Research Journal of Applied Sciences 8 (4): 230-234, 2013
ISSN: 1815-932X

Developing an Effective Automatic Web Pages Categorization Base on Ambiguity Weighting

Nivet Chirawichitchai
School of Information Technology,
Sripatum University Chonburi Campus, 20000 Chonburi, Thailand

Abstract: Web pages categorization is the process of automatically assigning predefined categories. Feature weighting which calculates feature (term) values in web pages is an important preprocessing technique in Web pages categorization. In this study, researchers proposed developing an effective automatic web pages categorization base on ambiguity weighting focusing on the comparison of various term weighting schemes. Researchers found ambiguity weighting most effective in the experiments with SVM, NB and DT algorithms. Researchers also discovered that the ambiguity weighting is suitable for combination with the information gain feature selection method. The ambiguity weighting with classification algorithms yielded the best performance with the F-measure over all algorithms. Based on the experiments, the classification algorithm with the information gain feature selection yielded the best performance with the F-measure of 99%.

Key words: Web pages categorization, ambiguity weighting, algorithm gain, F-measure, Thailand

INTRODUCTION

In recent years, researchers have seen an exponential growth in the volume of web pages available on the World Wide Web. While more and more textual information is available online, effective retrieval is difficult without organization and summarization of document content. Web pages categorization is one solution to this problem, Web pages from one or more Web sites are assigned to predefined categories according to their content. Web pages categorization is one of the essential techniques for Web mining. Specifically, classifying Web pages of a user-interesting class is the first step of mining interesting information from the Web. Traditionally, this task is performed manually by domain experts. However, human categorization is unlikely to keep pace with the rate of growth of the World Wide Web. Hence, as the World Wide Web continues to increase, the importance of automatic Web pages categorization becomes abvious. Moreover, automatic categorization is much cheaper and faster than human categorization. The techniques which have been applied to Web pages categorization have been studied extensively in recent decades and most of them are usually from the traditional machine learning such as support vector machines, K-nearest neighbor, Decision Tree, Naive Bayes, Neural Network, Linear Regression, etc.

Vector Space Model (VSM) (Salton and Lesk, 1968) is a major method for representing Web pages categorization. In this model, each Web pages is considered to be a vector in the feature space. Thus, one major characteristic of VSM is calculation of feature values in Web pages vectors. The processing that yields feature values is called feature weight. A widely used method for feature weight is tf x idf (Salton and Buckley, 1988), tf is the denote for term-frequency which stands for the capacity of features expressing Web pages content. idf is the denote for inverse document frequency which stands for the capacity of features discriminating similar Web pages. The motivation behind idf is that terms appearing frequently in many documents have limited discrimination power. Because methods of feature selection evaluate feature by scores, researchers can also adopt these methods for feature weight.

In this study, researchers purpose developing an effective Web pages categorization framework base on ambiguity weighting focusing on the comparison of various term weighting schemes.

FEATURE EXTRACTION

Web pagess can be represented in various ways. Maybe the simplest way to represent a Web pages is to extract the text found within the BODY element. This representation does not exploit the peculiarities of Web pagess, i.e., HTML structure and the hypertextual nature of web pagess. By exploiting HTML structure for Web pages representation researchers can choose how a term is representative of the pages considering the HTML

element it is present in. For obtaining good performance in Web pages representation exploiting HTML structure is important to know where the more representative words can be found. For example, researchers can think that a word present in the TITLE element is generally more representative of the document's content than a word present in the BODY element. Researchers tested text sources for Web page representation, namely: BODY, the content of the BODY tag; META, the meta-description of the META tag; TITLE, the page's title.

The first step in Web pages categorization is to transform HTML element which typically are strings of characters into a representation suitable for the learning algorithm and the classification task. With English languages, a text string can easily be tokenized into terms by observing the word delimiting characters such as spaces, semicolons, commas, quotes and periods. To prepare a feature set for Web pages corpus, researchers must first apply a tokenize text strings into series of terms. Once a set of extracted words are obtained from the training news corpus, the removal of HTML tags, removal of stop-words and then word stemming. The stop-words are frequent words that carry no information, (i.e., pronouns, prepositions, conjunctions, etc.). By word stemming researchers mean the process of suffix removal to generate word stems. This is done to group words that have the same conceptual meaning such as walk, walker, walked and walking (Aas and Eikvil, 1999).

TERM WEIGHTING SCHEME

All Web pages are segmented into words or tokens that are inputs for next steps. Every Web pages document which is input is firstly transformed into a list of words obtained by selecting only those which are not present in a list of stopwords. Then, the words are matched against the term dictionary. Each entry in dictionary includes current text, term frequency, number of documents containing the term. In the Vector Space Model, documents are represented by vectors of words. Usually, one has a collection of documents which is represented by a word by word document matrix where each entry represents the occurrences of a word in a document (Aas and Eikvil, 1999).

In this study, researchers discuss different approaches for term weighting scheme. There are several ways of determining the weight w_{ik} of word i in document k, let tf_{ik} be the frequency of word i in document k, N the number of documents in the collection, n the total number of times word i occurs in the whole collection, log the base of logarithmic operation is 2. The following describes 7 different weighting schemes that are based on these quantities.

Boolean weighting: The simplest approach is to let the weight be 1 if the word occurs in the document and 0 otherwise:

$$w_{ik} = \begin{cases} 1 & \text{if } tf_{ik} > 0 \\ 0 & \text{otherwise} \end{cases}$$

Term frequency weighting (tf): The baseline method for computing the weight of a term in a document is to count the number of times the term occurs in the document:

$$w_{ik} = tf_{ik}$$

Tf x idf weighting: The most popular term weighting approach which has been widely used in information retrieval and has consequently been adopted by researchers in Web pages categorization. Assigns the weight to word i in document k in proportion to the number of occurrences of the word in the document and in inverse proportion to the number of documents in the collection for which the word occurs at least once (Salton and Buckley, 1988):

$$tfidf_{ik} = tf_{ik} \times \log\left(\frac{N}{n_i}\right)$$

Tfc-weighting: The tf x idf weighting does not take into account that documents may be of different lengths. The tfc-weighting is similar to the tf x idf weighting except for the fact that length normalisation is used as part of the word weighting equation (Salton and Buckley, 1988):

$$tfc_{ik} = \frac{tf_{ik} \times \log\left(\frac{N}{n_i}\right)}{\sqrt{\sum_{j=1}^{M}\left[tf_{ij} \times \log\left(\frac{N}{n_i}\right)\right]^2}}$$

Ltc-weighting: A slightly different approach uses the logarithm of the word frequency instead of the raw word frequency thus reducing the effects of large differences in frequencies (Buckley *et al*., 1994):

$$ltc_{ik} = \frac{\log\left(tf_{ik}+1\right) \times \log\left(\frac{N}{n_i}\right)}{\sqrt{\sum_{j=1}^{M}\left[\log\left(tf_{ij}+1\right) \times \log\left(\frac{N}{n_i}\right)\right]^2}}$$

Entropy weighting: Entropy-weighting (Dumais, 1991) is based on information theoretic ideas and is the most sophisticated weighting scheme. In the entropy weighting scheme, the weight for word i in document k is given by:

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

the average uncertainty or entropy of word i. This quantity is -1 if the word is equally distributed over all documents and 0 if the word occurs in only one document.

Ambiguity weighting: Ambiguity weighting (Chirawichitchai, 2010) is based on probabilistic theory ideas and term weighting ratio scheme. In the ambiguity weighting, the weight for word i in document k is given by:

$$\text{Ambiguity}_{ik} = 1 + \log(tf_{ik}) \times \log\left(1 + \frac{b^2}{a \times c}\right)$$

Where:

a = The number of documents belonging to category where the term does not occur
b = The number of documents belonging to category where the term occurs at least once
c = The number of documents not belonging to category where the term occurs at least once
d = The number of documents not belonging to category where the term does not occur

DIMENSIONALITY REDUCTION

A central problem in statistical Web pages categorization is the high dimensionality of the feature space, standard classification techniques cannot deal with such a large feature set since processing is extremely costly in computational terms and the results become unreliable due to the lack of sufficient training data. Feature selection attempts to remove non-informative words from documents in order to improve categorization effectiveness and reduce computational complexity (Aas and Eikvil, 1999).

Information Gain (IG) (Chirawichitchai *et al.*, 2009) (Yang and Pedersen, 1997) measures the number of bits of information obtained for category prediction by knowing the presence or absence of a word in at document. Let $c_1;...; c_K$ denote the set of possible categories. The information gain of a word w is defined to be:

$$IG(w) = -\sum_{j=1}^{K} P(c_j)\log P(c_j) + P(w)\sum_{j=1}^{K} P\left(\frac{c_j}{w}\right)\log P\left(\frac{c_j}{w}\right) + P(\bar{w})\sum_{j=1}^{K} P\left(\frac{c_j}{\bar{w}}\right)\log P\left(\frac{c_j}{\bar{w}}\right)$$

Here, $P(c_j)$ can be estimated from the fraction of documents in the total collection that belongs to class c_j and P(w) from the fraction of documents in which the word w occurs. Moreover, $P(c_j/w)$ can be computed as the fraction of documents from class c_j that have at least one occurrence of word w and $P(c_j/w)$ as the fraction of documents from class c_j that does not contain word w. The information gain is computed for each word of the training set and the words whose information gain is less than some predetermined threshold are removed.

CLASSIFICATION ALGORITHMS

The goal of categorization is to build a set of models that can correctly predict the class of the different objects. The input for these methods is a set of objects, the classes which these objects belong to (i.e., dependent variables) and a set of variables describing different characteristics of the objects (i.e., independent variables). Once such a Predictive Model is built, it can be used to predict the class of the objects for which class information is not known. This study gives a brief introduction to three well-known algorithms that are widely used for Web pages categorization, i.e., Naive Bayes, Support Vector Machine and Decision Tree.

Support Vector Machine (SVM) (Joachims, 1998): The SVM algorithm is based on the structure risk minimization principle. It has been shown in earlier researches to be effective for Web pages categorization. SVM divides the term space into hyperplanes or surface separating the positive and negative training samples. An advantage of SVM is that it can research well on very large feature spaces, both in terms of the correctness of the categorization results and the efficiency of training and categorization algorithm. However, a disadvantage of SVM training algorithm is that it is a time consuming process especially training with a large corpus.

Naive Bayes (NB) (Lewis, 1998): The NB algorithm has been widely used for Web pages categorization and shown to produce very good performance. The basic idea

is to use the joint probabilities of words and categories to estimate the probabilities of categories given a document. NB algorithm computes the posterior probability that the document belongs to different classes and assigns it to the class with the highest posterior probability. The posterior probability of class is computed using Bayes rule and the testing sample is assigned to the class with the highest posterior probability. The naive part of NB algorithm is the assumption of word independence that the conditional probability of a word given a category is assumed to be independent from the conditional probabilities of other words given that category.

Decision Tree (DT) (Quinlan, 1986): The DT algorithm is a common method used in data mining. The goal is to create a model that predicts the value of a target variable based on several input variables. Each interior node corresponds to one of the input variables; there are edges to children for each of the possible values of that input variable. Each leaf represents a value of the target variable given the values of the input variables represented by the path from the root to the leaf.

EXPERIMENT AND RESULTS

Researchers performed experiments using a collection of the WebKB corpus contains Web pages collected from computer science departments of 4 universities (Cornell, Texas, Washington and Wisconsin) (http://www.cs.cmu.edu/~webkb/). The 4159 Web pages collected were manually classified into 4 categories: student, faculty, course and project. Researchers used WEKA an open-source machine learning tool to perform the experiments.

Researchers used the default settings for all algorithms. For SVM, the default kernel function is Linear kernel. Classification effectiveness is usually measured using precision and recall. Precision is the proportion of truly positive examples labeled positive by the system that were truly positive and recall is the proportion of truly positive examples that were labeled positive by the system. The F-measure function which combines precision and recall is computed as (http://www.cs.waikato.ac.nz/ml/weka/):

$$\text{F-measure} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

Researchers tested all algorithms using the 10 fold cross validation. The results in terms of precision, recall and F-measure are the averaged values calculated across

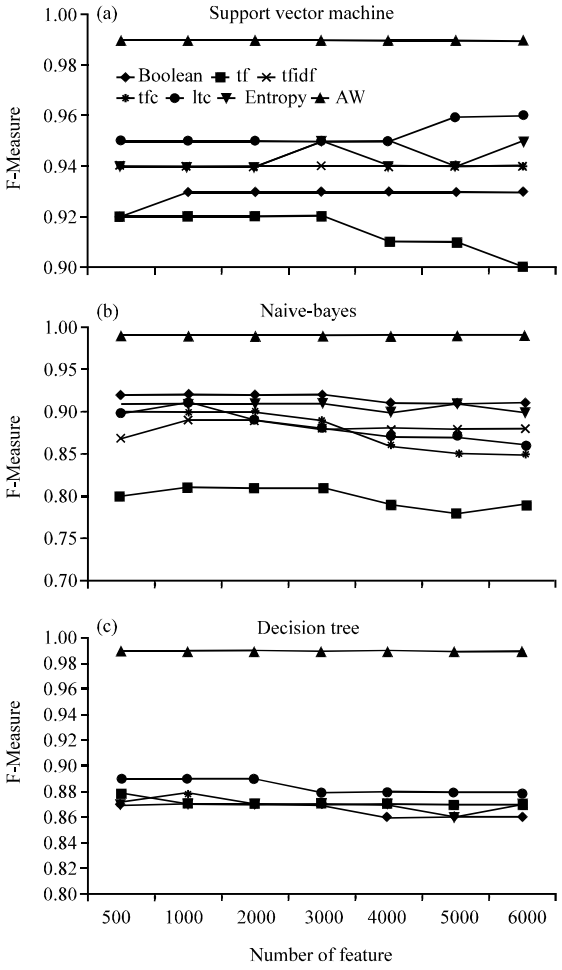

Fig. 1: Results of different weighting schemes on classification algorithm

all 10 fold cross validation experiments. The experimental results of these 7 term weighting scheme with respect to F-measure on Web pages corpus in combination with three learning algorithms are reported from Fig. 1.

Figure 1 shows the Web pages categorization on the information gain feature selection method results using three learning algorithms on WebKB corpus after feature weighting via Boolean, Tf, Tf-idf, Tfc, Ltc, entropy and ambiguity weighting, respectively. Three observations from the Web pages categorization results follows. First, performance of the different term weighting schemes with a small feature size can not be summarized in one sentence but the trends are distinctive that the F-measure points of different term weighting schemes increase as the number of the features grows. Second, ambiguity weighting is more effective than another weighting. Finally, the best F-measure points on ambiguity weighting with SVM, NB and DT were 99% based on 10 fold cross validation.

CONCLUSION

In this study, researchers proposed developing an effective Automatic Web pages categorization base on ambiguity weighting focusing on the comparison of various term weighting schemes. Researchers found ambiguity weighting most effective in the experiments with SVM, NB and DT algorithms. Researchers also discovered that the ambiguity weighting is suitable for combination with the information gain feature selection method. The ambiguity weighting with classification algorithms yielded the best performance with the F-measure over all algorithms. Based on the experiments, the classification algorithm with the information gain feature selection yielded the best performance with the F-measure of 100%. The experimental results also reveal that feature weighting methods have a positive effect on Web pages categorization.

ACKNOWLEDGEMENTS

Researchers would like to thank Sripatum University Chonburi Campus for scholarship support. The research is not possible without the data from World Wide Knowledge Base project of the CMU text learning group.

REFERENCES

Aas, K. and L. Eikvil, 1999. Text categorization: A survey. Norwegian Computing Center.

Buckley, C., G. Salton, J. Allan and A. Singhal, 1994. Automatic query expansion using SMART:TREC 3. Proceedings of the 3rd Text Retrieval Conference, November, 1994, National Institute of Standards, Gaithersburg, MD., pp: 69-80.

Chirawichitchai, N., 2010. Thai document categorization using ambiguity ratio term weighting technique. Ph.D. Thesis, King Mongkut's University of Technology, North Bangkok.

Chirawichitchai, N., P.S. Nguansat and P.Meesad, 2009. An experimental study on feature reduction techniques and classification algorithms of Thai documents. J. Sci. Ladkrabang, Vol. 18.

Dumais, S.T., 1991. Improving the retrieval information from external sources. Bahaviour Res. Methods Instrum. Comp., 23: 229-236.

Joachims, T., 1998. Text categorization with support vector machines: Learning with many relevant features. Proceedings of the 10th European Conference on Machine Learning, Chemnitz, Germany, April 21-23, 1998, Springer, Berlin, Heidelberg, pp: 137-142.

Lewis, D.D., 1998. Naive (Bayes) at forty: The independence assumption in information retrieval. Proceedings of the 10th European Conference on Machine Learning Chemnitz, Germany, April 21-23, 1998, Springer Berlin, Heidelberg, pp: 4-15.

Quinlan, J.R., 1986. Induction of decision trees. Mach. Learn., 1: 81-106.

Salton, G. and C. Buckley, 1988. Term-weighting approaches in automatic text retrieval. Inform. Process. Manage., 24: 513-523.

Salton, G. and M.E. Lesk, 1968. Computer evaluation of indexing and text processing. J. ACM., 15: 8-36.

Yang, Y.M. and J. Pedersen, 1997. A comparative study on feature selection in text categorization. Prceedings of the 14th International Conference on Machine Learning (ICML'97), Morgan Kaufmann, pp: 412-420.