Advances in Economics, Business and Management Research (AEBMR), volume 26
International Conference on Economics, Finance and Statistics (ICEFS 2017)

A Comparison of Statistical and Data Mining Techniques for Enrichment Ontology with Instances

Aurawan Imsombut[1,*] and Jesada Kajornrit[1]

[1] College of Innovative Technology and Engineering, Dhurakij Pundit University, Bangkok, Thailand

*aurawan.ims@dpu.ac.th

Abstract. Enriching instances into an ontology is an important task because the process extends knowledge in ontology to cover more extensively the domain of interest, so that greater benefits can be obtained. There are many techniques to classify instances of concepts, however, two popular ones are the statistic and data mining methods. This paper compares the use of these two methods to classify instances to enrich ontology having greater domain knowledge. This paper selects conditional random field for the statistics method and feature-weight k-nearest neighbor classification for the data mining method. The experiments were conducted on the tourism ontology. The results showed that conditional random fields methods provided greater precision and recall value than the other, specifically, F1-measure is 74.09% for conditional random fields and 60.04% for feature-weight k-nearest neighbor classification.

Keywords: Ontology Enrichment; Statistical Technique; Classification; Conditional Random Fields (CRFs); Feature-weighted k-Nearest Neighbor.

1. Introduction

Ontology consists of concepts in a domain-of-interest, such as tourism, medicine, and agriculture. In an ontology, the concepts are interconnected by semantic relations. The ontology can be implemented in various domains which are referred to systems and subs-systems that require in-depth meaning of the information, for example, information retrieval and recommendation systems. Furthermore, ontology learning consists of different tasks. They are term extraction and normalization synonym identification, concept and instance recognition and relation extraction [1]. The identifying instance is an important task for the ontology learning to expand knowledge in the ontology for implementing the ontology in various domains. However, the ontology instance extraction consumes both computational time and expert efforts. Therefore, automatic or semi-automatic ontology instance extraction is needed and should be investigated.

This paper focuses on instances of concepts relating to Attractions. Each concept is classified into sup-concepts. For example, the attraction concept consists of Cultural, Argo, Natural, and Shopping sub-concepts. Such information is mostly searched by users, and has been used for decision making. Basically, a word representing, for instance, for each concept in the ontology is a specific name called Name Entity (NE). This NE is a proposition used to identify things such as persons, organizations or locations [2]. However, NE in the Thai language does not have orthographical information: for example, the capital letters at the beginning of the sentence as used in the English language, or special characters such as Kanji, Katakana as used in the Japanese language. Then, there is a challenging task to extract NE in Thai language.

There are many techniques to extract instance of concepts (i.e. NE), however, two popular ones are the statistics and data mining methods (classification). This paper compares these two techniques to

---

The authors wish to thank Michael McAleer for helpful comments and suggestions.
This research was supported by Dhurakij Pundit Research Service Center, Dhurakij Pundit University.

classify instances, that is, Conditional Random Fields (CRFs) for statistics methods and feature-weight k-Nearest Neighbor (KNN) classification of data mining methods for extracting ontology instances [2], [3]. The CRF technique is recommended for recognizing classes for the sequence data, especially, the natural language processing (sequence of words).

On the other hand, KNN, one of many classification techniques in data mining methods, is selected in this paper because the features of data normally are nominal and boolean data types. This feature contains words that usually stay around the interested words. Thus, techniques such as Artificial Neural Network (ANN) or Support Vector Machine (SVM) cannot be applied. Moreover, the data used in this paper are unbalanced data. If the traditional technique such as kNN is used to classify, the problems related to majority class bias occur. Therefore, feature-weighted kNN is proposed for improving the performance of unbalanced data categorization problems, so that the feature-weighted kNN can improve the classification performance.

The remainder of the paper is organized as follows: Section 2 reviews related works, Section 3 presents a brief review of the methods used, the data and experiments are presented in Section 4, Section 5 presents the results and discussion, and Section 6 gives some concluding comments.

2. Related Works

There are many research studies concerning ontology enrichment (i.e., define and classify instances). Most of those studies apply NLP techniques with Information Extraction (IE) techniques and Machine Learning (ML) techniques.

Martinez et al. [5] proposed a combination of NLP and IE techniques by using GATE tools for extracting NE from restaurant and hotel corpus, and use heuristic algorithm for solving different kinds of ambiguities to populate the instances into tourism ontology. Faria et al. [6] presented another combination of NLP and IE to create rules for automatic population of ontologies from text. Their study was conducted on legal and tourism corpora.

Zhang et al. [7] applied NLP and ML techniques called Maximum Entropy to extract relationships between entities for the field of tourism. Nanba et al. [8] applied NLP and used CRF as ML in order to identify travel blog, and extract travel information relating to the relationships between location names and local products. Carlson et al. [9], Giuliano and Gliozo [10], Cimiano et al. [11] and Etizioni et al. [12] applied the NLP, IE and ML techniques to the ontology population.

3. Brief Review of Used Methods

In this section, we will briefly review the literature of statistical techniques: Conditional random fields (CRFs) and data mining technique: feature-weight k-Nearest Neighbor classification

3.1 Statistical Technique

Conditional random fields (CRFs) is a statistical technique that is usually used for pattern recognition, especially in the natural language processing area. CRFs [12] are undirected graphical models that are often used to predict sequences of labels for sequences of input samples, such as natural language text. When applying CRFs to the named entity recognition problem, an observation sequence is the token sequence in the document, and state sequence is its corresponding label sequence.

The conditional probability of a state sequence $s=<s_1, s_2, ..., s_T>$, given an observation sequence $o=<o_1, o_2, ..., o_T>$, is defined as:

$$P(s|o) = \frac{1}{Z_o} exp(\sum_{t=1}^{T} \sum_{k=1}^{K} \lambda_k f_k(s_{t-1}, s_t, o, t) , \quad (1)$$

where $f_k(s_{t-1}, s_t, o, t)$ is a feature function and is a learned weight for each feature function. Z_o is a normalization factor over all state sequences, and is defined as:

$$Z_o = \sum_s exp\, (\sum_{t=1}^{T} \sum_{k=1}^{K} \lambda_k f_k(s_{t-1}, s_t, o, t) \ . \quad (2)$$

3.2 Data Mining Technique

There are several classification techniques. Nevertheless, kNN is chosen to use in this research because the data type of input features are hybrid, which are nominal and boolean. The other classification techniques, such as a decision tree, is appropriate for nominal data type, whereas SVM and neural networks are suitable for numeric data types.

kNN is a simple classification technique to determine the class. It finds K-nearest neighbors from supervised learning data. Then it chooses the class from maximum score according to (3), where c_i denotes to class i, referred to correct class of $Sim(e, e_j)$. It represents the similarity of sample e which is testing data, and e_j is the sample of supervised learning data with K-nearest neighbor characteristics. It calculates the similarity in all feature k in the sample from k=1 to k=n, and $\delta(e_j, c_i)$ =1, if e_j contains class i, otherwise it is set to zero:

$$score(e, c_i) = \sum_{e_j \in KNN(e)} Sim(e, e_j)\delta(e_j, c_i) \tag{3}$$

$$Sim(e, e_j) = \sqrt{\sum_{k=1}^{n}(|e_k - e_{jk}|)} \tag{4}$$

$$\delta(e_j, c_i) = \begin{cases} 1 & e_j \in c_i \\ 0 & e_j \notin c_i \end{cases} \tag{5}$$

However, there are some limitations of kNN. It tends to classify the data based on the majority class. Therefore, it is not appropriate to classify with unbalanced data. The feature-weighted kNN classifier can mitigate the problem by using the weight of feature. The weight will be determined differently regarding the importance of the classification. The more important features will be weighted higher than the less important features. Thus, this can decrease the overall significance of the classification:

$$Sim(e, e_j) = \sqrt{\sum_{k=1}^{n} w_k(e_k - e_{jk})} \ , \tag{6}$$

where w_k is the weight of feature k and the weight of the feature can be calculated using correlations based on class attributes. It can be seen that the higher weight of the feature gains greater relevance in the considered class. In addition, a correlation lies between -1 and +1. This can also measure the relationship degree between two considered features. A positive value means a positive relationship, whereas a negative value refers to a negative association:

$$Correlation\ coefficient = \frac{\sum(X-\bar{X})(Y-\bar{Y})}{(n-1)\sqrt{S_x^2 S_y^2}} \ , \tag{7}$$

where X and Y have means, $\bar{X}$ and $\bar{Y}$, and standard deviations, S_x and S_y , respectively.

3.3 Benefits and Limitations of Each Method

CRFs are learned from the Corpus. They transform an input text to a feature vector, and then create all possible nodes. Then they select the most possible node for the answer. CRFs technique are able to solve the labeled bias problem because CRFs are discriminative models. The mathematical representation of CRFs is an undirected graphical model, and it evaluates the probability of the next label by using all previous labels that have event sequence as criteria to calculate the weights of features from different states. Thus, the state bias problem is reduced.

However, the limitation of CRF depends on the number of training data. If the number of data is large, the amount of memory used is increased. This limitation causes CRF technique not to suitable for large data. Some techniques such as feature selection is needed to reduce this limitation.

Feature-weighted kNN is a classification technique. The *k*NN performs fast as its uncomplex mechanism, and classifies data by using *k*-closed training data. The feature-weighted step calculates more weights to the features that have more effects than the features that have less effects to the classification. However, if the training data are unbalanced or noisy, the classification error can increase.

4. Data and Experiments

The data source used for the experiments was obtained from Thai tourism websites. One hundred randomly selected webpages were used to create a dataset as training data. The ontology instance extraction process is composed of three sequential phrases, as follows: Pre-processed, Feature Extraction, and Instance Extraction, as will be explained below. Finally, the data which have identified the domain of NE and its type is derived. 10-fold cross validation will be used to separate the training and testing data.

Pre-processed is the step to remove HTML tags with HTML parser from the documents. Then the documents are fed into Natural Language Processing (i.e. word segmentation and part of speech tagging) by using developed own tools. Word segmentation uses longest matching and defines POS with Hidden Markov Model (HMM).

Feature Extraction is the step to extract important features that are used by the system to learn a classification boundary, and identify types of noun-identified propositions. The characteristics are as follows:

Lexical & POS features, consist of:

- Words and POS of the current word
- Words and POS of 3 words before the current word
- Words and POS of 3 words after the current word

Dictionary features, consist of:

- Is current word in the cue word list? (e.g. Temple, Park)
- Are previous n-words before the current word in the cue word list? (e.g. Temple, Park)
- Are the words not in the dictionary?
- Do the words appear in a location dictionary?

Repeated occurrence:

- Do the words occurring before and after the considering word occur together more than 3 times?

In addition, the value of Lexical&POS features are nominal, but the value of the dictionary features and repeated occurrence are 0 or 1.

Instance Extraction is a step to extract noun-identified propositions. Noun-identified propositions are instances of concepts in ontology. This study identifies the boundary of NE and classifies types of NE by recognition technique CRFs and feature-weight kNN classification, a supervised learning that learns from class-labelled examples. The classified types are Cultural, Argo, Natural, Shopping, and others.

5. Results and Discussion

In the experiments, 100 Thai documents (or 40,000 words approximately) from the Thai tourism websites were used. The contents in the website were, for example, attractions, accommodations, and activities. Those documents were pre-processed and performed instance extraction. To evaluate the performance of the classification of attraction category, F1, which was proposed by van Rijsbergen [14], will be used. It applied precision and recall as follows:

$$Precision = \frac{number\ of\ correct\ positive\ predictions}{number\ of\ positive\ predictions} \times 100 \qquad (8)$$

$$Recall = \frac{number\ of\ correct\ positive\ predictions}{number\ of\ positive\ examples} \times 100 \tag{9}$$

$$F1 = \frac{2 \times Precision \times Recall}{(Precision + Recall)} \tag{10}$$

The results of the extracted instance of ontology concept can be seen from Table 1 below.

Table 1

Experimental results for instance extraction

Attraction class	**Precision**		**Recall**		**F1**	
	CRF	**feature-weight kNN**	**CRF**	**feature-weight kNN**	**CRF**	**feature-weight kNN**
Cultural	**80.17%**	52.59%	74.62%	**67.25%**	77.29%	59.02%
Agro	66.67%	75.81%	43.24%	33.33%	52.46%	46.31%
Natural	79.25%	**78.70%**	**87.50%**	62.63%	**83.17%**	**69.75%**
Shopping	66.67%	76.00%	53.33%	33.93%	59.26%	46.91%
	77.62%	60.84%	70.87%	59.26%	**74.09%**	**60.04%**

The preliminary experiment focused on k-value adjustment for the kNN classification, and was adjusted from k=1 to k=10. The results show that k=8 gained the maximum F1 value. In addition, the features that have maximum weight are features of repeat occurrence, namely the word after the current word and the word before the current word.

The results of the cultural attraction extraction process with the CRFs technique showed the highest precision because most of their names were specific names, such as temple names and monument names. As a result, it was not difficult for the classification module to clarify them.

On the other hand, the feature-weighted kNN provided less accuracy to classify class cultural because key features for classifying are contaminated by some common words. For example, the word “วัด” (in Thai) is a common word, but the classification system evaluates this general word into the cultural group.

In the case of location names in the Natural group, some words that begin with “Mountain”, “Fountain”, “Cave”, or “Hill” usually appear with the location name, and cause featured-weight kNN to classify correctly.

For the F1 value, considering average precision and recall values in every class, one can see that the CRFs technique showed higher F1 value than featured-weight kNN because the CRFs technique can reduce the bias problem of unbalanced data in the experiments.

6. Conclusion

This paper presented a comparison of instance extraction in ontology between the statistical method (Conditional Random Fields) and data mining method (feature-weighted kNN classification). The results showed that the CRFs technique provided greater precision and recall value than the feature-weighted kNN method because the CRFs technique is a more suitable technique to direct class for sequential data, and because CRFs can handle unbalanced data better. The data used in the experiments come from websites, and contain common words more frequently than location names. As a result, CRFs show superior results than feature-weighted kNN. In future work, more machine learning will be investigated, with extended concepts in the experiment.

References

[1] Z. Zhang and F. Ciravegna, 2011. Named Entity Recognition for Ontology Population using Background Knowledge from Wikipedia", in Ontology Learning and Knowledge Discovery Using the Web: Challenges and Recent Advances, IGI Global.

[2] Chinchor, Nancy. 1998. MUC-7 Named Entity Task Definition (Version 3.5). MUC-7. Fairfax, Virginia.

[3] Imsombut, A. and Sirikayon, C. 2016. An Alternative Technique for Populating Thai Tourism Ontology from Texts Based on Machine Learning. Proceeding of 15th IEEE/ACIS International Conference on Computer and Information Science (ICIS 2016).

[4] Imsombut, A. and Paireekreng, W. 2016. Extract Knowledge for Populating Thai Tourism Ontology from Texts Using Feature-weighted k-Nearest Neighbor. Proceeding of 1st International Conference on Information Technology.

[5] Martinez, J. et al. 2011. Ontology Population: An Application for the e-tourism Domain. International Journal of Innovative Computing, Information and Control.

[6] Faria, C., Girardi, R. and Novais, P. 2012. Using Domain Specific Generated Rules for Automatic Ontology Population. Proceeding of 12th International Conference on Intelligent Systems Design and Applications.

[7] Zhang, Y., et al. 2009. Automatic Entity Relation Extraction for the Field of Tourism. Journal of Computational Information System.

[8] Nanba, H., et al. 2009. Automatic Compilation of Travel Information from Automatically Identified Travel Blogs. Joint Conference of the 47th Annual Meeting of the Association for Computational Linguistics and 4th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing (ACL-IJCNLP).

[9] Carlson, A., J. Betteridge, R. Wang, Jr. E. Hruschka and T. Mitchell. 2010. Coupled Semi-Supervised Learning for Information Extraction. In Proceedings of the third ACM International Conference on Web Search and Data Mining (WSDM '10).

[10] Giuliano, C., and Gliozo, A. 2008. Instance-Based Ontology Population Exploiting Named Entity Substitution. In Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008).

[11] Cimiano P., Ladwig, G., and Staab. 2005. Gimme' The Context: Context-driven Automatic Semantic Annotation with C-PANKOW. In Proceedings of the 14th World Wide Web Conference (WWW).

[12] Etizioni, O., Cafarella, M., Downey, D., Kok, S., Popescu, A. M., Shaked, T., Soderland, S., Weld, D., and Yates A., Web-scale information extraction in KnowItAll. In Proceedings of the 13th World Wide Web Conference (WWW).

[13] Lafferty, J., A. McCallum, and F. Pereira. 2001. Conditional Random Fields: Probabilistic Models for Sequence Data. Proceeding of 18thICML. San Francisco.

[14] C. J. van Rijsbergen. 1979. Information Retrieval. Butterworth, 2nd edition.