Vietnam J Comput Sci (2015) 2:85–94
DOI 10.1007/s40595-014-0033-6

REGULAR PAPER

An ontology-based approach to auto-tagging articles

Gridaphat Sriharee

Received: 6 June 2014 / Accepted: 5 November 2014 / Published online: 20 November 2014

Abstract This paper proposes an auto-tagging methodology using tags defined in the ontology. The auto-tagging methodology consists of two main processes: classification process and tag selection process. The classification process concerns semantic analysis which includes the term-weight matrix and cosine similarity. The tag selection process focuses on the selection of appropriate ontological tag—tag defined in the ontology, for the article. The ontology weight computing is proposed for tag suggestion. A technique for ontology building using blog articles is also presented as well as an extensive experiment. The experiment results show that the proposed approach is an alternative methodology for auto-tagging articles in which the obtained tag is not just the piece of text but it presents the meaning of the articles.

Keywords Tag · Tagging · Auto-tagging · Ontology

1 Introduction

Tagging is a mechanism for linking to relevant resources. Tagging is implemented in internet forums, blogs, collaboration systems (e.g., Wikipedia), and social networks (e.g., Flickr). The tag can be in-text keyword (e.g., [1], Wikipedia) or out-of-text keyword labeled by word or phrase. The in-text keyword tagging methodology focuses on some keywords in the content that may link to other resources. In contrast, the out-of-text keyword maintains tags out of the content body. Tagging in article sharing system is similar to keyword indexing of web search system. However, the tagging is focused on on-site retrieval particularly. The user specifies keyword and later the articles that are tagged with such keyword are retrieved.

In tagging management, when there are many articles posted into the system manual-tagging may take time because the administrator is required to read the content of the article and to specify relevant tags. Thus, auto-tagging is desired and it is expected in returning accurate tags to the articles and such tags should represent semantics or meaning of the article detail and may link to similar or related resources.

Ontology is an information model that provides a formal explicit specification of a shared conceptualisation of a domain [2]. Many research works use ontology as a shared information model for participant collaboration. The participants agree on the shared information model and realise the existence of things and their relations in the domain. The ontology can be seen as a controlled vocabulary model by which the terms are categorised into a hierarchy with regard to the relation of the terms. The term (called concept) in the ontology may be a generic term or a specific term. The specific term can be named entity.

This paper has the main contribution to propose ontology-based auto-tagging methodology by which out-of-text keyword is focused. The proposed auto-tagging system suggests ontological tags—terms defined by the ontology, for the articles. The auto-tagging methodology includes pre-processing and tagging process. The former is the process of data preparation for tagging. The tagging process includes classification and tag selection process. The pre-processing process creates the term-weight matrix that is used in the classification process. The term-weight matrix describes the TF-IDF weight of terms in the domains. The article is classified into relevant category by cosine similarity computing. The tag selection relies on ontology weight computing. This paper is extended from the previous paper [3] in which, an extensive experiment that applies the proposed methodology with blog

---

G. Sriharee (✉)
Department of Computer and Information Science,
King Mongkut's University of Technology North Bangkok,
1518 Pibulsongkram Road, Bangsur, Bangkok 10800, Thailand
e-mail: gridaphat.s@sci.kmutnb.ac.th

Fig. 1 An example of the content from Wikipedia

Jasmine rice (Thai: ข้าวหอมมะลิ; RTGS: Khao Hom Mali; Thai pronunciation: [kʰâːw hɔ̌ːm malíʔ]), sometimes known as *Thai fragrant rice*, is a long-grain variety of rice that has a nutty aroma and a subtle *pandan*-like (*Pandanus amaryllifolius*-leaves) flavor caused by 2-acetyl-1-pyrroline.[1] Jasmine rice is originally from Thailand. It was named as Kao Horm Mali 105 variety (KDML105) by Sunthorn Seehanern, an official of the ministry of agriculture in the Chachoengsao Province of Thailand in 1954.[*citation needed*] The grains will cling when cooked, though it is less sticky than other rices as it has less amylopectin. It is also known as Thai Hom Mali. To harvest jasmine rice, the long stalks are cut and threshed. The rice can then be left in a hulled form and sold as brown rice or shucked and sold as white rice. Most Southeast Asians prefer the white variety of jasmine rice.

articles is presented and the semantic analysis for building ontology is discussed.

The remainders of the paper are as follows. Section 2 describes the motivation for tagging using ontology. Section 3 describes some related research works. Section 4 is the detail of the proposed methodology. The pre-processing process and tag selection process are explained. Section 5 presents the evaluation and result of the experiment using the proposed auto-tagging methodology and the discussion of the proposed classification process and auto-tagging accuracy. The experience of applying the proposed ontology-based auto-tagging methodology for blog articles and the enhancing of auto-tagging system are presented in Sect. 6 and Sect. 7 is a conclusion with some discussion of the proposed auto-tagging methodology.

2 Motivation for tagging using ontology

Tagging system is implemented on many online forums and social networks. The system supports the framework with different purposes. For example, tags are used to describe sharing resources, attract attention, self-presentation, and opinion expression. There are many web sites that use tagging as a mechanism for resource and content retrieval, for example, Delicious, Flickr, Blogger, Wordpress and Wikipedia. The tags are used as the linkage information to relevant resources. In social tagging system, the users specify tags to the published resources such as to images, news, and articles.

The tags may be organised and managed as the part of folksonomy system and that may be simple terms or ontological terms [4]. They are represented as free-form texts specified by the user or the system. Tagging with ontological terms, the tags are represented by the concepts defined in the taxonomy. Tags are typically short textual labels, which provide an easy way to categorise, search, and browse the information they describe. Tags may be represented by a representation language that enables for querying. Retrieval across some application can be implemented with tag linking.

Figure 1 depicts an example of the tagged content from Wikipedia. Wikipedia defines a tag as a free-text keyword and tagging as an indexing process for assigning tags to resources. In this example, the content is annotated by keywords (see underlined terms): ข้าวหอมมะลิ (rice named in Thai), RTGS (refers to Royal Thai General System), long grain, and rice. Some tags are proper names and some tags may associate with the other for example, long-grain rice is a particular kind of rice.

With regard to tag, there are many kinds of tags such as content-based tag, context-based tag, attribute tag, ownership tag and purpose tag [5]. The tagging system provides a particular kind. For content-based tags, the suggested tags may be significant terms because of their term frequency. For example, information in Fig. 1 is suggested with the tag *Rice* because the term rice has maximum frequency. Regarding semantic similarity, multiple tags may have the same meaning or may refer to the same thing. For example, *Kao Horm Mali* is jasmine rice in Thai with English spelling. The users must rely on their own intuition to pick the appropriate tags when multiple tags represent the same meaning. In this paper, the auto-tagging methodology is proposed and this concerns both term frequency and semantic similarity.

3 Related work

Auto-tagging is implemented in many research works. An automatic in-text keyword tagging is proposed by [1]. The tagging system selects candidate keywords from the keyword dictionary by comparing the input document with all terms in the provided dictionary. A tool to suggest tags for weblog is introduced in [6]. The tool finds the similar tagged posts and suggests some set of the associated tags to a user for selection. The research work of [7] follows [6] to provide automatic tag suggestions for the blog post but they focused on performance of tag suggestion system. The system that provides tag recommendation for tagging picture is addressed in [8]. The system recommended the tags for the posted pictures with Flickr web site. The recommended tags are those from similar tagged pictures. The users are able to add one or more tags in the system. Regarding ontology-based tagging approach, some research works proposed the

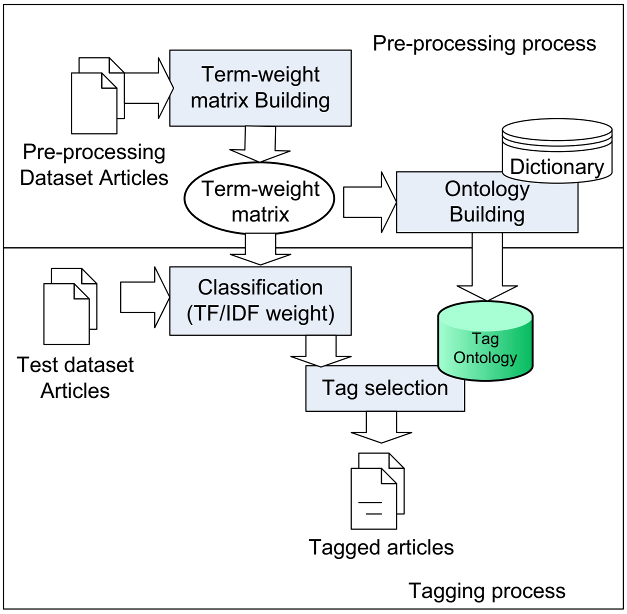


Fig. 2 The process of auto-tagging

formalisation of the tag description. For example, [9] discussed the approach for collaborative tagging at a semantic level. The tags are described by some metadata languages and this enables collaboration across the tagging systems and [4] proposed a formal representation model for tagging and this is represented by OWL. Regarding tag suggestion technique, [10] proposed a voting model where each feature in the resources votes for their favourite tags, and [11] proposed content-based similarity metrics for tagging. From the mentioned works, various approaches and techniques are applied in auto-tagging system using some data and description as aids for the tagging process.

This paper proposes a novel methodology for auto-tagging using ontological tags. The tagging system relies on both IR concept and ontology. Tag suggestion using semantic similarity is presented. The ontological tags are given to the article. The given tags represent semantic information that is acquired from the articles.

4 The proposed ontology-based auto-tagging methodology

Figure 2 depicts the proposed auto-tagging methodology. It includes two main processes: pre-processing process and tagging process. The details are as follows.

(i) The pre-processing process is the process for preparing data. The data are used in classification and tagging process. It consists of the term-weight matrix building. The term-weight matrix contains TF-IDF weight of terms in relevant domains. The obtained term-weight matrix will be used for article classification to find its relevant category/domain. Pre-processing process also includes tag ontology building. Some research works used dictionary and tagged contents to support tagging. The words in dictionary and the tags of the tagged contents are suggested to the article by some analysis. Both techniques rely on the quality of the dictionary or the tagged content. In this paper, ontology is required, however, there is no standard ontology and thus the ontology is provided particularly. In this paper, the ontology is provided manually and the ontology is created from the extracted terms of the train data specified in the pre-processing process.

(ii) The tagging process consists of two steps: article classification and tag selection. Article classification has the main objective to classify the article into relevant domain while tagging process focuses on tag suggestion. The term-weight matrix is used for the classification in tagging process. The article is assigned into a particular domain and it is tagged with the tag ontology of the domain. In related works, there is no obvious work that proposed classification as a step for tagging. Most of researched works assume that the tagging articles are in the relevant domain. In this paper, classification is used as a filtering process to assign the article into the relevant domain and later the ontology of the domain is retrieved for tag suggestion. Thus, classification has no effect to tag suggestion, but it makes tagging process more refined because the tag terms are specified into more specific domain. The article is assigned to the domain by cosine similarity computing. Later, the tagging process uses ontology of the assigned domain for tag selection. The tag selection process computes ontology weight for tag suggestion.

4.1 Pre-processing process

To prepare the data for tagging in runtime, the train dataset are used for the term-weight matrix building. The pre-processing process has three steps as follows.

(i) We extracted the text of the train dataset in part of title, abstract, and content. Lexitron dictionary [12] is adopted for use in this step. The train dataset articles are specified tags manually.

(ii) We built the term-weight matrix. The term-weight matrix contains the TF-IDF weight of the terms for the domains. The TF/IDF weight of the extracted terms from previous step is computed by:

$$\text{TF\_IDFweight}_{i,j} = tf_i \times \text{IDF}_i \tag{1}$$

where $\text{TF\_IDFweight}_{i,j}$ is a TF-IDF weight of the term i in the domain j, tf_i is term frequency of the term i in

Table 1 Term, term frequency, TF-IDF weight of terms in relevant domains

Terms	TF1	TF2	TF3	TF4	*df*$_i$	*D/df*$_i$	IDF$_i$	*W*1	*W*2	*W*3	*W*4
Wheel	0	0	0	208	1	4	0.602	0	0	0	125.229
Geer	0	0	0	160	1	4	0.602	0	0	0	96.330
Car	0	3	0	174	2	2	0.301	0	0.903	0	52.379
Machine	0	0	0	190	1	4	0.602	0	0	0	114.391
Oil	4	0	0	315	2	2	0.301	1.204	0	0	94.824
CVT	0	0	0	71	1	4	0.602	0	0	0	42.746
Speed	0	0	0	81	1	4	0.602	0	0	0	48.767
Break	0	0	0	59	1	4	0.602	0	0	0	35.522
Mitsubishi	0	0	0	57	1	4	0.602	0	0	0	34.317
Steering wheel	0	0	0	53	1	4	0.602	0	0	0	31.909
Mirage	0	0	0	51	1	4	0.602	0	0	0	30.705
Honda	0	0	0	49	1	4	0.602	0	0	0	29.501
Rate	3	0	0	41	2	2	0.301	0.903	0	0	12.342
Save	2	2	0	43	3	1.3	0.125	0.249	0.249	0	5.372
Nissan	0	0	0	45	1	4	0.602	0	0	0	27.093
Power	0	0	0	36	1	4	0.602	0	0	0	21.674
Gas	0	0	0	36	1	4	0.602	0	0	0	21.674
Air	0	0	0	35	1	4	0.602	0	0	0	21.070

the articles of the domain j, $\text{IDF}_i = \text{Log}\frac{D}{df_i}$ by which D is the number of the domain of the train dataset, and df_i is the number of the domains that have the term i. Table 1 depicts an example of term, term frequency, and TF-IDF weight of terms in relevant domains: food, tourism, sport and car; indicated by 1, 2, 3 and 4, respectively. Each term is analysed for its TF/IDF weight on each domain. For example, the term *Oil* is a significant term in domain car because of its highest score.

(iii) We built the tag ontology with terms from the term-weight matrix. The ontology can be enhanced by adding concepts from the domain dictionary. The tags are organised into a hierarchy by considering on generalised and specialised relation. Figure 3 depicts some concepts defined in the tag ontology of food domain. This ontology represents a semantic relation of the concepts regarding broader and narrower meaning of them.

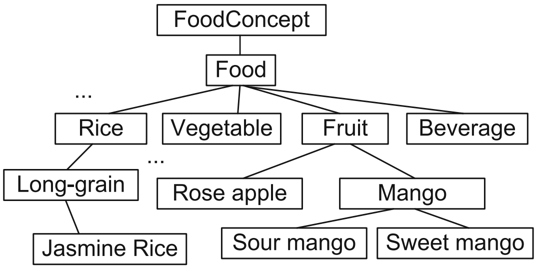


Fig. 3 Tag ontology of food domain

4.2 Auto-tagging process

In this paper, the auto-tagging has two processes: classification and tag selection with following detail.

- *Article classification* The article is classified into relevant domain using cosine similarity. The system compares the article with the train dataset articles. The article is assigned to the domain of the train article that has maximum cosine similarity. The cosine similarity function is computed by:

$$\text{Similarity}(A, D) = \frac{\sum_{i=1}^{n} w_{A_i} w_{D_i}}{\sqrt{\sum_{i=1}^{n} w_{A_i}^2 \times \sum_{i=1}^{n} w_{D_i}^2}} \tag{2}$$

where A is the tagging article, D is the article in the train dataset, w_{A_i} is TF/IDF weight of term i in article A, w_{D_i} is TD/IDF weight of term i in article D.

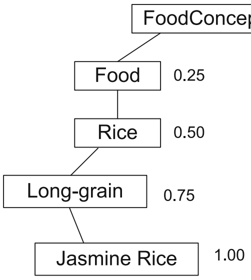
- *Tag selection* Tag selection has two steps as follows.

(i) The extracted terms of the article are matched with concepts defined in tag ontology of relevant domain. The matched terms are considered for tag suggestion in the next step.

(ii) Ontology weight is computed to specify tag's significance. The tags are ranked and suggested by their significance.

Fig. 4 Ontology weight of ontological tags

Jasmine rice (Thai: ข้าวหอมมะลิ; RTGS: Khao Hom Mali; Thai pronunciation: [kʰâːw hɔ̌ːm malíʔ]), sometimes known as *Thai fragrant rice*, is a long-grain variety of rice…
Ontology Weight: *Jasmine Rice, Long-grain, Rice*
Ontology Weight TF: *Rice, Jasmine Rice, Long-grain*



In this paper, the ontology weight is computed with term frequency and without term frequency. We follow edge-based method [13] and propose ontology weight computing as follows.

$$\text{OntoWeight}_{t,d} = \frac{N_t}{D_{N_t}} \tag{3}$$

$$\text{OntoWeightTF}_{t,i} = \text{OntoWeight}_{t,i} \times \text{TF}_t \tag{4}$$

where N_t is the number of edges from root to tag t, D_{N_t} is the number of edges from root to the descendant node (the leaf node) of tag t, and TF_t is the number of term frequency of tag t in the tagging article.

Figure 4 depicts the suggested tags of the content from Fig. 1. With the ontology weight score without TF, the suggested tags are *Jasmine Rice* (weight score = 1.00), *Long grain* (0.75), and *Rice* (0.50), respectively. In contrast, the suggested tags are *Rice* (weight score = 5.00, TF = 10), *Jasmine Rice* (weight score = 4.00, TF = 4) and *Long grain* (weight score = 0.75, TF = 1), respectively.

5 The evaluation

In this paper, we conduct two kinds of evaluation with two different purposes:

(i) To check whether the auto-tagging suggests tags that include manual-tags or not.
(ii) To compare auto-tagging accuracy with manual-tagging accuracy. The recall and precision are computed for both.

Figure 5 depicts an example of the manual-tags (left) and suggested tags from auto-tagging system (right). In this example, the suggested tags from auto-tagging system include the manual-tags. With the purpose (i), this shows that the auto-tagging system suggests tags that include the manual-tags.

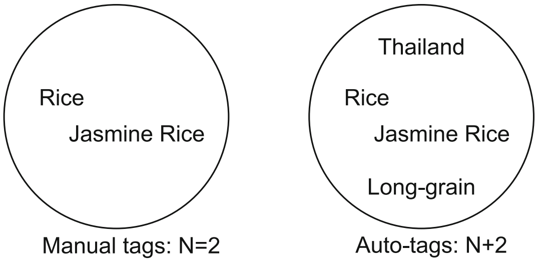


Fig. 5 The set of manual-tags (*left*) and auto-tags (*right*)

In this experiment, 140 articles are used for this evaluation. The articles are in Thai language. There are 70 articles in the train dataset and 140 articles for the test dataset (the formers are included). Although, the test data are included in this experiment but the tagging evaluation based on accuracy is not affected by classification process. Both datasets are articles collected from vcharkarn.com web site. The articles are in Thai and categorised according to the mentioned domains.

Table 2 depicts the results of the evaluation for the purpose (i). The suggested tags from auto-tagging are compared with the manual-tags. For example, if the article has N manual-tags, the length of tag suggestions in auto-tagging: $N+1$, $N+2$, $N+3$, $N+4$ and $N+5$, are evaluated. With the proposed ontology weight computing, most tags with specific meaning are tagged before the tags with generic meaning. The auto-tagging provides tags that include the manual-tags when the length of tag suggestion is increased. From Table 2, the length $N+5$ tag suggestion produces high accuracy, whereas the shorter length of tag suggestion has low accuracy.

Table 3 shows the result of the evaluation regarding the recall and the precision. In this paper, the auto-tags and manual-tags are evaluated with different lengths of tag suggestions. The experiment is implemented by querying articles

Table 2 Evaluation result of purpose (i)

Tagging methods	N	$N+1$	$N+2$	$N+3$	$N+4$	$N+5$
Ontology	37.10	52.90	64.30	71.40	80.00	85.70
Ontology $\times tf$	67.10	81.40	84.00	87.10	88.60	94.30

Table 3 Evaluation result of purpose (ii)

Domain	Manual-tagging		Auto-tagging $N=4$		Auto-tagging $N=6$		Auto-tagging $N=8$		Auto-tagging $N=10$	
	R	P	R	P	R	P	R	P	R	P
Car	0.93	0.79	0.85	0.58	0.99	0.60	0.99	0.66	1.00	0.65
Sport	0.98	0.95	0.97	0.92	0.99	0.92	0.99	0.93	1.00	0.93
Food	0.89	0.95	0.95	0.87	0.97	0.86	0.99	0.86	1.00	0.86
Tourism	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Average	0.95	0.92	0.94	0.84	0.99	0.85	0.99	0.86	1.00	0.86

Note that R is recall value and P is precision value

using a set of 10 keywords for each particular domain. The system retrieves the articles using such keywords. The accuracy is evaluated by ontology weight computing with term frequency. The average recall and precision of this experiment are 0.98 and 0.85, respectively. Also, the accuracy of the classification is 90 %. The proposed auto-tagging methodology returns high recall but precision may be lowered accordingly when the length of tag suggestion is increased. However, tagging is expected a better recall rather than precision.

With the proposed auto-tagging methodology, the classification process supports retrieving specific information but tagging is focused on how to choose appropriate tags for the article. In this paper, classification is high because the test dataset (140 articles) is comprised of the train dataset (70 articles) that are used for classification. However, tag selection is not affected by such train dataset. Because tagging is implemented by semantic analysis of the article's content by which TF-IDF weight and ontology weight are focused.

6 An experience of ontology-based auto-tagging with blog articles

6.1 Ontology building

In previous sections, the experiment uses the provided ontologies for the rough four article domains and such are obtained from a small dataset. Here, an extensive experiment is conducted by focusing on ontology building with blog articles. A collection of 308 blog articles is collected from http://www.travelfish.org/blogs/thailand for this experiment. Most articles in Travelfish are classified into associate place (i.e. province) and associate sub-categories. For example, an article is classified into category *Bangkok* with six sub-categories: accommodation, sightseeing and activity, art and culture, transport, bar and nightlife, and event and festivals. However, some categories have no sub-categories due to a small number of the articles. With no standard ontology available, the ontology is created particularly for this experiment. Here, the Autotags tool (v.1.3) [14] is used. The Autotags is a tool for tag generation. It provides semantic analysis based on term frequency. In this paper, the Autotags is an aid for keyword extraction from articles. The tool generates some tags according to some weight scores that are rated based on some characteristics of terms such as capitalised terms, white space term. The suggested tags can be simple term and complex term (i.e. term with white space). From this experiment, 10 suggested tags are obtained for each article. However, Autotags may generate some misuse tags by which those are slang, author' speech opinion, and Thai word (pronounced in English). Thus, these tags are removed manually. In this experiment, the obtained tags (keywords) are analysed their relevancy according to six sub-categories mentioned above.

Table 4 is an example of semantic analysis focusing on the relationship between the term and domains. For example, *Wat Phra Kaew* can be recommended as a point of interest and a historic landmark for sightseeing and activity, and art and culture domain, respectively; *Boat* may associate to transportation by boat for transport domain and it may represent a particular museum—boat museum (e.g., Thai boat museum) for art and culture domain; *River* can be recommended as a point of interest for sightseeing and activity domain and water transportation for transport domain; and *Museum* may represent a point of interest in art and culture domain.

Ontology building needs the knowledge and view in regard to the phenomena of the domain. The term can be derived

Table 4 Semantic analysis of relationship between term and domains

Terms	Sightseeing and activity	Even and festival	Transport	Accommodation	Bar and nightlife	Art and culture
Wat Phra Kaew	Point of interest	–	–	–	–	Historic landmark
River	Point of interest	–	Water transportation	–	–	–
Museum	–	–	–	–	–	Point of interest
Boat	–	–	Transportation by boat	–	–	Point of interest
Songkran	–	Festival	–	–	–	–
Bungalow	–	–	–	A kind of accommodation	–	–
Live show	–	–	–	–	Event	–

from the generic term into the specific term. For example, transportation can be derived into water transportation, which can be derived further into boat transportation, cruise transportation, and ferry transportation. Figure 6 depicts some concepts defined in tourism ontology. The concepts are determined into particular sub-domains (e.g., sightseeing and activity, and art and culture). For art and culture domain, the point of interest can be derived into museum, historic landmark and religious worship. For sightseeing and activity domain, the point of interest can be islands and park, but shopping can be defined as the activity of the domain. In this experiment, 1,688 terms are obtained from Autotags and the ontological tag 1,459 tags are derived from the former.

Building ontology can be implemented with three approaches: bottom-up approach, top-down approach and combination approach [15]. In bottom-up approach, the information is derived from the instance or the specific term to the generic term. In this experiment, the combination approach is the suitable methodology. Figure 6 depicts an example of deriving concepts by considering on *is-a* relationship using the obtained information from semantic analysis (see Table 4). In addition, each concept can be defined with equivalent property for example, the concept *Temple* is the equivalent concept of *Wat* (means temple in Thai language), and *Phu-Khao* (means mountain in Thai) is the equivalent concept of *Mountain* and in vice versa.

Building ontology requires experience and skill of the ontology engineer to analyse semantics of terms and relations between them and the domains. This process is usually implemented manually and may use some knowledge-base and dictionary as aids for the analysis. It is difficult to judge if ontology is a well-built ontology even it is created by the ontology engineer who has particular expertise. However, the ontology can be evaluated after used and can be improved to support the processing of application. Building ontology means the creating of concept, instance, and relations between them [16]. In this paper, concept and instance creation are concentrated particularly by which the concept may have is-a relation (specified by rdfs:subClassOf) with another concept. Figure 7 shows an example of some concepts of tourism domain defined with Protégé [17].

To maintain information, the ontology can be described by a language that is available for the system to query. Here, the ontology is presented by OWL [18]. OWL is a standard language for ontology creation proposed by W3C. Figure 8 depicts an example of the instance description that describes the historic landmark *Wat Phra Kaew*. The instance is the member of class *Historic_Landmark* and it associates to two keywords: *Wat Phra Kaew* and *Wat Phra Si Rattana Satsadaram*; the former is the short well-known name and the latter is the official name. These keywords are used for matching in the tagging process. The use of the OWL-based ontology profile is explained in Sect. 6.3.

6.2 Classification using ontological information

In Sect. 4, tagging is based on classification using supervised information (i.e. the train dataset) with a small dataset. From the experiment, cosine similarity computing may take time when there are many train data. Moreover, the selection of the train data is a critical task. Thus, preparing the train data may need another efficient technique such as support vector machine to determine the classifiers for the particular domain and this technique is appropriate for a large dataset.

In this paper, an extensive experiment applying for blog articles is implemented. The unsupervised approach for the classification process is focused and the classification process relies on the built ontology. The article is classified into relevant domain. Here, the relevancy of an article to a domain is represented by the number of the matched terms. For example, the short message "From Pattaya, it is a little over an hour to the Bangkok Airport. You could catch an early flight to KL operated by Thai Airway or other low cost airline to Malaysian city and return in the evening" is matched with

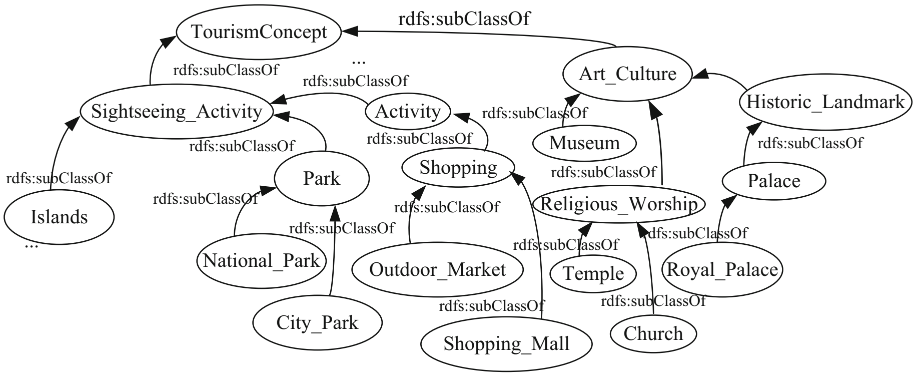


Fig. 6 Some concepts tourism ontology

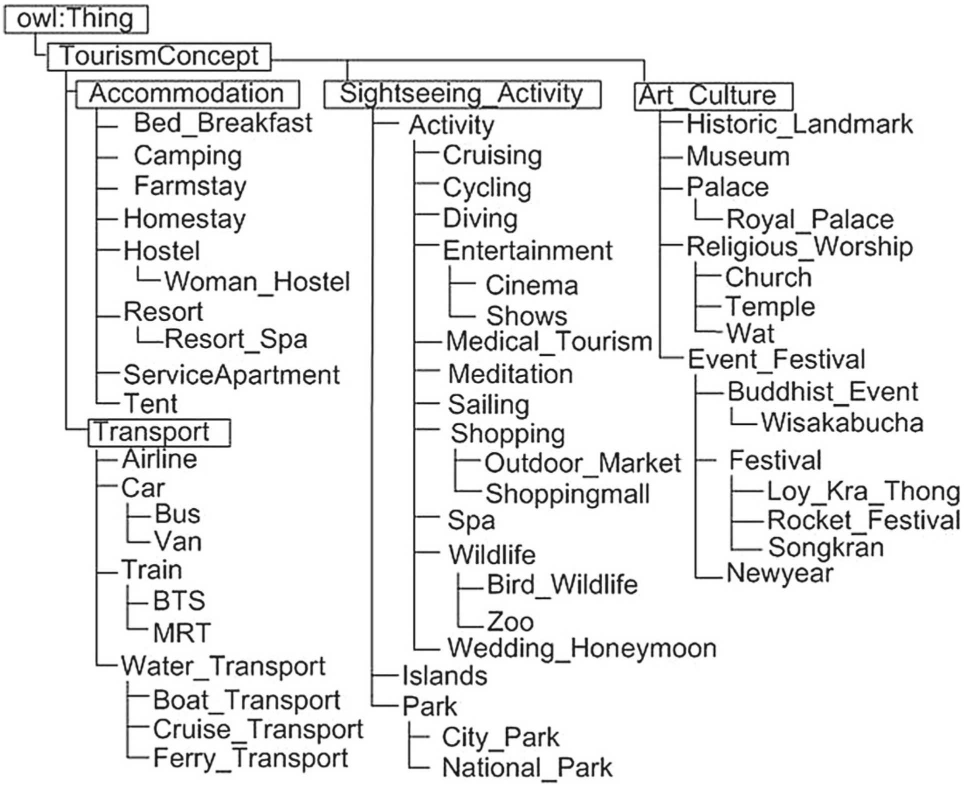


Fig. 7 Some concepts defined in tourism domain

Fig. 8 An example of OWL description of some concepts of tourism ontology

```
<owl:Class rdf:ID="Art_Culture">
    <rdfs:subClassOf rdf:resource="#TourismConcept"/>
</owl:Class>
<owl:Class rdf:ID="Religious_Worship">
    <rdfs:subClassOf rdf:resource="#Art_Culture"/>
</owl:Class>
<owl:Class rdf:ID="Temple">
    <rdfs:subClassOf rdf:resource="#Religious_Worship"/>
    <owl:equivalentClass rdf:resource="#Wat"/>
</owl:Class>
<!—Instance Description -->
<Historic_Landmark rdf:ID="Historic_Landmark_1">
    <hasKeyword rdf:datatype="&xsd;string">
            Wat Phra Kaew</hasKeyword>
    <hasKeyword rdf:datatype="&xsd;string">
            Wat Phra Si Rattana Satsadaram</hasKeyword>
</Historic_Landmark>
```

Table 5 The results of extensive experiment

Domains	No. of articles (i)	No. of classified articles (ii)	Accuracy (iii)	No. of classified articles[a] (iv)	Tags (v)	Onto-tags (vi)
Sightseeing and activity	87	84	0.96	41	669	578
Art and culture	22	20	0.91	5	228	206
Event and festival	24	23	0.96	0	192	156
Accommodation	13	12	0.92	15	110	87
Transport	32	31	0.97	5	234	213
Bars and nightlife	35	35	1.00	4	255	219
Total	213	205	0.95	70	1,688	1,459

[a] The articles without pre-defined relevant category

three terms: *Bangkok Airport*, *Thai Airway*, and *low cost airline* of transport domain. Note that term matching is implemented with insensitive case matching by which the *N*-gram matching can be considered to enhance the precision of the matching terms. Here, the article can be assigned into one or more domains if the numbers of the matched terms of those domains are equivalent. The article may have no relevant domain if there is no the matched terms for all domains. In this experiment, an article is classified into six domains with brief descriptions as follows.

- *Sightseeing and activity* domain includes the articles that provide information regarding recommended place to visit and some other activities such as shopping, cycling, journey, and park.
- *Art and culture* domain includes the places that are historic landmark, religious worship, museum and including language learning.
- *Event and festival* domain includes the political events, national festivals and religious festivals.
- *Accommodation* domain includes the blog articles that outline about recommended resorts or hotels and accommodation guidance.
- *Transport* domain includes the articles that give some information of travelling to some places and the trip planning.
- *Bar and nightlife* domain includes the articles recommended nightlife, restaurants, bars or clubs and live show.

Table 5 shows the result of this extensive experiment. The 308 blog articles (English articles) are collected from TravelFish website by which 213 articles have their relevant categories (classified by Travelfish.org) with the number of articles shown in column (i). Column (ii) shows the number of articles that are classified into particular domain. Classification accuracy is shown in column (iii). There are 95 articles that have no relevant categories (six domains) and these are classified using the ontology (Sect. 6.2). From this experiment, 25 articles are not matched with any domains and 70 articles (iv) are classified into relevant domains by which 61 articles associate with multiple domains and the rest 9 articles associate with single domain. The number of keywords obtained from Autotags is in column (v) and the number of the derived ontological tags is in column (vi) and this is analysed by semantic analysis (Sect. 6.1). Most of ontological tags are named entities of places in Thailand.

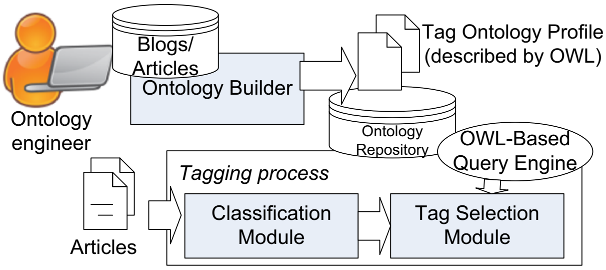


Fig. 9 The components of the enhanced ontology-based auto-tagging system

6.3 The enhancing of auto-tagging system

This section gives the detail of the enhancing of the ontology-based auto-tagging system (see Fig. 9). With the OWL-based ontology profile, information maintenance can be managed in the system. For example, the new concepts/topics can be introduced in the system. The system may maintain the amount of the articles for ontology building. The article may be a set of the articles the system providing. Ontology engineer interacts with the ontology builder tool. The ontology builder provides text extraction (e.g., using Autotags) for semantic analysis. The ontology engineer specifies the term for ontology creating. The ontology builder generates the OWL-based tag ontology profile. The profile is available for query.

With the proposed ontology weight computing (Sect. 4.2), it is possible to query depth of the concepts using the SPARQL query [19]. There are some OWL-based query engines available such as RAP API [20] and Jena [21]. The classification module can be implemented using ontology as the basis information (Sect. 6.1). The tag selection module computes ontology weight with and without term frequency (Sect. 4.2).

7 Conclusion

This paper proposed ontology-based auto-tagging methodology using semantic approach. The auto-tagging consists of classification process and tag selection process by which the former is a step for filtering the articles into relevant domains. The classification process is evaluated with supervised and unsupervised approach. With supervised approach, the cosine similarity is implemented and for the large set of the articles the unsupervised approach is more suitable. The technique of ontology building is presented in this paper. It is quite obvious that the lightweight ontology is appropriate for the application. The results from the experiment with blog articles show that the classification process using ontology can be implemented with the ease computing, but produces the effective results.

With ontology-based tagging, the suggested tags are ranked according to semantic analysis and this concerns not only term frequency but also similarity measured by ontology. Using ontology, the suggested tags are meaningful tags and these also present semantics of the article.

References

1. Kim, J., Jin, D., Kim, K., Choe, H.: Automatic in-text keyword tagging based on Information retrieval. J. Inf. Process. Syst. **5**(3), 159–166 (2009)
2. Gruber, T.: A translation approach to portable ontology specifications. Knowl. Acquis. **5**(2), 199–220 (1993)
3. Rattanapanich, R., Sriharee, G.: Auto-tagging articles using latent semantic indexing and ontology. In: Proceedings of the 6th Asian Conference on Intelligent Information and Database Systems ACIIDS, pp. 153–162 (2014)
4. Knerr, T.: Tagging ontology—towards a common ontology for Folksonomies. https://tagont.googlecode.com/files/TagOntPaper.pdf (2013). Retrieved 4 Nov 2013
5. Gupta, M, Li, R., Yin, A., Han, J.: Survey on social tagging techniques, ACM SIGKDD Explorations Newsletter, vol. 12, Issue 1, ACM New York, USA, pp. 58–72 (2010)
6. Mishne, G.: AutoTag: a collaborative approach to automated tag assignment for web log posts. In: The 15th International World Wide Web Conference 2006. Edinburgh, Scotland (2006)
7. Sood, S.C., Owsley, S.H., Hammond, K.J., Birnbaum, L.: TagAssist: automatic tag suggestion for Blog Posts. In: International Conference on Weblogs and Social Media, Boulder, Colorado, USA, March, pp. 26–28 (2007)
8. Garg, N., Weber, I.: Personalized, interactive tag recommendation for Flickr. In: The 8th ACM Recommender Systems Conference. Lausanne, Switzerland (2008)
9. Kim, H.L., Scerri, S., Breslin, J.G., Decker, S.: The state of the art in tag ontologies: a semantic model for tagging and folksonomies. In: Proc. Int' l Conf. on Dublin Core and Metadata Applications (2008)
10. Si, X., Liu, Z., Li, P., Jiang, Q., Sun, M.: Content-based and graph-based tag suggestion. In: Proceedings of ECML PKDD (The European Conference on Machine Learning and Princi ples and Practice of Knowledge Discovery in Databases) Discovery Challenge 2009, Bled, Slovenia, September 7 (2009)
11. Byde, A., Wan, H., Cayzer, S.: Personalized tag recommendations via tagging and content-based similarity metrics. In: International Conference on Weblogs and Social Media, Boulder, Colorado, USA, March, pp. 26–28 (2007)
12. LEXITRON. http://lexitron.nectec.or.th/. Accessed 17 Nov 2014
13. Wu, Z., Palmer, M.: Verb semantics and lexical selection. In: Proceedings of the 32nd annual meeting of the associations for computational linguistics (1994)
14. Autotags. http://mrolafsson.github.io/autotags/. Accessed 17 Nov 2014
15. Gómez-Pérez, A., Fernandez-Lopez, M., Corcho, O.: Ontological engineering: with examples from the areas of knowledge management, e-commerce and the semantic web. In: Advanced information and knowledge processing, 1st edn. Springer, Berlin (2010)
16. Noy, N.F., McGuinness, D.L.: Ontology development 101: a guide to creating your first ontology. Stanford Knowledge Systems Laboratory Technical Report, March (2001)
17. The Protégé Ontology Editor and Knowledge Acquisition System. http://protege.stanford.edu/
18. McGuinness, D.L., Harmelen, F. V.: OWL web ontology language overview. http://www.w3.org/TR/owl-features (2004). Accessed 17 Nov 2014
19. SPARQL query language for RDF. http://www.w3.org/TR/rdf-sparql-query/. Accessed 17 Nov 2014
20. RAP—RDF API for PHP V0.9.6. http://sourceforge.net/projects/rdfapi-php/. Accessed 17 Nov 2014
21. Jena a semantic web framework for Java. http://jena.sourceforge.net/. Accessed 17 Nov 2014