Automatic Advertisement System using Text Mining Technique

Chatklaw Jareanpon

POLAR LAB, Department of Computer Science, Faculty of Informatics, Mahasarakham University, 44150, THAILAND

*Corresponding Author: chatklaw.j@gmail.com, chatklaw.j@msu.ac.th

Abstract

Nowadays, there are a wide variety of advertising types. Indeed, one of them is the Internet advertising and it is growing rapidly. However, with the different behaviors of consumers who visit a web site, many websites are developed and proposed in order to meet the needs of many different people. Indeed, if a website is interested by many people, rating of the website is increased. However, after accessing to the website, if the Internet advertisings found on the website does not relate with the consumer interests, these online advertising may not be effective. As the result, this work presents a method of text mining approach to suggest the online advertisements that are relevant to the content of the website. After experimenting by the Random Forest and Bayesian Network classifier, the result of Random Forest is better than the result of Naïve Bayes, where the Random Forest shows the accuracy at 83.97%

Keywords: Advertising, Text Mining, Fit Ensemble (Random Forest) algorithm, Naive Bayes algorithm, Classification

1. Introduction

The internet advertising is very growing rapidly. Many websites allow to show and relate the advertising such as Google Adwords. The cost comes from user's click called PPC (Pay per Click). Sometimes, the advertising did not match with the content of websites, then user is not interested in the advertising.

Many research interested in Text Mining techniques. Un⁽¹⁾ proposed the Text Mining with Information Extraction called DISCOTEXT. This algorithm is able to discover the patterns in extracted text, strict matching of string. Jacopo et al⁽²⁾ proposed the text mining in computational

advertising using n-grams, topic models and sentiment analysis. Vishal⁽³⁾ proposed the survey of text mining classification. Text mining is similar to the data mining, but text mining is able to work with unstructured data sets such as emails, documents and HTML. Hans⁽⁴⁾ proposed the Decision Trees and Text mining technique for extending taxonomics for ontology. Xiao⁽⁵⁾ proposed the text mining in music mood classification by collecting the 8,839 songs with tags and lyrics. Bjornar and Chinatsu⁽⁶⁾ proposed the fast and effective text mining using k-means linear-time document clustering.

From the previous research, the text mining is used for classification the document in different objective, and able to use for the unstructured data such as HTML. This research interested in the content and tried to extract the characteristic of website using text mining technique. Moreover, this research compares the accuracy rate of selected classification algorithms to create the automatic Advertisement System.

2. Automatic Advertisement System using Text Mining Technique

This system consisted of 4 steps that are Pre-processing, Feature Extraction, Model generating, and Evaluation as shown in Figure 1.

2.1 Pre-Processing

The system firstly collects the content from webpage. This research creates the training set from the 1600 webpages that separates and labels into 8 classes (200 pages per class): immovable property, sport, car, computer, education, games, health, and travel.

2.1.1 HTML tag removing is the process to remove the tag of HTML such as <body>, <head>, <html>.

The HTML tag will effect with the frequency of content but it is unable to represent the character of the class.

2.1.2 Word Segmentation is the process to separate the content from sentence to word. This research interested in THAI language. The best algorithm for THAI segmentation is Longest Word Matching Algorithm⁽⁷⁾ as shown in Figure 2.

Fig. 1. Automatic Advertisement System.

2.1.3 Stop word is the process to remove the unnecessary word which is not effecting with the content meaning, but it is usually found in the content that will effect with the frequency feature extraction. The stop word is the preposition, conjunction, pronouns, adverb, and interjection. The example of the frequency stop word is that and which.

From this process, the content will separate to clear word removing the HTML tag and stop word.

2.2 Feature Extraction

This process is extracting the word to the feature. Firstly, all word of all class will count the frequency, and select the top 100 words to create the feature called Bag of

words. The weight of feature extraction is found in many ways such as

Fig. 2. Longest Word Matching Algorithm.

2.2.1 Boolean weighting is calculated from equation (1)

$$W_{ik} = \begin{cases} 1 & \text{if } Tf_{ik} > 0 \\ 0 & \text{otherwise} \end{cases} \quad (1)$$

2.2.2 Tf-weighting (Term frequency) is calculated from equation (2)

$$W_{ik} = Tf_{ik} \quad (2)$$

2.2.3 Tf*Idf-weighting (Term frequency-inverse document frequency) is calculated from equation (3)

$$TfIdf_{ik} = Tf_{ik} \times \log\left(\frac{N}{n_i}\right) \quad (3)$$

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

2.2.4 Tfc-weighting is calculated from equation (4)

$$Tfc_{ik} = \frac{Tf_{ik} \times \log\left(\frac{N}{n_j}\right)}{\sqrt{\sum_{j=1}^M \left[Tf_{ij} \times \log\left(\frac{N}{n_1}\right) \right]}} \quad (4)$$

2.2.5 ltc-weighting is calculated from equation (5)

$$ltc_{ik} = \frac{\log(Tf_{ik} + 1.0) \times \log\left(\frac{N}{n_1}\right)}{\sqrt{\sum_{j=1}^M \left[\log(Tf_{ij} + 1.0) \times \log\left(\frac{N}{n_1}\right) \right]}} \quad (5)$$

2.2.6 Entropy weighting is calculated from equation (6)

$$\text{entropy}_{ik} = \log(Tf_{ik} + 1.0) \times \left(1 + \frac{1}{\log(N)} \sum_{j=1}^M \left[\frac{Tf_{ij}}{n_i} \times \log\left(\frac{f_{ij}}{n_i}\right) \right] \right) \quad (6)$$

After this process, the feature will select by frequency of the word, and calculate the weight. The next process is try to generate the model by the classification.

2.3 Model Generating

The classifier of text mining can be classify by three ways that are unsupervised, supervised and semi supervised classifier for example Bayesian network classifier, Decision Tree, K-nearest neighbor (KNN), Support Vector Machines (SVMs), Neural Network. This research selected to use and compare the 2 different classifier that are Bayesian network and Random forest classifier.

2.3.1 Bayesian Classifier⁽⁸⁾ is based from the priori probability and conditional probability. The advantage of the Bayesian classifier is that it requires a small training data, and it is robust to the missing data. The Bayesian Classifier can be calculated from equation (7)

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \quad (7)$$

where P(A) Prior probability of hypothesis A

P(B) Prior probability of training data B

P(A|B) probability of A given B

P(B|A) probability of B given A

2.3.2 Random forest Classifier⁽⁹⁾ is as follows:

1. N is the number of instances, and M is the attributes.

2. m of input variable to be used to determine the decision at a node of the tree; m is less than M.

3. Choosing N times with replacement from all N for this tree.

4. Each node in the tree randomly chooses m variable base the decision at that node. Each tree is fully grown and not pruned.

The advantage of random forests are it produces a highly accuracy rate, it is able to handle the large number of attributes, and fast.

The main objective is trying to test between the required small training data algorithm and handled large number of attribute data.

2.4 Evaluation

This paper will test the data using K-Fold cross validation and Hold-out cross validation, and evaluate the model using accuracy rate calculated by Table 1 and equation 8.

Table 1. Accuracy variables.

		Predicted Label	
		Positive	Negative
Known Label	Positive	TP	FN
	Negative	FP	TN

where

$$\text{Accuracy rate} = \frac{TP + TN}{TP + TN + FP + FN} \quad (7)$$

3. Experimental Result

This research collects the 1600 webs from the web directory divided into 8 classes (200 webs per class), and separates the experiment into 5 experiments. JAVA language is used to develop and test the algorithm.

3.1.1 Experimental time is divided in 3 steps showed in Table 2.

Table 2 Experimental time.

Data Preprocessing	20.63 Minutes
- Bag of words	
- Feature Extraction	
Model Generating	17.18 Minutes
Model Usages Averages	0.215 Seconds

This research tested the various popular feature weights that are 1) Term Frequency (TF), Term Frequency-Inverse Document Frequency (TF*IDF) and Boolean.

3.1.2 Accuracy rate tested by Hold-out cross validation. The confusion matrix of Bayesian classifier will show in Table 3(a-c) and Figure 3, and of Random forest classifier will show in Table 4(a-c) and Figure 4.

Table 3(a). Confusion of Bayesian Network using TF.

		Classification System Output							
		Car	Computer	Education	Game	Health	Immovable	Sport	Tourism
Target Output	Car	27	1	1	1	2	3	1	3
	Computer	1	29	1	1	3	5	2	0
	Education	3	1	32	4	3	2	2	0
	Game	2	4	1	29	1	2	5	0
	Health	4	2	2	0	28	1	1	1
	Immovable	0	3	0	0	0	26	1	1
	Sport	0	0	1	4	2	0	28	1
	Tourism	3	0	2	1	1	1	0	34
Accuracy		72.81%							

Table 3(b). Confusion of Bayesian Network using TF*IDF.

		Classification System Output							
		Car	Computer	Education	Game	Health	Immovable	Sport	Tourism
Target Output	Car	26	2	1	1	2	4	2	2
	Computer	1	26	4	2	3	4	2	0
	Education	4	3	29	3	3	3	0	0
	Game	3	4	0	29	1	1	5	0
	Health	2	2	2	0	28	0	2	1
	Immovable	1	3	0	0	0	27	0	1
	Sport	0	0	1	4	2	0	29	1
	Tourism	3	0	3	1	1	1	0	35
Accuracy		71.56%							

Table 3(c). Confusion of Bayesian Network using Boolean.

		Classification System Output							
		Car	Computer	Education	Game	Health	Immovable	Sport	Tourism
Target Output	Car	29	2	2	1	0	0	1	1
	Computer	7	32	2	6	6	15	11	1
	Education	1	3	31	1	6	0	3	0
	Game	1	1	1	28	0	1	4	0
	Health	0	1	2	1	27	1	2	2
	Immovable	0	1	1	0	1	23	0	1
	Sport	1	0	0	3	0	0	18	0
	Tourism	1	0	1	0	0	0	1	35
Accuracy		69.69%							

Table 4(a). Confusion of Random Forest using TF.

		Classification System Output							
		Car	Computer	Education	Game	Health	Immovable	Sport	Tourism
Target Output	Car	35	1	2	2	0	0	1	0
	Computer	2	35	3	6	0	9	3	0
	Education	0	1	30	1	2	2	3	0
	Game	1	1	1	29	0	0	3	0
	Health	0	2	2	0	37	1	2	1
	Immovable	0	0	1	0	1	27	1	1
	Sport	1	0	1	1	0	0	26	0
	Tourism	1	0	0	1	0	1	1	38
Accuracy		80.31%							

Table 4(b). Confusion of Random Forest using TF*IDF.

		Classification System Output							
		Car	Computer	Education	Game	Health	Immovable	Sport	Tourism
Target Output	Car	35	1	1	3	0	0	0	0
	Computer	2	33	4	3	1	7	3	0
	Education	0	1	29	2	0	3	4	0
	Game	1	2	1	27	0	0	1	0
	Health	0	0	2	0	39	1	1	2
	Immovable	0	2	1	0	0	28	1	0
	Sport	1	1	1	2	0	0	30	0
	Tourism	1	0	1	3	0	1	0	38
Accuracy		80.94%							

Table 4(c). Confusion of Random Forest using Boolean.

		Classification System Output							
		Car	Computer	Education	Game	Health	Immovable	Sport	Tourism
Target Output	Car	30	1	1	4	0	0	1	0
	Computer	2	34	3	5	0	6	3	0
	Education	2	1	31	2	3	2	4	0
	Game	1	0	0	27	0	0	4	0
	Health	0	2	3	0	36	2	2	3
	Immovable	2	1	2	0	1	28	1	0
	Sport	1	1	0	2	0	1	24	0
	Tourism	2	0	0	0	0	1	1	37
Accuracy		77.19%

Figure 3. The comparison of accuracy of various feature weight.

Figure 4. The comparison of accuracy of various feature weight.

3.1.3 Accuracy rate tested by K-fold cross validation. This research selected to use $k = 5$ as shown in Table 5 and 6. The comparison between Bayesian Classifier and Random Forest will show in Figure 5.

Table 5. Accuracy Bayesian Classifier by K-Fold Cross Validation

K	Accuracy Bayesian Classifier by K-Fold Cross Validation		
	TF	TF*IDF	Boolean
K = 1	67.81	70.00	72.18
K = 2	72.81	71.87	73.75
K = 3	71.87	70.62	73.75
K = 4	69.68	70.00	70.31
K = 5	72.81	71.56	69.68
Average	71.00	70.81	71.93
Standard Deviation	2.19	0.87	1.89

Table 6. Accuracy Random Forest by K-Fold Cross Validation

K	Accuracy Random Forest by K-Fold Cross Validation		
	TF	TF*IDF	Boolean
K = 1	84.37	84.37	83.75
K = 2	80.31	79.37	79.37
K = 3	79.06	79.06	78.43
K = 4	80.00	77.50	79.06
K = 5	80.31	80.93	80.62
Average	80.81	80.25	83.75
Standard Deviation	2.05	2.60	2.11

Figure 5. The comparison between Bayesian Classifier and Random Forest classifier.

3.1.4 Paired T-Test Statistics used to test the significant of feature weight selected the accuracy rate to test from the K-Fold cross validation experiment. The significant rate (α) is equal to 0.05. The paired T-Test experimental result will show in Table 7 and Table 8.

Table 7. T-Test Bayesian Network Classifier Result

T-Test Bayesian Classifier			Result
Feature		P-Value	
TF	TF*IDF	0.7904	Accepted main hypothesis
TF	Boolean	0.4818	Accepted main hypothesis
TF*IDF	Boolean	0.2684	Accepted main hypothesis

Table 8. T-Test Random Forest Classifier Result

T-Test Random Forest Classifier			Result
Feature		P-Value	
TF	TF*IDF	0.3602	Accepted main hypothesis
TF	Boolean	0.0705	Accepted main hypothesis
TF*IDF	Boolean	1.0000	Accepted main hypothesis

From Table 7 and Table 8, The P-Value is greater than significant rate that means the various feature weights for each classifier is not significant effected accuracy rate.

3.1.5 Paired T-Test Statistics, this research interested in significant of comparison algorithms. The significant rate (α) is equal to 0.05. The paired T-Test experimental result will show in Table 9.

Table 9. Paired T-Test between Random Forest and Bayesian Network Classifier Result

T-Test			Result
Algorithm		P-Value	
Random Forest	Bayesian Network	0.0038	Not Accepted main hypothesis

From Table 9, The P-Value is less than significant rate that means the algorithms is significant with the accuracy.

4. Discussion and Conclusions

From the experimental result is able to summarize as follows:

The highly accuracy from Hold out cross validation of Bayesian Network is from Term Frequency (T) weight at 72.81%, and of Random Forest is from Term Frequency-Inverse Document Frequency (TF*IDF) at 80.93%.

The highly accuracy from K-fold cross validation of Bayesian Network is from Boolean weight at 71.94%, and of Random Forest is from Boolean weight at 89.75%.

However, the paired T-test algorithm results found that the different feature weight is not effecting with the accuracy rate, but the different algorithm is effecting.

The Random Forest is showing the higher performance than Bayesian Network Classifier, because of highly number of Bag of Words. The future work will combine the sentiment analysis from social network such as Facebook because of the various contents. It is difficult to classify.

Acknowledgment

This research supported by Faculty of Informatics Mahasarakham University and Mahasarakham University.

References

- (1) Un Yong Nahm, Text Mining with Information Extraction, Technical report, UTEXAS, 2011.
- (2) Jacopo Soriano, Timothy Au and David Banks, Text mining in computational advertising, Statistical Analysis and Data Mining: The ASA Data Science Journal, Special Issue: Special Issue on Statistical Learning, Volume 6, Issue 4, pages 273–285, August 2013.
- (3) Vishal Gupta, A Survey of Text Mining Techniques and Applications, Journal of Emerging Technologies in Web Intelligence, Volume 1, No. 2, August, 2009.
- (4) H.F. Witschel. Using decision trees and text mining techniques for extending taxonomies. In Proc. of Learning and Extending Lexical Ontologies by using Machine Learning Methods, Workshop at ICML-05, 2005.
- (5) Xiao Hu, J. Stephen Downie, and Andreas F. Ehmann, Lyric Text Mining in Music Mood Classification, The Tenth International Society for Music Information Retrieval Conference (ISMIR 2009), pp. 411-416, 2009.
- (6) Bjornar Larsen, and Chinatsu Aone, Fast and effective text mining using linear-time document clustering,
- (7) Choochart Haruechaiyasak, Sarawoot Kongyoung, and Matthew N. Dailey, A Comparative Study on Thai Word Segmentation Approaches, In Proceedings of ECTI-CON, 2008. Proceeding
- (8) KDD '99 Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 16-22, 2009.
- (9) Tom M. Mitchell, Machine Learning. The McGraw-Hill Companies, Inc., 1997.
- (10) Leo Breiman, Random Forest, Machine Learning, Vol 45(1), pp. 5-32, 2001.