Advanced Materials Research Vols. 931-932 (2014) pp 1360-1364
Online available since 2014/May/09 at www.scientific.net
© (2014) Trans Tech Publications, Switzerland
doi:10.4028/www.scientific.net/AMR.931-932.1360

Classification of Tourism Web with Modified Naïve Bayes Algorithm

Naruepon Panawong[1,a], Chakkrit Snae Namahoot[1,b*] and Michael Brückner[2,c]

[1]Department of Computer Science and Information Technology, Naresuan University, Thailand

[2]Department of Educational Technology and Communication, Naresuan University, Thailand

[a]jnaruepon.p@gmail.com, [b]chakkrits@nu.ac.th, [c]michaelb@nu.ac.th

Keywords: Classification, Information Retrieval, Naïve Bayes, Tourism Ontology, Web Analysis

Abstract. In this paper we report results of a research aimed at classification Web contents on tourism with a modified Naïve Bayes algorithm. We used Web pages relating touristic information about Thailand. An appropriate light-weight tourism ontology with related terms was used to improve the results, which were categorized into six categories (attractions, accommodation, dining, local product markets, One Tambon One Product (OTOP) shops, and events). The Naïve Bayes algorithm generates results for each category, but Web pages can contain diverse information about tourism spanning over groups. The initial Web classification system could not categorize 130 sites (27.40%) out of 475 tested pages, because those Web pages contain words from more than one category. Therefore, we modified the Naïve Bayes algorithm to improve the efficiency of Web classification, which was then tested with the help of F-Measure: the results show 100% for precision, 97.39% for recall, and 98.58% for F-measure.

Introduction

Currently, a lot of Web pages provide services for searching for travel information. However, tourists need to know more about touristic places and areas, e.g. relating attractions, hotels, dining, local product markets, One Tambon One Product (OTOP) shops, and events. Most tourists search for interesting areas with a search engine, but the set of search results is often difficult to consume and confusing to understand because of the overload of information, mixed unwanted information, and uncategorized incoherent presentation. This causes time wasted extracting all relevant information and leads to inconvenient information gathering from a single information source.

The Naïve Bayes algorithm is often applied to this problem and uses probability for document categorization; e.g. in [1] document clustering by ten separate groups including Requisition, Complaint, Acknowledgement, Inquiry, Sales, Purchase, Goodwill, Feedback, Maths, and Computer is presented. A review of machine learning for online document classification of online news, blogs, e-mail, digital library [2] refers to Rocchio's Algorithm, K-Nearest Neighbor, Decision Tree Decision, Rules Classification, Naïve Bayes Algorithm, Artificial Neural Network, Fuzzy correlation, Genetic Algorithm, and Support Vector Machine and states that Naïve Bayes is the best for e-mail, numeric and document classification. Naïve Bayes also needs less data for the learning step and is not too complex. Also, Naïve Bayes is convenient to develop when compared to other algorithms [3].

According to the Classification Site Attractions classification training of tourism Web pages on Thailand in [4] and [5] 1,048 web directory data from True hits Web page were used for learning and 475 result Web pages from Google as a set of test data for evaluating the classification results. This methodology applied keywords and six categories (attractions, accommodation, dining, local product markets, OTOP shops, and events). Each Web page was calculated applying Naïve Bayes algorithm with a threshold value for classification and correction. The experimental results showed 72.60% for precision, 70.99% for recall, and 71.61% for F-Measure. However, with Naïve Bayes 130 Web pages (or 27.40% of the test data) could not be categorized because of low volume in word frequency for the right category. Moreover, keywords did not cover synonyms and similar words that describe the same ideas, and this caused misclassified texts as well. Some Web pages

All rights reserved. No part of contents of this paper may be reproduced or transmitted in any form or by any means without the written permission of TTP, www.ttp.net. (ID: 202.28.21.98-11/05/14,03:30:08)

cover specific details for more than one subject, so they had to be classified belonging to several categories. Naïve Bayes algorithm calculates probability as a single value for classification Web pages and gives the minimum value, which affects results for the mixed detail Web pages. This leads to wrong classification and decreased precision, recall, and F-Measure.

In this paper we reduce the error in Web page classification and apply sets of words, thereby introducing an improvement to Naïve Bayes algorithm not only by adding synonyms and related words into the classification technique but also allowing the categorization of a Web page into more than one category. For example, ทาน (than), กิน (kin), and ดื่ม (duem) represent Dining, whereas นอน (norn) and พักผ่อน (phakphorn) represent accommodation, and so on. In the following we lay out the methodology of this research, report on the testing and results, and finally draw conclusions and indicate some directions of future work.

Methodology

In this paper we propose a modification of the Naïve Bayes algorithm resulting in the classification of Web documents as follows:

1. Apply Naïve Bayes algorithm for learning with 1,048 tourism Web pages from Truehits and six categories including Attraction (233 Web pages), Accommodation (200 Web pages), Dining (318 Web pages), Souvenir (54 Web pages), OTOP (88 Web pages), and Events (155 Web pages).
2. Modified Naïve Bayes algorithm for a Thailand tourism Web classification for increasing efficiency.
3. Test of Web classification applying 475 Web pages from Google search results with Naïve Bayes and the modified Naïve Bayes algorithms.
4. Efficiency comparison between Naïve Bayes algorithm and modification Naïve Bayes using F-Measure.

Naïve Bayes algorithm

The Naïve Bayes algorithm is based on Eq. (1).

$$\text{Cmap} = \underset{c \in C}{\operatorname{argmin}} \left(C_i\right) \tag{1}$$

$$\text{Define} \quad C_i = \log(P(c)) + \sum_{i=1}^{6} \log(P(t_i, | c))$$

Where Camp is web classification result. Ci is probability of Attraction (At), Accommodation (Ac), Dining (D), Souvenir (S), OTOP (O) and Events (E). ti is word frequency, C is category, P(c) is a probability each of category and i is category counter 1 to 6 (including Attraction, Accommodation, Dining, Souvenir, OTOP and Event)

Modification of the Naïve Bayes algorithm

Naïve Bayes algorithm generates one (partial) probability for a Web page belonging to each of the six categories. Thailand Tourism Web Clustering System using Naïve Bayes algorithm not cover words for classification. So, we add many words from variation word and dictionary and modification Naïve Bayes algorithm for web classification efficient. We calculate minimum and maximum probability values for all Web pages in the data set using Naïve Bayes algorithm for tourism classification. In Fig.1 two steps are indicated: Learning and Testing.

Learning

As the learning data set we used 1,048 Web pages from Truehits Web directory and divide into six categories (Attractions, Accommodation, Dining, Souvenir, OTOP, and Events). We removed HTML Tags and calculate with Naïve Bayes (Ci value from equation 1) using terms from tourism ontology for tourism Web classification. Then, we calculated minimum and maximum values for the appearance of words in each of category. Finally, we defined CiMin minimum and CiMax maximum values of each category for Web classification automatically (Table 1).

Testing

1. Use Web pages from test data set for web classification using modification Naïve Bayes (Ci value from equation 1) and use words from tourism ontology [6,7] for web classification.
2. Compare Ci with CiMin and CiMax in Table 1.
3. If Ci between CiMin and CiMax then categorize
4. Repeat for all categories and summarize Web classification results.

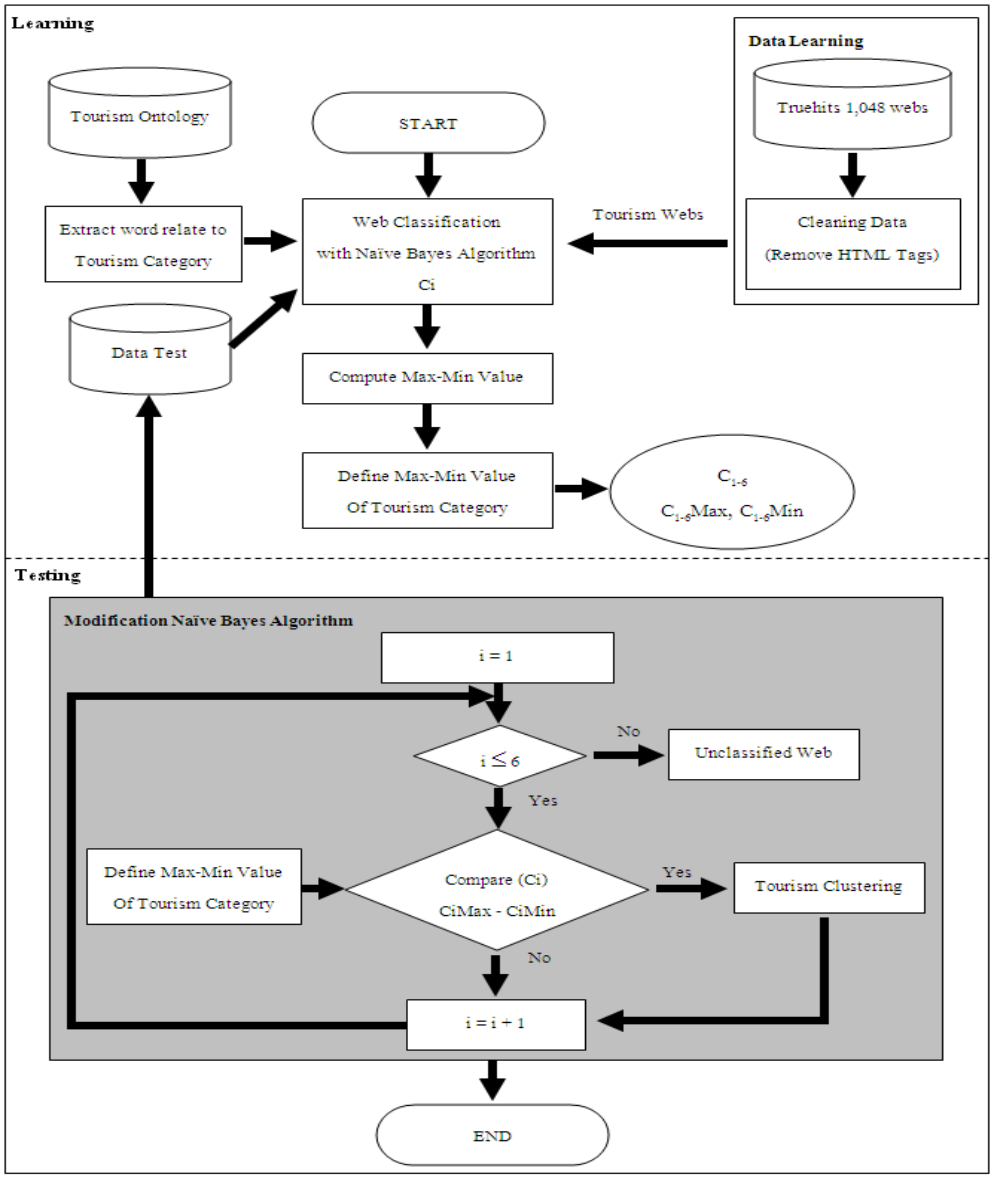

Fig. 1 Learning and Testing Process

Table 1. Minimum and Maximum value for each category using Naïve Bayes (1,048 web pages)

Category	Minimum Value	Maximum Value
Attraction	-91.85	-1.01
Accommodation	-23.61	-1.16
Dining	-50.44	-1.01
Souvenir	-5.47	-1.65
OTOP	-7.04	-1.45
Events	-21.12	-1.21

The modified Naïve Bayes algorithm for Web classification can classify a Web page into more than one category, which is more in line with the diverse content on tourism information Web pages. This is an improvement of efficiency.

Testing and Result

We used 475 tourism Web pages for classification. We calculated Ci value from Eq. (1). After Web analysis for tourism Web classification we calculated the F-Measure [6] for measuring the efficiency by Eq. (2), for which results are shown in Table 2.

$$F-\text{measure} = 2*\left(\frac{P*R}{P+R}\right) \tag{2}$$

P is True Positive/(True Positive + False Positive)
R is True Positive/(True Positive + False Negative)
True Positive is web site in category and system was in that category
False Positive is web site not in category and system was in that category
False Negative is web site in category and system was not in that category

Table 2 shows a comparison of Web analysis for tourism Web classification using Naïve Bayes algorithm and modified Naïve Bayes algorithm. Precision, recall and F-Measure have been calculated. Naïve Bayes algorithm results in 72.60% for precision, 70.99% for recall, 71.61% for F-Measure, because Naïve Bayes algorithm classifies the Web pages into one category. Modified Naïve Bayes algorithm leads to 100% for precision and Web pages can be in more than one category. Moreover, the F-Measure data show that the modified Naïve Bayes algorithm is more efficient in classification Web pages for Thailand tourism information.

Table 2. Comparison of Web analysis

Category	Naïve Bayes			Modification Naïve Bayes		
	P	R	F	P	R	F
Attraction	72.12%	84.07%	79.90%	100%	95.63%	97.69%
Accommodation	79.72%	80.85%	80.28%	100%	100%	100%
Dining	76.00%	70.37%	73.08%	100%	100%	100%
Souvenir	65.38%	62.96%	64.15%	100%	100%	100%
OTOP	61.11%	64.71%	62.86%	100%	95.24%	97.44%
Events	77.27%	62.96%	69.39%	100%	93.46%	96.34%
Average	**72.60%**	**70.99%**	**71.61%**	**100%**	**97.39%**	**98.58%**

Conclusion

This paper introduces a modified Naïve Bayes algorithm with increased efficiency, which was tested for Thailand tourism Web classification. We used tourism Web pages as a learning data set

and found minimum and maximum values for probabilities for each of the six categories used in the training. We use Google search results for testing the Naïve Bayes and the modified Naïve Bayes algorithms. Modified Naïve Bayes algorithm results in 100% for precision, 97.39% for recall and 98.58% for F-Measure. As further work we will use a temporal ontology for a recommendation system for tourist Web sites that takes date and time into account.

References

[1] K.K Sureshkumar, M. Umadevi, N.M. Elango, Divisive Clustering method using Naïve Bayes Algorithm for Text Categorization. International Journal of Advanced Research in Computer and Communication Engineering. 2:4 (2013), 1747-1753.

[2] A. Khan, B. Baharudin, L.H. Lee, K. Khan, A Review of Machine Learning Algorithms for Text-Documents Classification, Journal of Advances in Information Technology. 1:1 (2010) 4-20.

[3] T.M. Nogueira, S.O. Rezende, H.A. Camargo, On The Use of Fuzzy Rules to Text Document Classification, International Conference on Hybrid Intelligent Systems. (2010) 19-24.

[4] K. Chatcharaporn, T. Angskun, J. Angskun, Tourist Attraction Categorization Models using Machine Learning Techniques, Suranaree Journal of Science and Technology. 6:2 (2011) 35-58.

[5] N. Panawong, C. Snae Namahoot, Thailand Tourism Web Clustering System using Naive Bayes Algorithm, The 9th National Conference on Computing and Information Technology. (2013) 83-89.

[6] N. Panawong, C. Snae Namahoot, Performance Analysis of an Ontology-Based Tourism Information System with ISG Algorithm and Name Variation Matching. NU Science Journal. 9:2 (2013) 47-64.

[7] N. Panawong, C. Snae, Search System for Attractions in Thailand with Ontology and Name Matching. Journal of Information Science and Technology. 1:2 (2010) 60-69.

KKU International Engineering
10.4028/www.scientific.net/AMR.931-932

Classification of Tourism Web with Modified Naive Bayes Algorithm
10.4028/www.scientific.net/AMR.931-932.1360