the overall accuracy of the query response using the F-measure value which shows the accuracy and effectiveness of retrieving the information stored in the ontology, expressed as a percentage. This calculation is done using Equation (4). The result from the calculation illustrates that the accuracy in retrieving information from the knowledge base is 94%.

$$F = 2 \cdot \frac{precision \cdot recall}{precision + recall} \tag{4}$$

These results from the evaluation show that our designed technique and ontology can retrieve information requested in Thai natural language with a high degree of success. This was not achieved in the research previously published which we identified in our literature review.

5. **Conclusions.** This paper describes a cultural tourism ontology which we developed which is based on the definitions of the UNWTO and on other tourism ontologies, allowing interoperation with other tourism ontologies. Natural language processing of Thai language queries using a semantic search approach is a feature of the ontology, enabling particularly Thai tourists to query the desired information in a user-friendly manner. This is a significant improvement over keyword searches given the variety of cultural attractions and information on those attractions which varies from region to region, often based on local language usage and naming conventions. We are confident of the contribution of this research to the tourism industry which makes information more accessible to the tourist, and which will allow tourists to successfully search for their information and be provided with more exact information with fast responses. By making online information more accessible in this way will contribute to the economies of local regions and enhance the business success of local tour and service providers. Further work is recommended to analyze the efficiency of each natural language processing method which would result in

greater accuracy in extracting and segmenting words in a given sentence in Thai natural language. Improvements in this would affect the accuracy of retrieving information.

REFERENCES

[1] *United Nations World Tourism Organization (UNWTO)*, http://www.wto.org.

[2] O. Daramola, M. Adigun and C. Ayo, Building an ontology-based framework for tourism recommendation services, *ENTER 2009*, Amsterdam, Netherlands, pp.135-147, 2009.

[3] S. Mouhim, A. E. Aoufi et al., A knowledge management approach based on ontologies: The case of tourism, *Int. J. Comput. Sci. Emerg. Technol.*, vol.4, no.3, pp.362-369, 2011.

[4] S. Ou, V. Pekar, C. Orasan, C. Spurk and M. Negri, Development and alignment of a domain-specific ontology for question answering, *Proc. of the 6th International Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco, pp.2221-2228, 2008.

[5] M. Archana, K. Akshatha, S. Apoorva and J. Anitha, A survey on existing tourism ontologies, *IJRET: International Journal of Research in Engineering and Technology*, vol.4, no.14, pp.20-23, 2015.

[6] N. Tachapetpaiboon and K. Kularbphettong, Ontology based knowledge management for cultural tourism, *Journal of Theoretical and Applied Information Technology*, pp.384-388, 2015.

[7] A. Salaiwarakul, A cultural tourism ontology for lower northern Thailand, *KMUTNB Int. J. Appl. Sci. Technol.*, vol.10, no.1, pp.1-6, 2017.

[8] C. Wutiwiwatchai, P. Mittrapiyanuruk, T. Potipiti and V. Sornlertlamvanich, The state of the art in Thai language processing, *Proc. of the 38th Annual Meeting on Association for Computational Linguistics*, pp.1-2, 2000.

[9] P. Sojka and D. Antoš, Context sensitive pattern based segmentation: A Thai challenge, *Proc. of EACL Workshop on Computational Linguistics for South Asian Languages – Expanding Synergies with Europe*, Budapest, Hungary, pp.65-72, 2003.

[10] T. Mahmud, K. M. Azharul Hasan, M. Ahmed and T. H. C. Chak, A rule based approach for NLP based query processing, *The 2nd International Conference on Electrical Information and Communication Technologies (EICT)*, Khulna, pp.78-82, 2015.

[11] J. Kazimierczak, An approach to natural language processing in the rule-based expert system, *ACM Annual Conference on Cooperation*, pp.215-222, 1990.

[12] M. Lesk, Automatic sense disambiguation using machine readable dictionaries: How to tell a pine cone from an ice cream cone, *Proc. of the 5th Annual International Conference on Systems Documentation*, Toronto, Ontario, Canada, pp.24-26, 1986.

[13] S. Liu et al., An effective approach to document retrieval via utilizing WordNet and recognizing phrases, *Proc. of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp.266-272, 2004.

[14] H. T. Ng and J. Zelle, Corpus-based approaches to semantic interpretation in natural language processing, *AI Magazine*, vol.18, no.4, pp.45-64, 1997.

[15] S. Armstrong, Corpus based methods for NLP and translation studies, *Interpreting: International Journal of Research and Practice in Interpreting*, vol.2, no.1, pp.141-162, 1997.

[16] K. Prantner, OnTour: The ontology, *Deri Insbruck*, http://ontour.deri.org/ ontology/ontour-02.owl, 2004.

[17] M. Dell'Erba, O. Fodor, F. Ricci and H. Werthner, Harmonise: A solution for data interoperability, *Proc. of the IFIP Conference on Towards the Knowledge Society: E-Commerce, E-Business, E-Government*, pp.433-445, 2002.