

Comparing Statistical and Data Mining Techniques for Enrichment Ontology with Instances

Aurawan Imsombut¹ and Jesada Kajornrit^{2,*}

¹College of Creative Design and Entertainment Technology, Dhurakij Pundit University, Bangkok, Thailand

²College of Innovative Technology and Engineering, Dhurakij Pundit University, Bangkok, Thailand

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 with two popular techniques being the statistical and data mining methods. The paper compares the use of the two methods to classify instances to enrich ontology having greater domain knowledge, and selects a conditional random field for the statistical method and feature-weight k-nearest neighbor classification for the data mining method. The experiments are conducted on tourism ontology. The results show that conditional random fields methods provide 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 (Zhang and Ciravegna 2011). 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 effort. 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 (Chinchor 1998). 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 and Katakana as used in the Japanese language. Then, there is a challenging task to extract NE in the Thai language.

There are many techniques to extract instances of concepts (that is, NE). However, two popular ones are the statistics and data mining methods (classification). This paper compares these two techniques to classify instances, that is, Conditional Random Fields (CRFs) for the statistics methods and feature-weight k-Nearest Neighbor (KNN) classification of data mining methods for extracting ontology instances (Imsombut and Sirikayon 2016; Imsombut and Paireekreng 2016). 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 are normally 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 traditional techniques such as kNN are used to classify, the problems related to majority class bias occur. Therefore, feature-weighted kNN is

*Address of correspondence to this author at the College of Innovative Technology and Engineering, Dhurakij Pundit University, Bangkok, Thailand; Tel: (+66) 029547300; Fax: (+66) 025899605; E-mail: jesada.kaj@dpu.ac.th

JEL: C38, C44, C82.

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, and Section 3 presents a brief review of the methods used. The data and experiments are presented in Section 4, and Section 5 presents the results and discussion. Section 6 provides some concluding comments.

2. RELATED WORKS

There are many studies concerned with ontology enrichment (that is, define and classify instances). Most of these studies apply NLP techniques with Information Extraction (IE) techniques and Machine Learning (ML) techniques.

Martinez *et al.* (2011) proposed a combination of NLP and IE techniques by using GATE tools for extracting NE from restaurant and hotel corpus, and used heuristic algorithm for solving different kinds of ambiguities to populate the instances into tourism ontology. Faria *et al.* (2012) 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.* (2009) applied NLP and ML techniques called Maximum Entropy to extract relationships between entities for tourism. Nanba *et al.* (2009) applied NLP and used CRF as ML in order to identify travel blogs, and extracted travel information relating to the relationships between location names and local products. Carlson *et al.* (2010), Giuliano and Gliozo (2008), Cimiano *et al.* (2005) and Etzioni *et al.* (2004) applied the NLP, IE and ML techniques to the ontology population.

3. BRIEF REVIEW OF METHODS

This section briefly reviews the literature of statistical techniques: Conditional random fields (CRFs) and data mining technique; and feature-weight k-Nearest Neighbor classification

3.1. Statistical Techniques

Conditional random fields (CRFs) is a statistical technique that is usually used for pattern recognition, especially in the natural language processing area. CRFs (Lafferty *et al.* 2001) 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 $s = \langle s_1, s_2, \dots, s_T \rangle$, given an observation sequence $o = \langle o_1, o_2, \dots, o_T \rangle$, is defined as:

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

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:

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

3.2. Data Mining Techniques

There are several classification techniques. Nevertheless, kNN is chosen for this paper 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 e_j 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 the 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:

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

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

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

However, there are some limitations of kNN as it tends to classify data based on the majority class.

Therefore, it is not appropriate to classify using unbalanced data. The feature-weighted kNN classifier (Vivencio *et al.* 2007) can mitigate the problem by using the weight of feature, where 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, which can decrease the overall significance of the classification as follows:

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

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, correlations lie 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:

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

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, create all possible nodes, and select the best possible node for the answer. CRFs technique is 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 the 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 numbers of data are large, the amount of memory used is increased. This limitation causes the CRF technique be unsuitable for large data. Some techniques, such as feature selection, are needed to reduce this limitation.

Feature-weighted kNN is a classification technique. The kNN performs fast as its simple mechanism, and classifies data by using k -closed training data. The

feature-weighted step return greater weights on features that have more effects than on features that have fewer 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 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 (that is, 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 to 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 values 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 paper 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 approximately 40,000 words) from Thai tourism websites were used. The contents in the website were, for example, attractions, accommodations, and activities. These documents were pre-processed and performed instance extraction. In order to evaluate the performance of the classification of attraction category, F1, which was proposed by van Rijsbergen (1979), will be used. It applied precision and recall as follows:

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

$$Recall = \frac{\text{number of correct positive predictions}}{\text{number of positive examples}} \times 100, \quad (9)$$

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

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

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

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%
Average	77.62%	60.84%	70.87%	59.26%	74.09%	60.04%

kNN method because the CRFs technique is a more suitable technique to direct class for sequential data. Moreover, CRFs can handle unbalanced data better.

The data used in the experiments were obtained 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 experiments.

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

REFERENCES

- 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). <https://doi.org/10.1145/1718487.1718501>
- Chinchor, N. 1998. "MUC-7 named entity task definition dry run version, version 3.5" Proceedings of the Seventh Message Understanding Conference (MUC-7) (to appear). Fairfax, Virginia: Morgan Kaufmann Publishers, Inc.
- 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). <https://doi.org/10.1145/1060745.1060796>
- Etzioni, O., Cafarella, M., Downey, D., Kok, S., Popescu, A. M., Shaked, T., Soderland, S., Weld, D., and Yates A. 2004. "Web-scale information extraction in KnowItAll" In Proceedings of the 13th World Wide Web Conference (WWW-04), pp. 100-110. <https://doi.org/10.1145/988672.988687>
- Faria, C., Girardi, R. and Novais, P. 2012. "Using Domain Specific Generated Rules for Automatic Ontology Population" Proceedings of 12th International Conference on Intelligent Systems Design and Applications. <https://doi.org/10.1109/isda.2012.6416554>
- 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). <https://doi.org/10.3115/1599081.1599115>
- Imsombut, A. and Paireekreng, W. 2016. "Extract Knowledge for Populating Thai Tourism Ontology from Texts Using Feature-weighted k-Nearest Neighbor" Proceedings of 1st International Conference on Information Technology.
- Imsombut, A. and Sirikayon, C. 2016. "An Alternative Technique for Populating Thai Tourism Ontology from Texts Based on Machine Learning" Proceedings of 15th IEEE/ACIS International Conference on Computer and Information Science (ICIS 2016). <https://doi.org/10.1109/ICIS.2016.7550762>
- J.M. Ruiz-Martínez, J.A. Miñarro-Giménez, D. Castellanos-Nieves, F. García-Sánchez, R. Valencia-García. 2011. "Ontology Population: An Application for the e-tourism Domain" International Journal of Innovative Computing, Information and Control (IJICIC), 7 (11) (2011), pp. 6115–6134.
- Lafferty, J., A. McCallum, and F. Pereira. 2001. "Conditional Random Fields: Probabilistic Models for Sequence Data" Proceeding of 18th ICML. San Francisco.
- Nanba, H., Taguma, H., Ozaki, T., Kobayashi, D., Ishino, A. & Takezawa, T. 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).
- Van Rijsbergen, C. J. 1979. "Information Retrieval" Butterworth, 2nd edition.
- Vivencio, D.P., Hruschka Jr., E.R., Nicoletti, M., Santos, E., Galvao, S. 2007. "Feature-weighted k-nearest neighbor classifier" In Proceedings of IEEE Symposium on Foundations of Computational Intelligence (FOCI 2007).
- Yihao Zhang, Jianyi Guo, YU Zhengtao, Yao Xianming, Zhang Shaomin. 2009. "Automatic Entity Relation Extraction for the Field of Tourism" Journal of Computational Information System, vol. 5, no. 6, pp. 1653-1659.
- Zhang, Z. and Ciravegna, F. 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.

Received on 16-02-2017

Accepted on 13-05-2017

Published on 09-06-2017

DOI: <https://doi.org/10.6000/1929-7092.2017.06.39>

© 2017 Imsombut and Kajornrit; Licensee Lifescience Global.

This is an open access article licensed under the terms of the Creative Commons Attribution Non-Commercial License (<http://creativecommons.org/licenses/by-nc/3.0/>) which permits unrestricted, non-commercial use, distribution and reproduction in any medium, provided the work is properly cited.