ISBN 978-93-84422-80-6
17th IIE International Conference on Computer, Electrical, Electronics and Communication Engineering (CEECE-2017)
Pattaya (Thailand) Dec. 28-29, 2017

Framework for an Automatic Important Dates Calendar for International Conferences: Related Webpage Classification

Isaranon Promdee[1] and Kritsada Sriphaew[1]

College of Information and Communication Technology, Rangsit University, Pathumtani, Thailand
(corresponding author's phone: +669 8624 9996; e-mail: isaranon.p58@rsu.ac.th)

***Abstract*:** *We aim to build a framework to create a datahub portal that provides important date information for international conferences. To develop an automatic important dates calendar for international conferences, we crawl the related webpages, regardless of whether they include the important dates, and classify the pages using supervised machine learning techniques, including decision tree, k-nearest neighbor algorithm, support vector machine, and Naïve Bayes. We found that the decision tree provided the most accurate classification results (98.33%).*

Keywords: *supervised machine learning, text classification, text mining, webpages classification*

1. Introduction

This study aims to collect important dates for international conferences from webpages to construct a conference event calendar, similar to that available at www.conferencealert.com. However, this website requires manual handling of data. Therefore, real-time updates are not available to users.

The proposed framework of the important date calendar for international conferences includes a process to automatically update event information. Initially, we used a webpage crawler to identify webpages with content about international conferences. We classified the pages as positive or negative. Pages classified as positive contained information about an international conference, including important dates. Pages classified as negative also contained information about an international conference but did not include important dates. Note that webpages with content not related to an international conference were also classified as negative. Pages classified as positive were information extracted and created a datahub portal.

The remainder of this study is organized as follows. In Section 2, we define the webpage classification problem and provide a general overview of the text classification method used in the proposed framework. Section 3 describes the proposed methodology. The data used in the study and experimental results are presented in Section 4. Conclusions and suggestions for future work are provided in Section 5.

2. Related Work

Ozel and Sarac performed webpage classification using a vector space model for web-based searches, where users can search a database to retrieve an event that matches their preferences [1]. Abbasi et al. used 30 websites to detect fake medical websites using recursive trust labeling, and their results demonstrated high accuracy of 94% [2]. Kazemian and Ahmed used unsupervised machine learning (K-means and affinity propagation) and supervised machine learning techniques (k-nearest neighbor (K-NN), support vector machine (SVM), and Naïve Bayes) to detect malicious webpages. Their results indicated that the unsupervised machine learning techniques did not perform better than the supervised machine learning techniques, which provided greater than 89%

accuracy [3]. Therefore, we have chosen to employ supervised machine learning techniques in the proposed framework.

We examined the use of supervised machine learning techniques for text classification because the content of the webpages to be classified by the proposed framework is stored in a text file. In 2005, Ikonomakis et al. employed a supervised machine learning technique for text classification, and their results showed that the supervised machine learning technique demonstrated good performance [4]. In 2014, Menaka S. and Radha N. introduced journal data for text classification using decision tree, K-NN, and Naïve Bayes techniques, as well as keyword extraction using WordNet [5],[6]. Their results showed that the decision tree provided the best performance. In 2017, Lucini et al. studied the use of the early emergency department patient records for text classification using decision tree, K-NN, SVM, and Naïve Bayes techniques, and their results indicated that an SVM gave the best performance [7].

A decision tree construction algorithm takes advantage of the sparsity of text data and a rule simplification method converts the decision tree into a logically equivalent rule set [8]. K-NN is a method that can improve the performance of text classification using effectively estimated parameters [9]. Some variants of the K-NN method with different decision functions, k values, and feature sets have been proposed and evaluated to determine effective parameters. When applied to text classification tasks, SVMs provide excellent precision but poor recall. One means of customizing SVMs in order to improve recall is to adjust the threshold associated with the SVM. Shanahan and Roma described an automatic process to adjust the thresholds of a generic SVM with better results [10]. Naïve Bayes techniques are often used in text classification applications and experiments owing to their simplicity and effectiveness [11]. However, the performance of such methods is often degraded because they do not effectively model text. Schneider addressed such problems and demonstrated that they can be solved by some simple corrections [12]. Klopotek and Woch presented the results of an empirical evaluation of a Bayesian classifier based on a new method for learning very large tree-like Bayesian networks. Their results suggests that tree-like Bayesian networks can handle a text classification task with 100,000 variables at sufficient speed and accuracy [13].

In this study, we propose the use of supervised machine learning (decision tree, K-NN, SVM, and Naïve Bayes) techniques for webpage classification.

3. Proposed Framework

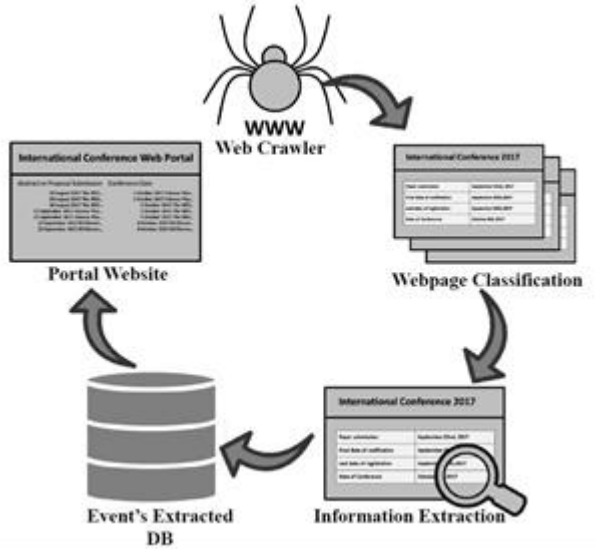

Fig. 1 Proposed event calendar framework

Figure 1 shows that the proposed framework for an automatic important date calendar for international conferences includes five components, i.e., a web crawler, webpage classification, information extraction, a database of extracted events, and a portal website. In this study, we focus on webpage classification.

4. Methodology

Our research methodology for webpage classification has six steps, including data collection, labeling, preprocessing, classification, evaluation, and determination of the most appropriate algorithm, as shown in Figure 2.

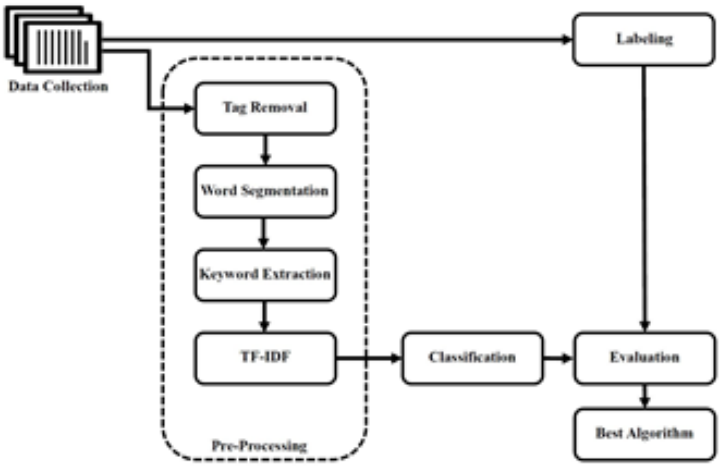

Fig. 2: Research methodology

4.1. Data Collection

We used RapidMiner Studio 7 to store text file from URLs obtained between April and December 2017. We ran a web crawler on 11,417 webpages (from 376 websites) and stored the data in a MySQL database as a text files. We then use same tool for duplicated the data to reduce data duplication.

4.2. Labeling

After data collection, we asked two experts to label two types of webpages. Each expert labeled the same dataset, and then we measured the inter-rater agreement (Cohen's kappa coefficient was 1).

4.3. Preprocessing

Our preprocessing involved the important steps of tag removal, word segmentation, keyword extraction, and determination of the term frequency-inverse document frequency (TF-IDF).

Tag removal: At this stage, the researchers write the program with PHP language to remove HTML tags for improving classification accuracy

Word segmentation: We used spaces as the basis for word division with RapidMiner Studio 7.

Keyword extraction: This involves identifying important or frequently used words and selecting words with the same meaning to reduce the number of attributes and increase the accuracy of the feature weight. In this study, we used Porter stemming algorithm and WordNet dictionary to perform keyword extraction.

TF-IDF is a numerical statistic intended to reflect how important a word is to a document in a collection or corpus. It is often used as a weighting factor in information retrieval, text mining, and user modeling. The TF-IDF value increases proportionally to the number of times a word appears in the document but is often offset by the frequency of the word in the corpus, which helps adjust for the fact that some words appear more frequently in general. TF-IDF is a popular term-weighting scheme. For example, 83% of text-based recommender systems in the digital libraries domain use TF-IDF. Variations of the TF-IDF weighting scheme are often used by search

engines as a central tool in scoring and ranking a document's relevance given a user query. TF-IDF can be successfully used for stop-word filtering in various subject areas, including text summarization and classification [14].

4.4. Classification

The classification process took 5,575 examples (Table 1) from preprocessing to create a model with the four selected classification techniques to predict the two types (positive or negative classes) of webpages. Note that we used 10-fold cross validation to build the models.

TABLE I: Number of Text File in this Work

DESCRIPTION	POSITIVE	NEGATIVE	TOTAL
Collected data	1,921	9,496	11,417
Duplicated data	-508	-3,216	-3,698
Deleted empty	-0	-2,118	-2,118
Used in this work	1,413	4,162	5,575

4.5. Evaluation

In an evaluation, we used RapidMiner Studio 7 to measure the accuracy, precision and recall for the positive class.

5. Result and Discussion

Table 2 shows that webpage classification with supervised machine learning techniques demonstrates good performance. It was found that keyword extraction using Porter's algorithm with a decision tree gave the highest accuracy (98.33%). This technique found notif, deadlin, congress, and counsel as the top keywords for classifying a related webpage, as shown in Figure 3. We then performed a significance test between Porter and WordNet. In this test, we set the level of significance to 0.05. The determined level of significance was 0.076. In summary, the two keyword extraction techniques did not differ significantly.

TABLE II: Compare Webpage Classification Performance

TECHNIQUES		ACCURACY	PRECISION	RECALL
STEMMING	CLASSIFICATION			
PORTER	DECISION TREE	98.33	98.89	99.42
	K-NN	87.22	87.05	84.17
	SVM	95.23	96.89	99.45
	NAÏVE BAYES	84.32	88.54	81.14
WORDNET	DECISION TREE	97.92	98.62	99.30
	K-NN	88.68	91.92	86.18
	SVM	95.37	96.68	99.21
	NAÏVE BAYES	75.41	80.82	69.41

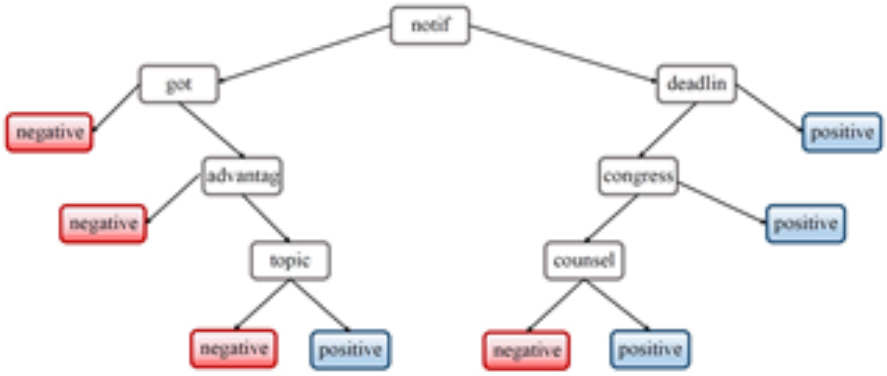

Fig. 3 Webpage classification tree

6. Conclusion and Future Work

Text classification is one of the major applications of machine learning. The proposed method uses text mining algorithms for webpage classification. The Porter and WordNet dictionary are used to calculate the semantic distances between keywords, i.e., words with the greatest similarity. Then, documents are classified based on the extracted keywords using machine learning algorithms, i.e., decision tree, K-NN, SVM, and Naïve Bayes. The performance analysis of supervised machine learning algorithms for text classification shows that the decision tree algorithm gave better results based on prediction accuracy when compared to the other three algorithms. In future work, we planning to use the best algorithm from this study to classify positive webpages for information extraction and create a datatahub portal.

7. References

[1] S. A. Ozel and E. Sarac, "Focused crawler for finding professional events based on user interests," in *2008 23rd International Symposium on Computer and Information Sciences*, 2008, pp. 1–4.

[2] A. Abbasi, F. ``Mariam'' Zahedi, and S. Kaza, "Detecting Fake Medical Web Sites Using Recursive Trust Labeling," *ACM Trans. Inf. Syst.*, vol. 30, no. 4, p. 22:1--22:36, Nov. 2012.

[3] H. B. Kazemian and S. Ahmed, "Comparisons of machine learning techniques for detecting malicious webpages," *Expert Syst. Appl.*, vol. 42, no. 3, pp. 1166–1177, 2015.

[4] E. Ikonomakis, S. Kotsiantis, and V. Tampakas, "Text Classification Using Machine Learning Techniques," *WSEAS Trans. Comput.*, vol. 4, pp. 966–974, 2005.

[5] Menaka S. and Radha N., "Text Classification using Keyword Extraction Technique," *Int. J. Adv. Res. Comput. Sci. Softw. Eng.*, vol. 3, no. 12, pp. 734–740, 2013.

[6] G. A. Miller, "WordNet: A Lexical Database for English," *Commun. ACM*, vol. 38, no. 11, pp. 39–41, Nov. 1995.

[7] F. R. Lucini *et al.*, "Text mining approach to predict hospital admissions using early medical records from the emergency department," *Int. J. Med. Inform.*, vol. 100, no. Supplement C, pp. 1–8, 2017.

[8] D. Johnson, F. J. Oles, T. Zhang, and T. Goetz, "A Decision-Tree-Based Symbolic Rule Induction System for Text Categorization," *IBM Syst. J.*, vol. 41, pp. 428–437, 2002.

[9] H.-S. Lim, "Improving kNN Based Text Classification with Well Estimated Parameters," vol. 3316. pp. 516–523, 2004.

[10] J. Shanahan and N. Roma, "Improving SVM Text Classification Performance through Threshold Adjustment," vol. 2837. pp. 361–372, 2003.

[11] S.-B. Kim, H.-C. Rim, D. Yook, and H. Lim, "Effective Methods for Improving Naive Bayes Text Classifiers," in *Proceedings of the 7th Pacific Rim International Conference on Artificial Intelligence: Trends in Artificial Intelligence*, 2002, pp. 414–423.

[12] K.-M. Schneider, "Techniques for Improving the Performance of Naive Bayes for Text Classification," in *In Proceedings of CICLing 2005*, 2005, pp. 682–693.

[13] M. A. Kłopotek, "Very large Bayesian multinets for text classification," *Futur. Gener. Comput. Syst.*, vol. 21, no. 7, pp. 1068–1082, 2005.

[14] J. Han, J. Pei, and M. Kamber, *Data Mining: Concepts and Techniques*. Elsevier Science, 2011.