Available online at www.sciencedirect.com
ScienceDirect
Expert Systems with Applications 36 (2009) 260–272
Expert Systems with Applications
www.elsevier.com/locate/eswa

Two novel feature selection approaches for web page classification

Chih-Ming Chen [a,*], Hahn-Ming Lee [b], Yu-Jung Chang [c]

[a] Graduate Institute of Library, Information and Archival Studies, National Chengchi University, No. 64, Sec. 2, ZhiNan Road, Wenshan District, Taipei 116, Taiwan, ROC
[b] Department of Computer Science and Information Engineering, National Taiwan University of Science and Technology, Taipei, Taiwan, ROC
[c] Department of Computer Science and Information Engineering, National Taiwan University, Taipei, Taiwan, ROC

Abstract

To help the growing qualitative and quantitative demands for information from the WWW, efficient automatic Web page classifiers are urgently needed. However, a classifier applied to the WWW faces a huge-scale dimensionality problem since it must handle millions of Web pages, tens of thousands of features, and hundreds of categories. When it comes to practical implementation, reducing the dimensionality is a critically important challenge. In this paper, we propose a *fuzzy ranking analysis* paradigm together with a novel relevance measure, *discriminating power measure* (DPM), to effectively reduce the input dimensionality from tens of thousands to a few hundred with zero rejection rate and small decrease in accuracy. The two-level promotion method based on fuzzy ranking analysis is proposed to improve the behavior of each relevance measure and combine those measures to produce a better evaluation of features. Additionally, the DPM measure has low computation cost and emphasizes on both positive and negative discriminating features. Also, it emphasizes classification in parallel order, rather than classification in serial order. In our experimental results, the *fuzzy ranking analysis* is useful for validating the uncertain behavior of each relevance measure. Moreover, the DPM reduces input dimensionality from 10,427 to 200 with zero rejection rate and with less than 5% decline (from 84.5% to 80.4%) in the test accuracy. Furthermore, to consider the impacts on classification accuracy for the proposed DPM, the experimental results of China Time and Reuter-21578 datasets have demonstrated that the DPM provides major benefit to promote document classification accuracy rate. The results also show that the DPM indeed can reduce both redundancy and noise features to set up a better classifier.

Keywords: Feature selection; Fuzzy decision making; Web page classification; Discriminating power measure

1. Introduction

With the growing popularity of the World Wide Web, and the maturity and availability of related tools and techniques (like Web Servers, browsers, visual tools for Web page makers, dynamic HTML, Web-based databases and so on), more and more heterogeneous information is being "published" and added to the Web. The explosive growth in the number of Web pages has, in turn, contributed to the popularity of search tools. However, those search tools suffer from some problems. Search robots (like Openfind (Openfind), AltaVista (AltaVista)) often make users feel lost in *irrelevant search results*. Search tools based on manually maintained classified directories (like Yam (Yam), Yahoo! (Yahoo)) provide high-quality results but are hampered by low *production rates*. Since classification does improve search results but is time-consuming when done manually, *automatic* Web page classification should be considered to remedy the information-overloading problem.

An Automatic Web Page Classifier (AWPC) not only can relieve the slowness of manual classification, but could also guide the users of search tools through the various kinds of ambiguity by providing a list of topic paths. In order to achieve high-quality classification performance, both selection of effective features and selection of a

* Corresponding author. Tel.: +886 2 29393091x88024; fax: +886 2 29384704.
E-mail address: chencm@nccu.edu.tw (C.-M. Chen).

0957-4174/$ - see front matter © 2007 Elsevier Ltd. All rights reserved.
doi:10.1016/j.eswa.2007.09.008

classifier that can make good use of those features with limited training data, memory, and computing power are essential (Lippmann, 1989). In (Lippmann, 1987, 1989; Zurada, 1992; Nadler & Smith, 1993; Holmstom, 1997; Joshi, 1997), numerous pattern classification techniques (including statistical pattern recognition, neural networks, machine learning, neuro-biological, and neuro-fuzzy) are introduced, classified and compared. Importantly, many scholars were conscious of the subject of applying pattern classification techniques to Web page classification, and thus a growing number of classification models and machine-learning techniques have been applied to Web page classification in recent years, including multivariate regression models (Yang & Chute, 1994), nearest neighbor classification (Yang, 1994), Bayesian probabilistic approaches (Friedman, Geiger, & Goldszmidt, 1997), decision trees, neural networks (Musavi, Ahmed, Chan, Faris, & Hummels, 1992), symbolic rule learning (Cohen William & Singer, 1996), and inductive learning algorithms (Lewis, Schapire, Callan, & Papka, 1996). Moreover, lots of techniques were proposed to focus on the subject of Chinese Web page classification, such as the linear-based classifier (Chen, Liu, & Lee, 2001) (i.e. vector space model (VSM) classifier), neural network models (Chen, Lee, & Hwang, 2005), fuzzy theory (Yang & Hou, 1998), and so on. Besides, to improve Web page classification techniques, a novel proposed self-organizing HCMAC neural network classifier (Chen, 2003; Lee, Chen, & Lu, 2003) has been demonstrated its good performance for Web page classification. Actually, among the various kinds of classifiers, determining which ones are more appropriate for Web page classification is difficult and complex job.

In addition, when constructing an automatic Web page classifier, one still has to deal with the problem of huge-scale datasets. Namely, the AWPC must handle millions of Web pages (huge amount of instances), tens of thousands of features (extremely high input dimensionality), and hundreds or thousands of categories (high output dimensionality). Unfortunately, this situation is getting worse due to the explosive growth of Web pages. Consequently, effective feature selection mechanisms are critically important. Moreover, the VSM model (Chen et al., 2001; Salton, 1983, 1989) is simple and generally used for automatic document classification. To demonstrate the performance of the proposed feature selection approach for Web page classification, we focus on the combination of VSM classifier with the proposed feature selection method because an effective feature selection approach can generally promote classification accuracy rate for any classification models. To determine an appropriate criterion for feature selection, we emphasize that the given threshold value for extracting the informative feature terms has practically the uncertainty behavior. Therefore, we propose a *fuzzy ranking analysis* paradigm, which consists of *ranking analysis* steps to analyze and evaluate the uncertain behavior. Additionally, we also propose a *two-level promotion* technique to promote the performance of existing relevance measures, and present a novel relevance measure, named *discriminating power measure* (DPM), to obtain higher quality feature terms for document classification.

According to our experimental results, the *fuzzy ranking analysis* is useful for validating the uncertain behavior of each relevance measure. The *two-level promotion* techniques, which work under the restriction that relevance measures are limited and the perfect relevance measure is difficult to acquire by available sensing techniques, can show the trade-off between the rejection rate and the accuracy rate. Also, the experimental results for the DPM are very encouraging. The DPM greatly reduces input dimensionality, with zero rejection rate, while maintaining high classification accuracy. The DPM can reduce both redundancy and noise features.

2. Feature selection

In pattern classification, the so-called feature engineering process can be divided into three stages: feature generation stage, feature refinement stage, and feature utilization stage. In the feature generation stage, candidate features (i.e., the original feature set) are generated by pre-determined kinds of sensing techniques from the training set. For greater efficiency and even accuracy, the original feature set can be refined by feature selection and/or feature extraction. In feature selection, it is assumed that there are sufficient relevant features in the original feature set to discriminate clearly between categories, and that some irrelevant features can be eliminated to improve efficiency and even accuracy. For instance, elimination of redundancy will improve efficiency without losing accuracy, and elimination of noise will improve both efficiency and accuracy. In the feature extraction approach, it is supposed that the features in the original set are not all appropriate; nevertheless, sufficient information for classification is already captured by them. Feature extraction, which generates new features and measurements from the original features and measurements, is designed to handle this kind of situation. Examples include feature clustering, which may be the simplest way to implement feature extraction, and factor analysis by latent semantic indexing using singular value decomposition (Deerwester, 1990). The best way to test whether a representation is useful or not is simply to utilize it. In the feature utilization stage, features in the refined set are first used to represent each instance in the dataset. Then, an appropriate classification model is selected to make good use of these features.

2.1. Consideration of feature set quality

In pattern classification applications, when accuracy and/or efficiency are unacceptable, one tries to find the possible reasons and solve the problems. In Lewis (1992), Lewis enumerates six situations where feature set are of poor-quality so that obtaining a useful classifier is difficult or impossible:

(1) The feature set does not sufficiently distinguish instances.
For example, if the feature set only contains a single feature with binary values and there are three distinct classes, then it is impossible to find a classifier in this situation.
(2) The feature set excludes concepts from the hypothesis space.
The learning system might search a space of linear classifiers, but there may be no hyperplane that separates the positive from the negative instances in the space defined by the features.
(3) The feature set results in a "too big" hypothesis space.
The high input-dimensionality problem makes learning a classifier slow and difficult. This problem is especially serious in Web page classification due to the huge-scale dataset.
(4) The feature set violates explicit or implicit assumptions of the learning algorithm.
Even though the feature set includes concepts from the hypothesis space (i.e., there is a solution in the hypothesis space), learning a classifier may be impossible because the feature set violates assumptions of the learning algorithm (i.e., the feature set makes the search for solutions difficult or impossible). For example, the feature set may lead to a local minimal in neural networks.
(5) The feature set is noisy.
The definition of noise is well defined in statistical communication theory (Hamming, 1980), but much less well defined for machine-learning applications. Whatever the definition, noise leads a learning algorithm to derive a different, usually poorer classifier that it would be from noise-free data, and will reduce the accuracy of classifiers when applied to new instances.
(6) The feature set may contain redundancy.
Two different features may actually be measurements of the same property of an instance. Hence, we say these two features are alternative and we should keep one and eliminate the other.

From the description above, we see that feature selection is one way to improve the pattern representation but not the only way. It can ease or solve some of the above-mentioned problems of poor-quality feature sets; still, there are some restrictions or side effects.

2.2. The behavior of relevance measures

In practical implementation of feature selection, the design of relevance measures is crucial. With relevance measures, we can distinguish relevant features from irrelevant ones. A *relevance measure* is a feature evaluation function such that each feature has a quantitative description, which is usually a real value, to indicate whether the feature is relevant or not. In this paper, we combine some relevance measures and also propose a new relevance measure named *discriminating power measure*, described in Section 4, to help the system to detect noise and redundancy. The behavior of relevance measures can be divided into two categories according to the explicit or implicit viewpoints of feature selection (Littlestone, 1998; Kira & Rendell, 1992; Almuallim & Dietterich, 1991; Quinlan, 1983, 1993; Langley & Sage, 1994). The explicit viewpoint of feature selection is shown in Fig. 1. There exists a perfect relevance measure along with a perfect threshold such that relevant features and irrelevant ones can be divided distinctly.

The implicit viewpoint of feature selection reveals the uncertain behavior of relevance measures. In practice, it is not possible to design the perfect relevance measures with perfect thresholds shown in Fig. 1, especially when the sources of noise and redundancy are unknown or uncertain. Instead, almost perfect measures with fuzzy thresholds (shown in Fig. 2a) are more practical. Even if there exists only some useful but clearly imperfect measures with fuzzy thresholds (shown in Fig. 2b), feature selection can be achieved by combining those useful measures.

2.3. Useful relevance measures in text categorization

In Table 1, we list several commonly used relevance measures (Salton, 1989; Yang, 1993; Huang, 1997) in text

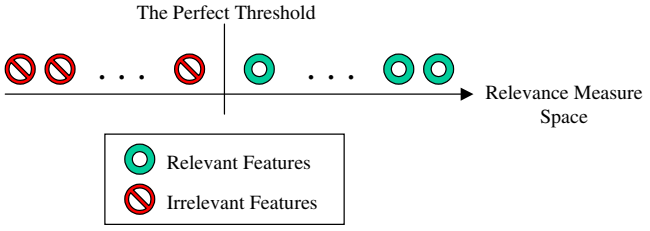


Fig. 1. The explicit viewpoint of feature selection.

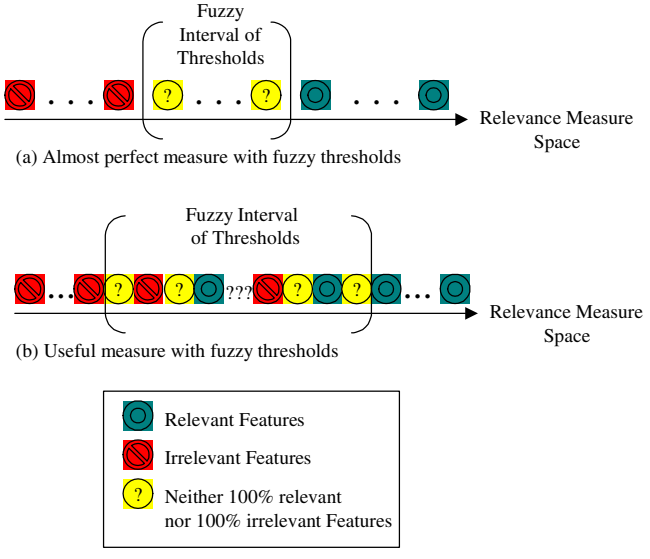


Fig. 2. The implicit viewpoint of feature selection.

Table 1
List of relevance measures in text categorization

Name	Definitions
Term occurrence number	• In a document: $n(\text{term}_k, \text{doc}_j)$: the occurrence number of term_k in doc_j • In a category: $n(\text{term}_k, \text{cat}_i)$: the occurrence number of term_k in cat_i • In a whole collection: $n(\text{term}_k, \text{collection})$: the number of occurrence of term_k in the whole collection
Term frequency	• In a document: $\text{TF} = n(\text{term}_k, \text{doc}_j)/n(\text{term}_{\text{all}}, \text{doc}_j)$: the frequency of term_k in doc_j • In a category: $n(\text{term}_k, \text{cat}_i)/n(\text{term}_{\text{all}}, \text{cat}_i)$: the frequency of term_k in cat_i
Document frequency (Salton, 1989)	• In a category: $\text{DF} = n(\text{doc}(\text{term}_k)_{\text{all}}, \text{cat}_i)/n(\text{doc}_{\text{all}}, \text{cat}_i)$: the frequency of documents containing term_k in cat_i • Inverse Document Frequency (IDF): $\text{IDF} = \log(1/\text{DF})$
Conformity (Yang, 1993, 1997)	• Inverse Cluster Frequency (ICF): $icf_k = -\sum_i P_{ki} \log P_{ki}$, where $P_{ki} = \frac{n(\text{doc}(\text{term}_k)_{\text{all}}, \text{cat}_i)}{\sum_i n(\text{doc}(\text{term}_k)_{\text{all}}, \text{cat}_i)}$
Uniformity (Huang, 1997)	$H_k = -\sum_j P_{kj} \log P_{kj}$, where $P_{kj} = \frac{n(\text{term}_k, \text{doc}_j)}{\sum_j n(\text{term}_k, \text{doc}_j)}$

categorization and state their definitions. The notation used in Table 1 is listed as follows:

term_k: the kth term; doc_j: the jth document; cat_i: the ith category; term_{all}: all terms; doc_{all}: all documents; cat_{all}: all categories; doc $(\text{term}_k)_{\text{all}}$: all documents that contain term_k; $n(A, B)$: the number of A in B.

All of the frequency measures in Table 1 (such as term frequency and document frequency) are based on the assumption that the more frequently potential feature is used, the more essential or emphasized it is. In addition to frequency, one must also consider discrimination. Here, both the conformity measure (Yang, 1993; Huang, 1997) and the uniformity measure (Huang, 1997) use the entropy concept to indicate the quality of feature distribution. The conformity measure prefers features whose distributions are centralized in certain categories. The uniformity measure factors features with flat distributions across documents of a specific category.

3. Feature selection by fuzzy ranking analysis

In text categorization or Web page classification, the original input dimensions (i.e., the number of original features) are often in the tens of thousands because every unique term, which is a basic element with complete semantic meaning (e.g., words in English), can be a feature. Even after feature selection, the number of features is typically still more than one thousand (Lewis, 1992; Blum & Langley, 1997). Such high input dimensionality makes some powerful machine-learning algorithms (like neural networks (Lippmann, 1989; Lippmann, 1987; Zurada, 1992; Nadler & Smith, 1993; Holmstom, 1997)), which can solve both linear-separable and nonlinear-separable distributions and have good generalization ability, computationally overloaded. Thus, we want to reduce the input complexity such that those complexity-sensitive learning algorithms can be applied to Web page classification.

3.1. The scalability analyses of input features

In order to achieve this goal of reducing input features, the *scalability analyses of input features* are needed because these results can help us to determine appropriate feature dimensions. Restated, we can *use as few features as possible to achieve almost the same classification accuracy*, and that is what we emphasize in this paper. Fig. 3 shows a likely analytic example of scalable feature selection under different scenarios. The original feature set may contain noise and redundancy. Redundancy decreases the classification efficiency. Noise decreases not only the classification efficiency, but also the classification accuracy. Generally, if noise and redundancy features can be successfully removed, then the curve of a typical case will be shifted closer to the curve of the best case as shown in Fig. 3. Namely, a good feature selection method should shift the curve of classification accuracy rates as close to the best case as possible.

3.2. Two-level promotion based on fuzzy ranking analysis

In everyday life, the need to select the relevant items from a set of alternatives arises frequently. One common strategy is to define a perfect relevance measure accompanied with a perfect threshold (shown in Fig. 1) to distinctly divide the alternatives into two groups (relevant set and irrelevant set). However, it is not easy to obtain the perfect case in real applications. In fact, partial relevance measures (shown in Fig. 2), which are not perfect and only focus on special relevance, are easier to obtain. Therefore, aggregating a set of partial relevance measures to approximate the perfect relevance measure is more practical.

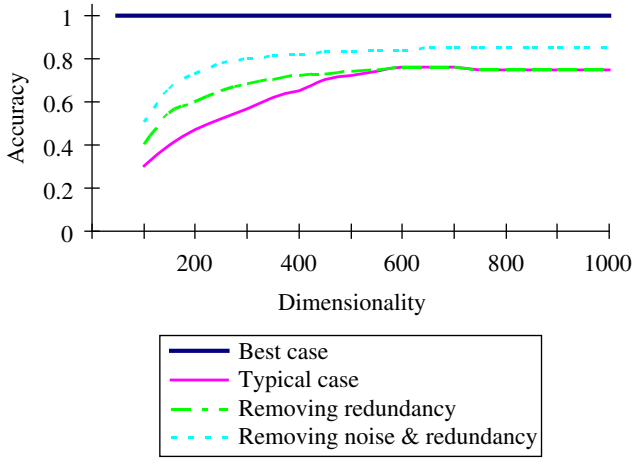


Fig. 3. Scalable feature selection under different scenarios.

Because the partial relevance measures are not globally perfect, i. e. their behavior is uncertain, so we need some promotion methods to enhance the selection activity. We propose the *fuzzy ranking analysis* paradigm, which consists of ranking analysis steps and a two-level promotion technique. When the behavior of relevance measures is uncertain, the feature selection activity can be modeled as multi-criteria decision making in a fuzzy environment. Features can be viewed as a set of alternatives. Thus, relevance measures are viewed as a set of criteria and judgments given on each alternative. The global criterion can then be formed by aggregating all criteria and a ranking of all alternatives is induced. Hence, we can choose a number of the top alternatives to be the selected subset, with the number determined by the trade-off between accuracy and efficiency. We list the promotion steps as follows:

Step 1. *Intra-level promotion*: Intra-level promotion aims at transforming a non-monotonic relevance measure into a monotonically increasing or decreasing (or as monotonic as possible) and normalized measure.

Step 2. *Inter-level promotion*: The remaining relevance measures are aggregated by aggregation operators, and a ranking of all alternatives is induced to promote relevance measures for feature selection.

3.2.1. Intra-level promotion step

A relevance measure with monotonically increasing or decreasing property represents that a candidate feature with a larger or smaller value of this measure is a better feature, which can be selected as feature for classification. Actually, a relevance measure with the property of monotonically increasing or decreasing is easier to determine the threshold of relevance measure than the non-monotonic relevance measures for feature selection. For example, the document frequency (DF) measure is a monotonically increasing measure and the inverse cluster frequency (ICF) measure is monotonically decreasing measure, but the term frequency measure widely used in information retrieval is not either monotonically increasing or decreasing measure. The proposed intra-level promotion can transform the term frequency as a monotonically increasing measure. In this step, we use fuzzy sets to model the uncertain behavior of relevance measures and make each relevance measure be as monotonically increasing or decreasing as possible. Two sub-steps are included as follows:

(1) Define a proper fuzzy set.
(2) Use the fuzzy set to transform the original measure into a promoted measure.

For each relevance measure, we define a proper fuzzy set, which is a function from the original relevance measure space to the membership space $[0,1]$ such that a larger value means a higher degree of relevance, transforming the original measure into a monotonically increasing or monotonically decreasing one. The proper fuzzy set can be defined according to the various distributions of candidate terms. Restated, the definition of proper fuzzy set is actually a heuristic work, and a prior statistical analysis of term's distribution is helpful for determining a proper fuzzy set. Of course, some optimization algorithms can help us to determine the proper fuzzy set for intra-level promotion, such as genetic algorithm, etc. However, this work is a time-consuming job for the proposed intra-level promotion approach. Generally, the term frequency measure as mentioned in Section 2.3, medium-frequency features are usually more relevant than low-frequency features or high-frequency ones. Also, the most of term frequency values are less than 0.5 in the China-Times news dataset. Thus, we define the fuzzy set, like the one in Fig. 4, to transform the term-frequency measure into monotonically increasing one by mapping the term frequency of each term into a fuzzy degree for the feature selection of China-Times news dataset.

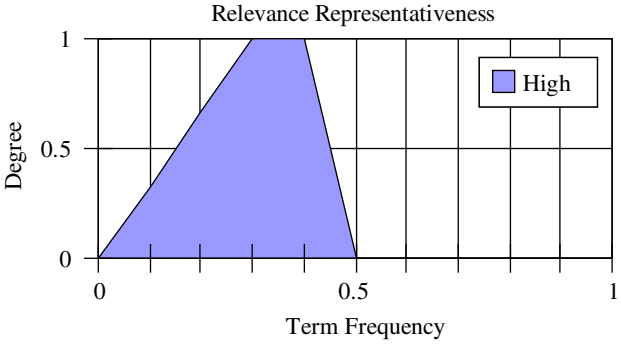


Fig. 4. An example of intra-level promotion (term-frequency measure).

3.2.2. Inter-level promotion step

This step uses multiple attribute decision making (MADM) (Zimmermann, 1987; Fodor, Marichal, & Roubens, 1995; Ribeiro, 1996) to model the feature selection process. The original feature set can be viewed as a set of alternatives. The relevance measures form the set of criteria and produce scores for each alternative. An aggregation operator determines the way to combine those scores into a total score for each alternative. In this paper, two aggregation operators, a minimum operator and a weighted average operator, are used. Two sub-steps are included in inter-level promotion:

(1) Obtain all scores.
Let $A = \{a_1, a_2, \ldots, a_n\}$ be a set of features and $C = \{c_1, c_2, \ldots, c_m\}$ be a set of relevance measures characterizing a decision situation. Then, the basic information involved in MADM can be expressed by the matrix **R** as follows:

$$\mathbf{R} = \begin{array}{c} \\ c_1 \\ c_2 \\ \vdots \\ c_m \end{array} \begin{array}{c} \begin{array}{cccc} a_1 & a_2 & \cdots & a_n \end{array} \\ \begin{bmatrix} r_{11} & r_{12} & \cdots & r_{1n} \\ r_{21} & r_{22} & \cdots & r_{2n} \\ \vdots & \vdots & \vdots & \vdots \\ r_{m1} & r_{m2} & \cdots & r_{mn} \end{bmatrix}_{m\times n} \end{array} \tag{1}$$

Because in the intra-level step, each relevance measure is fuzzified, all entries of this matrix are real numbers in [0,1]. Each entry r_{ij} expresses the relevance degree of a feature a_j scored by relevance measure c_i ($i \in N_m$, $j \in N_n$).

(2) Aggregate all scores into a total score for each alternative.

Two aggregation operators are used.

- Minimum operator:

$$r_j = h(r_{1j}, r_{2j}, \ldots, r_{mj}) = \min_i r_{ij} \quad (i \in N_m, j \in N_n) \tag{2}$$

- Weighted-average operator:

$$r_j = h(r_{1j}^{w_1}, r_{2j}^{w_2}, \ldots, r_{mj}^{w_m}) = \frac{\sum_{i=1}^{m} w_i r_{ij}}{\sum_{i=1}^{m} w_i} \quad (j \in N_n) \tag{3}$$

where $w_1, w_2, \ldots, w_m$ are weights that indicate the relative importance of relevance measures $c_1, c_2, \ldots, c_m$.

In summary, we infer that two factors may affect the result of intra-level promotion, but these two factors may not be completed. First, the definitions of the membership functions may not be proper. If the fuzzy sets used are not appropriate, the effect of intra-level promotion is not obvious for Web page classification. However, to determine the definitions of the membership functions currently is a heuristic work and is sensitive by different datasets because it is highly relevant with the data distribution of datasets. Second, the promotion performance may not be obvious if the quality of the relevance measure is poor. Experiments given later show that several tested relevance measures has slightly promotion of accuracy for Web page classification problem after using the intra-level promotion procedure for feature selection. To promote the classification accuracy rate more obviously, this paper tries to propose a more powerful relevance measure, i.e. *discriminating power measure* (DPM), to solve this problem. Furthermore, later experiments show that inter-level promotion can obviously reduce the rejection rates for Web page classification in training and testing phases.

4. A novel relevance measure: discriminating power measure

Obviously, when the quality of all the relevance measures is poor, the promotion from existing relevance measures will be limited. To relieve this situation, this section proposes a new relevance measure (called *discriminating power measure*, *DPM*) for the purpose of promoting feature selection.

4.1. Algorithm of discriminating power measure

The goal of the DPM measure is to select features that reveal larger differences among categories. To achieve this aim, two steps are applied:

Step 1. Consider each feature for each category, calculate the difference, in term of document frequency, between that category and others for that feature.

The document frequency (DF) inside a certain category and the DF outside the category are first calculated. Then, we can get the absolute difference between the inside DF and the outside DF as follows:

- Doc frequency inside the ith category:

$$\mathrm{DF}_i^{\mathrm{in}} = \frac{n(\mathrm{doc}(\mathrm{term}_k)_{\mathrm{all}}, \mathrm{cat}_i)}{n(\mathrm{doc}_{\mathrm{all}}, \mathrm{cat}_i)} \tag{4}$$

- Doc Frequency outside the ith category:

$$\mathrm{DF}_i^{\mathrm{out}} = \frac{n(\mathrm{doc}(\mathrm{term}_k)_{\mathrm{all}}, \mathrm{collection} - \mathrm{cat}_i)}{n(\mathrm{doc}_{\mathrm{all}}, \mathrm{collection} - \mathrm{cat}_i)} \tag{5}$$

where term_k is the kth term; cat_i is the ith category; $\mathrm{collection} - \mathrm{cat}_i$ are all documents in the whole collection except cat_i; $\mathrm{doc}(\mathrm{term}_k)_{\mathrm{all}}$ are all documents that contain term_k; $\mathrm{doc}_{\mathrm{all}}$ are all the documents; B is a variable that represents cat_i, $\mathrm{collection} - \mathrm{cat}_i$ or others; $n(\mathrm{doc}(\mathrm{term}_k)_{\mathrm{all}}, B)$ is the number of documents with term_k in B; $n(\mathrm{doc}_{\mathrm{all}}, B)$ is the total number of documents in B.

- The difference for cat_i:

$$\delta_i = |\mathrm{DF}_i^{\mathrm{in}} - \mathrm{DF}_i^{\mathrm{out}}| \tag{6}$$

Step 2. For the whole collection, sum up the total difference due to a feature.

- Discriminating power measure:

$$\mathrm{DPM} = \sum_i \delta_i \tag{7}$$

In this paper, we use our proposed DPM as a feature selection mechanism and test its performance of promoting feature selection. If the DPM values of some terms are beyond the given threshold DPM value, these terms are selected because they are better able to distinguish between different categories.

4.2. Characteristics of discriminating power measure

In order to compare the difference of the discriminating power measure with the other relevance measures for fea-

Table 2
A dataset of document classification

Category/Number of documents	Cat_A/ 10	Cat_B/ 20	Cat_C/ 30	Cat_D/ 40

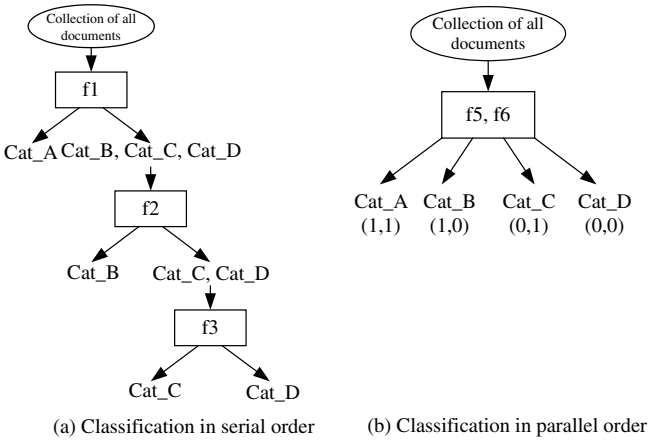
ture selection, we give an example as listed in Table 2 to show the proposed DPM measure which can select the feature terms with parallel order for document classification. This property is different from the other relevance measures, such as term frequency (TF) and inverse cluster frequency (ICF), to select the feature terms with serial order for document classification. Assume we have a dataset with one hundred of documents as listed in Table 2 for document classification, which belongs to four different categories, respectively. Moreover, assume there are totally seven candidate feature terms extracted from all documents. Table 3 illustrates the corresponding scores of seven candidate feature terms using three different relevance measures, respectively. To select most informative feature terms for document classification, Table 4 exhibits the ranking order of these candidate feature terms according to the scores of different relevance measures. We find that ICF measure must select the top three features to distinguish four categories because the selected feature terms give serial order decision regions for document classification. However, our DPM measure only needs to select the top two features to distinguish four categories due to the selected feature terms give parallel order decision regions for document classification. Namely, our proposed DPM measure can achieve high classification accuracy rate even using as few feature terms as possible. The concepts of serial and parallel order decisions in document classification are shown as Figs. 5a and b, respectively.

Finally, we summarize the characteristics of the DPM as follows:

(1) It has low computation cost. From Eqs. (4)–(7), we can find that the DPM uses the document frequency to obtain all useful information for feature selection. In contrast with the term frequency, document frequency has lower computation cost since document frequency does not need to compute the occurrence times of each feature term.

Table 4
The ranking order of seven candidate feature terms according to the scores of different relevance measures for feature selection

ICF	f1	f2	f3	f4	f6	f5	f7
DPM	f5	f6	f4	f7	f3	f2	f1



(a) Classification in serial order (b) Classification in parallel order

Fig. 5. Two different aspects of document classification according to the selected feature terms.

(2) It emphasizes both positive discriminating features and negative discriminating features.
In order to select the most informative feature terms for document classification, the DPM considers simultaneously the document frequency inside some category and outside some category in feature selection process. The measure of document frequency inside some category emphasizes positive discriminating features, and the measure of document outside some category emphasizes negative discriminating features. To consider both measures simultaneously will contribute more discriminating information for feature selection.
(3) It emphasizes classification in parallel order, as opposed to serial order.

From the above analyses, we find that most traditional relevance measures extract the feature terms with serial order decision like decision tree. However, our proposed DPM extracts the feature terms with parallel order decision.

Table 3
The scores of candidate feature terms using different relevance measures

Candidate feature terms	Evaluated items					
	Document frequency in Cat_A	Document frequency in Cat_B	Document frequency in Cat_C	Document frequency in Cat_D	ICF	DPM
f1	10	0	0	0	0	1.435
f2	0	20	0	0	0	1.841
f3	0	0	30	0	0	2.208
f4	0	0	0	40	0	2.516
f5	10	20	0	0	0.276	2.724
f6	10	0	30	0	0.224	2.69
f7	10	20	30	0	0.439	2.516

5. Experimental results

To demonstrate the performance of the proposed methods, the proposed feature selection approach is applied to identify informative feature terms for document classification and the similarity-based classifier, i.e. vector space model (VSM) classifier is utilized for document classification. We collect two topic structures, the China-Times and Reuters, as our datasets. The two datasets are described in Section 5.1. The proposed methods are examined to show if they can efficiently promote the classification accuracy rate of VSM classifier. Besides, the accuracy rate (the rate at which documents are classified accurately) and the rejection rate (the proportion of unrecognized documents) are used to analyze and compare the classification results for different relevance measures and our two-level promotion method.

5.1. Datasets

The major concern of this paper is the scalability of feature selection for automatic Web page classifiers. To test our proposed methods, we wrote a spider program in advance to automatically collect daily news from the China-Times Web site (http://news.chinatimes.com/). This site publishes about one hundred news articles in HTML format every day and divides the news into eight categories as shown in Table 5. We collected the daily news from 1998/12/29 to 1999/01/07 (a total of 10 days, about 1000 articles). Then we split the dataset into two parts: the training set and the test set. We use the first five days (1998/12/29–1999/01/02, a total of 511 articles) as the training set and the other five days (1999/01/03–1999/01/07, a total of 485 articles) as the test set.

Furthermore, Reuters dataset is from Reuters-21578 text categorization test collection Distribution 1.0 (Lewis, 1999). We select documents with single matched category and spit documents into training and testing sets according to the modified Lewis split. The training and testing datasets thus consist of 6552 and 2581 documents, respectively, and all the documents can be uniquely labeled by 59 categories.

Table 5
Categories in the China-Times dataset

Directory	Category Name
\cchina	兩岸三地 (News about Taiwan, Hong Kong and Mainland China)
\ceconomy	財經 (Economic News)
\cent	藝文 (Entertainment News)
\cintl	國際 (International News)
\cpolitic	政治 (Political News)
\csocial	社會綜合 (Social News)
\csports	體育 (Sports News)
\cstar	影劇 (Star News)

5.2. Experimental results of feature selection

In this work, we use the China-Times dataset to demonstrate the performance of the proposed feature selection approach. Two common relevance measures: term frequency (TF) and inverse cluster frequency (ICF) (shown and defined in Section 2.3) are applied to test our two-level promotion method. To focus on scalability analysis, each test shows only the results of the first one thousand feature rankings. In what follows, we describe the results of scalability tests shown in Figs. 6–13. For both training and testing, four kinds of rankings using different relevance measures or aggregation operations are tested and the accuracy rate versus dimensionality relationships are plotted. The eight experiments and their corresponding figures are shown in Figs. 6–13.

Figs. 6 and 7 compare intra-level promotion results for term frequency (TF) and promoted term frequency (PTF) in the training phase and testing phase, respectively. The figures show that the effect of intra-level promotion is only slightly positive. Two factors may contribute to this result. First, the definition of the fuzzy set function may be

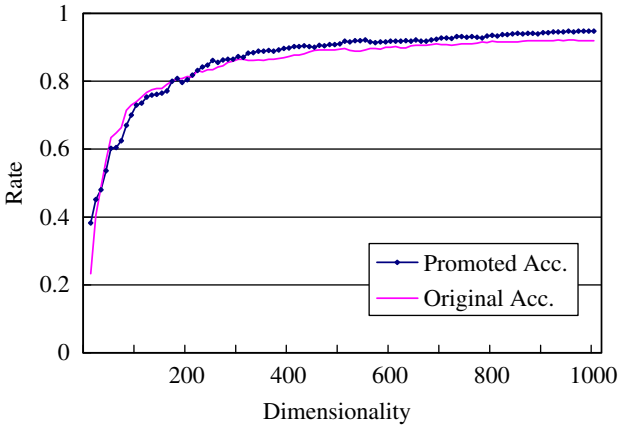


Fig. 6. An example of intra-level promotion (TF and PTF) in the training phase.

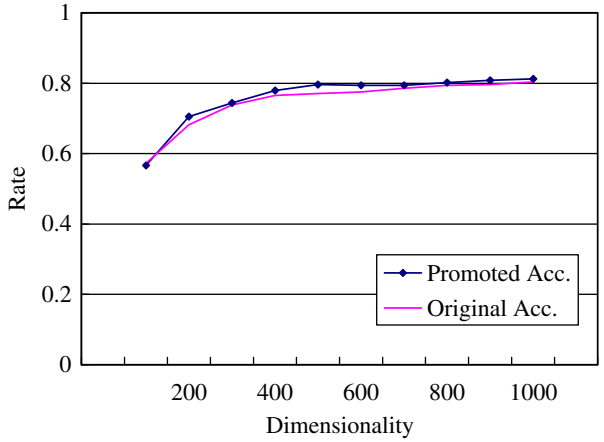


Fig. 7. An example of intra-level promotion (TF and PTF) in the test phase.

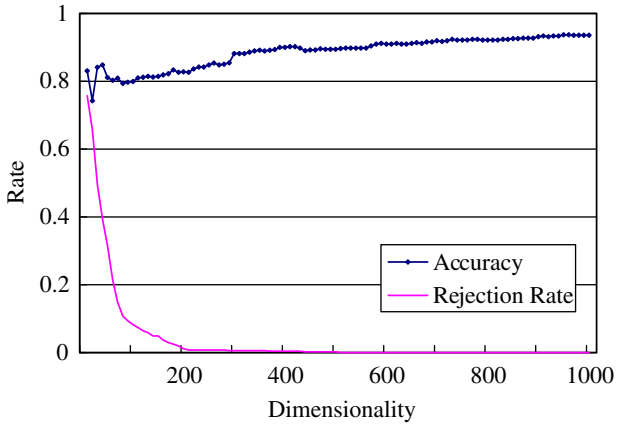


Fig. 8. Scalability test for minimum aggregation of PTF and ICF in the training phase.

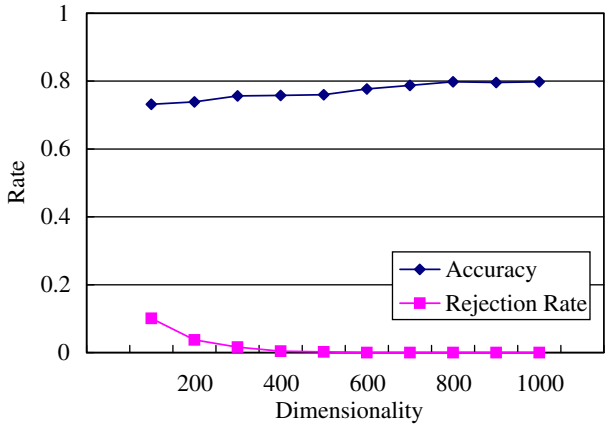


Fig. 11. Scalability test for weighted-average aggregation of PTF ($w = 0.5$) and ICF ($w = 0.5$) in the test phase.

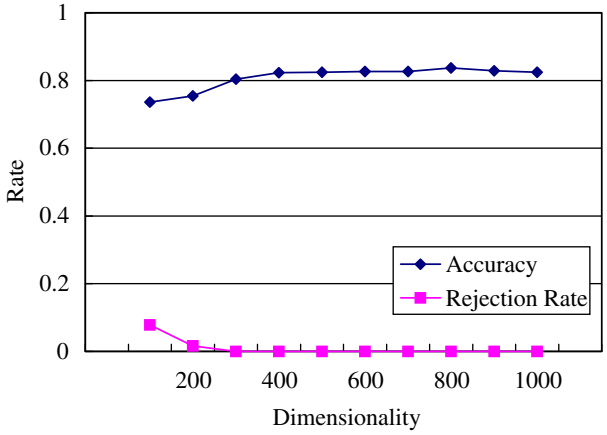


Fig. 9. Scalability test for minimum aggregation of PTF and ICF in the test phase.

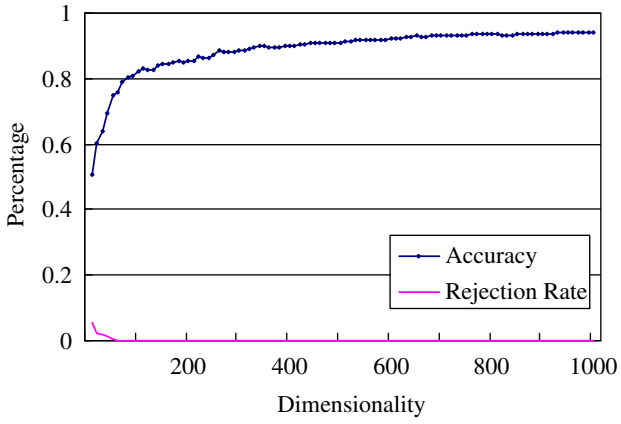


Fig. 12. The scalability test for DPM in the training phase.

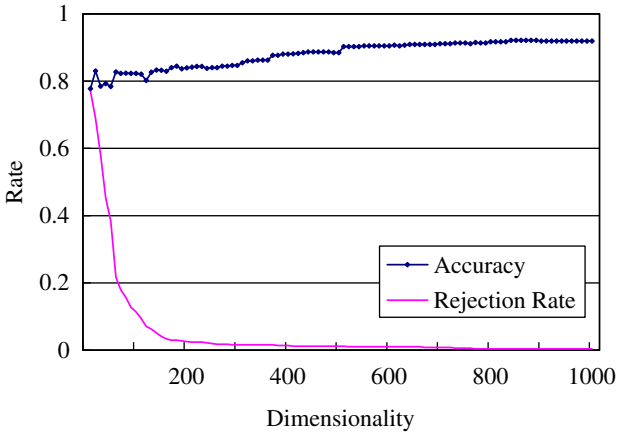


Fig. 10. Scalability test for weighted-average aggregation of PTF ($w = 0.5$) and ICF ($w = 0.5$) in the training phase.

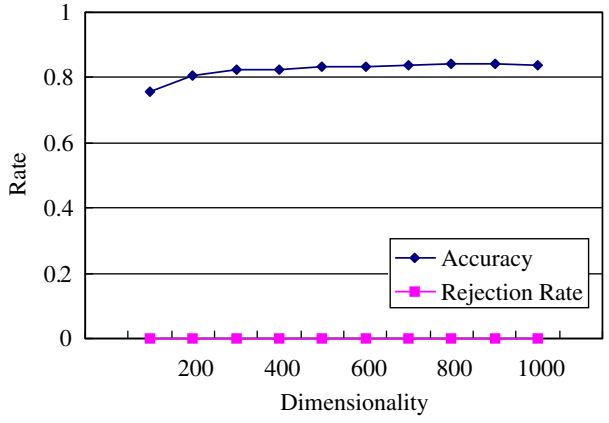


Fig. 13. The scalability test for DPM in the test phase.

improper. Second, the quality of the TF measure may be too low to promote significantly. The results of the minimum aggregator operator are shown in Figs. 8 and 9, and the results of the weighted-average aggregator with weights $= \langle 0.5, 0.5 \rangle$ are shown in Figs. 10 and 11. We also use some different weight combinations to observe the scalability results of weight aggregation in the training phase and the testing phase. We find from the results of the weighted-average operator with different weights do not show obvious promotion; nevertheless, the rejection rates, which indicate the side effects of relevance measures, vary greatly when the weights change.

In Figs. 12 and 13, the results are very encouraging. The discriminating power measure (DPM) reduces the input dimensionality by 78% (from 10,427 to 200) with zero rejection rate and with less than 5% degrading (from 84.5% to 80.4%) in the test accuracy. The 84.5% classification accuracy rate derives from no feature selection procedure (i.e. 10,427 feature terms selected) is applied to select significant feature terms, but the 80.4% classification accuracy rate comes from the proposed PDM measure is applied for feature selection (i.e. 200 feature terms selected). This experiment can demonstrate the DPM measure not only reduces redundancy but also reduces noise features.

5.3. Classification experiments on the China-Times and reuters datasets

To consider the impacts on classification accuracy for the proposed feature selection methods, we implement experiments on the China-Times dataset to demonstrate that the proposed feature selection methods also provide benefit to promote document classification accuracy rate. These experimental results are listed as Table 6. In Table 6, weights $= \langle 0.25, 0.75 \rangle$ mean that two relevance measures are aggregated as an integrated measure by the weighted-average operator with the corresponding weights 0.25 and 0.75, and so forth. According to our experimental results on the China-Times dataset, we find that the classification accuracy rate is improved after the proposed feature selection approach is applied. Importantly, for the China-Times dataset (eight categories, 511 training examples, 485 testing instances), the testing accuracy without feature selection (using 10,427 features) was 84.5%, and testing accuracies with feature selection (using 4000 features, which is the smallest to guarantee zero rejection for all the feature selection methods we used) by the DPM method and the weighted average aggregation of PTF ($w = 0.5$) and ICF ($w = 0.5$) were 85.4% (0.9% improved), 84.7% (0.2% improved), respectively. These results verify that the proposed DPM method is superior to other four tested feature selection methods.

Besides, in order to evaluate the classification performance of our methods for larger datasets, we also perform a comparison of classification accuracy rate on Reuters-21578 dataset for vector space model classifier with two different feature selection approaches, i.e. mutual information feature selection addressed by Han (1999) and our proposed discriminating power measure. First, we use stop words and the word stemming algorithm proposed by Porter (1980) to preprocess the documents. Then we apply our discriminating power measure and mutual information feature selection algorithm (Yang & Pedersen, 1997) respectively to select top 2000 features to set up a vector space model classifier (Chen et al., 2001) and compare the testing classification results listed in Table 7. The results show that the classification accuracy rate of vector space model classifier is obviously improved because of using the proposed discriminating power measure feature selection algorithm. That is, this implies that the proposed DPM indeed can efficiently identify useful keywords to set up a better classifier.

Table 7
Testing results of vector space model classifier using different feature selection methods

Feature selection method	Performed classifier	Testing classification accuracy rate (%)
Mutual Information (Yang & Pedersen, 1997)	Vector space model	87.68 (Han, 1999)
Discriminating power measure		88.83

6. Discussion

In this paper, we have proposed an efficient discriminating power measure for feature selection and a two-level promotion technique to improve the behavior of some relevance measures often used in text categorization. However, there remain some issues which need to be further discussed.

6.1. Methods of input-dimensionality reduction

Chakrabarti, Dom, Aagrawal, and Raghavan (1998) pointed out that a major characteristic, or difficulty, of text categorization problems is the high dimensionality of the feature space. In the machine-learning literature, Dietterich (1997) and Blum and Langley (1997) also emphasized selecting relevant information to scale up

Table 6
Testing results of vector space model classifier using different feature selection methods with 4000 selected feature terms on the China-Times dataset

Feature selection method	# of testing documents	# of accurate documents	# of error documents	# of rejection documents	Testing classification accuracy rate
No feature selection is applied (use all 10427 feature terms)	485	410	75	0	0.845361
Term frequency	485	408	77	0	0.841237
Weighted average-aggregation (0.25, 0.75)	485	408	77	0	0.841237
Minimum aggregation	485	400	85	0	0.824742
Entropy (ICF)	485	395	90	0	0.814433
Weighted average-aggregation (0.75, 0.25)	485	395	90	0	0.814433
Weighted average-aggregation (0.5, 0.5)	485	411	74	0	0.847423
Discriminating power measure	485	414	71	0	0.853608

machine-learning algorithms so that they can be applied to problems with millions of instances, thousands of features, and hundreds of categories. Because Automatic Web Page Classifier (AWPC) may be the largest scale application of machine learning ever, the dimensionality-reduction process is the crucial challenge when it comes to practical implementation.

The feature selection method is frequently adopted to deal with the input dimensionality-reduction problem in text categorization (or called text classification) (Salton, 1989, 1983; Lewis, 1992; Yang, 1993; Huang, 1997; Lin et al., 2002; Chekuri and Goldwasser, 1997; Lin and Chen, 1986; Yang, 1999; Niu and Ji, 2004; Chen et al., 2006). However, why a particular feature selection method was chosen is rarely mentioned and this creates confusion. In Lewis (1992), Lewis cleared the confusion and enumerated four approaches for input-dimensionality reduction in text categorization:

(1) Feature selection (term selection)
Directly reduce the dimensionality by selecting relevant features and eliminating irrelevant features.
(2) Feature clustering (term clustering)
Directly reduce the dimensionality by grouping individual words or small fragments into closely connected words.
(3) Pattern clustering (document clustering)
Automatically generate categories of documents based on the similarities between documents.
(4) Factor analysis (latent indexing)
Transform one representation of a collection of documents into a new representation with desirable mathematical properties.

Because of the huge-scale dataset problems presented by AWPCs, the efficiency of input-dimensionality-reduction methods is very important. The factor analysis approach is less suitable for AWPCs due to the computational expense. The pattern clustering approach is also undesirable because the effect on the pattern representation is difficult to predict and patterns cannot be clustered without some existing representation (Lewis, 1992). Generally, both feature selection and feature clustering are appropriate methods to reduce the input dimensionality in AWPCs. The feature selection approach assumes that some features are irrelevant for classification and can be eliminated to improve accuracy or efficiency. The feature clustering approach recognizes redundancies among features, and then clusters these features. Restated, by replacing individual features with the clustered features, dimensionality can be reduced. In this study, the proposed *fuzzy ranking analysis* paradigm and *discriminating power measure* (DPM) belong to feature selection approach to enhance feature selection procedure for Web page classification. The *fuzzy ranking analysis* paradigm aims at promoting feature selection measures by the intra-level promotion or inter-level promotion. In addition, the *discriminating power measure* can simultaneously consider both positive and negative features, and it emphasizes classification in parallel order rather than classification in serial order, so that feature selection procedure is obviously promoted.

6.2. Challenges of automatic content analyzers on the web

Due to the explosive growth in the numbers of Web pages, various kinds of automatic content analyzers on the Web, like automatic Web page classifiers, are increasingly in demand. However, the rapid change in the Web environment leads to some challenges for the designers of automatic content analyzers:

(1) Popular usage of new complex design techniques for Web pages (like DHTML, VRML, JavaScript, and VBScript) makes it more difficult to analyze the content of Web pages.
(2) Complex Web-based systems (like Web databases, systems powered by CGI or ASP) need new manipulating techniques instead of indexing.
(3) News site indices require daily updates. This is a heavy load for search engines.
(4) Non-textual information (like images, sound, animation and other objects) is still difficult to analyze or index.

7. Conclusion

In the paper, we raise the problems of feature selection in Web page classification and text categorization. We try to solve the problems by

(1) promoting existing relevance measures by a *two-level promotion technique* based on fuzzy ranking analysis;
(2) analyzing and evaluating the scalability behavior by *ranking analysis*; and
(3) proposing the *discriminating power measure* (*DPM*) for feature selection to improve efficiency and accuracy of classification.

Instead of passively eliminating irrelevant features, we emphasize active selection of relevant features. After improving the relevance ranking by our proposed methods, the scalability of input features can be raised, i.e., we can use much fewer features to achieve almost the same degree of classification accuracy. Hence, a reasonable trade-off between accuracy and efficiency exits and an optimal can be chosen according to the scalability tests. Also, we propose a new relevance measure, DPM. The DPM has low computation cost and emphasizes on both positive and negative discriminating features. Also, it emphasizes classification in parallel order, rather than classification in serial order. The experimental results for this measure are encouraging. It can reduce the input-dimensionality from tens of thousand to a few hundreds with zero rejection rate and with less than 5% degrading (from 84.5% to 80.4%) in

the test accuracy on the China-Times dataset. Moreover, the classification results of China-Time and Reuter-21578 datasets further confirmed that the DPM provides major benefit to promote document classification accuracy rate. Restated, the DPM measure is not only able to reduce redundancy but also capable of reducing noise features, and thus helping set up a better classifier.

Acknowledgement

Chang wants to thank Prof. Jan-Ming Ho and Prof. Cheng-Yan Kao for their valuable discussions and comments.

References

Almuallim, H., & Dietterich, T. G., (1991). Learning with many irrelevant feature. In *Proceedings AAAI-91*, Anaheim, CA (pp. 547–552).

AltaVista, Search Robots. <http://www.altavista.com>.

Blum, A. L., & Langley, Pat (1997). Selection of relevant features and examples in machine learning. *Relevance, Artificial Intelligence, 97*(1–2), 21–23 [Special issue].

Chakrabarti, S., Dom, B., Aagrawal, R., & Raghavan, P. (1998). Scalable feature selection, classification and signature generation for organizing large text databases into hierarchical topic taxonomies. *VLDB Journal, 7*, 163–178.

Chekuri, C. & Goldwasser, M. H., (1997). Web search using automatic classification. Poster presentation papers of the 6th international WWW conference, California, USA.

Chen, C.-M. (2003). Incremental personalized web page mining utilizing self-organizing HCMAC neural network. *IEEE/WIC International Conference on Web Intelligence* (pp. 47–53).

Chen, C.-M., Lee, H.-M., & Hwang, C.-W. (2005). A hierarchical neural network document classifier with linguistic feature selection. *Applied Intelligence, 23*, 277–294.

Chen, C.-M., Lee, H.-M., & Tan, C.-C. (2006). An intelligent web-page classifier with fair feature-subset selection. *Engineering Applications of Artificial Intelligence, 19*(18), 967–978.

Chen, J.-M., Liu, T.-C., Lee, & H.-M., (2001). A modularized hierarchical document classification with the ability of handling similar documents. In *Proceedings of the sixth national conference on artificial intelligence and application (TAAI-2001)*, Taiwan, 2001.

Cohen William, W. & Singer, Y., (1996). Context-sensitive learning methods for text categorization. In *Proceedings of the 19th annual international ACM SIGIR conference on research and development in information retrieval (SIGIR'96)* (pp. 307–315).

Deerwester, S. et al. (1990). Indexing by latent semantic analysis. *Journal of the American Society for Information Science, 41*(6), 391–407.

Dietterich, T. G. (1997). Machine-learning research: four current directions. *AI Magazine, 18*(4), 97–136.

Fodor, J. C., Marichal, J.-L., & Roubens, M. (1995). Characterization of some aggregation functions arising from problems. *Fuzzy Logic and Soft Computing*, 194–201.

Friedman, N., Geiger, D., & Goldszmidt, M. (1997). Bayesian network classifier. *Machine Learning, 29*, 131–163.

Hamming, R. W. (1980). *Coding and information theory*. Englewood Cliffs, NJ: Prentice-Hall.

Han, E. H. (1999). *Text categorization using weight adjusted K-nearest neighbor classification*. PhD thesis. University of Minnesota.

Holmstom, L. et al. (1997). On neurobiological, neuro-fuzzy, machine learning, and statistical pattern recognition techniques. *IEEE Transactions on Neural Networks, 8*(1), 5–17 [Special issue on Artificial neural network and statistical pattern recognition].

Huang, Y.-L., (1997). *A theoretic research of cluster indexing for mandarin chinese full text document—the construction of vector space model*. PhD thesis. Taipei, Taiwan: Department of Business Administration, National Taiwan University.

Joshi, A. et al. (1997). On neurobiological, neuro-fuzzy, machine learning, and statistical pattern recognition techniques. *IEEE Transactions on Neural Networks, 8*(1), 18–31 [Special issue on Artificial neural network and statistical pattern recognition].

Kira, K. & Rendell, L., (1992). A practical approach to feature selection. In *Proceedings 9th international conference on machine learning*, Aberdeen, Scotland (pp. 249–256).

Langley, P. & Sage, S., (1994). Oblivious decision trees and abstract cases. *Working notes of the AAAI-94 workshop on case-based reasoning*, Seattle, WA (pp. 113–117).

Lee, H.-M., Chen, C.-M., & Lu, Y.-F. (2003). A self-organizing HCMAC neural network classifier. *IEEE Transactions on Neural Networks, 14*(1), 15–27.

Lewis, D. D., (1992). Representation and learning in information retrieval. PhD Dissertation, Tech. Report UM-CS-1991-093, Department of Computer Science, University of Massachusetts, Amherst, MA. <ftp://ftp.cs.umass.edu/pub/techrept/techreport/1991/UM-CS-1991-093.ps>, <http://www.research.att.com/~lewis/papers/lewis91d.ps>.

Lewis, D. D., (1999). Reuters-21578 text categorization test collection distribution 1.0. <http://www.research.att.com/lewis>.

Lewis, D., Schapire, E., Callan, P., & Papka, (1996). Training algorithms for linear text classifier. In *Proceedings of the 19th annual international ACM SIGIR conference on research and development in information retrieval (SIGIR'96)* (pp. 298–306).

Lin, C., & Chen, H. (1986). An automatic indexing and neural network approach to concept retrieval and classification of multilingual (Chinese–English) documents. *IEEE Transactions SMC—Part B: Cybernetics, 26*(1), 75–88.

Lin, S. H., Chen, M. C., Ho, J. M., Ko, M. T., & Huang, Y. M. (2002). ACIRD: Intelligent internet documents organization and retrieval. *IEEE Transactions on Knowledge and Data Engineering, 14*(3), 599–614 [Special issue on web technologies].

Lippmann, R. P. (1987). An introduction to computing with neural networks. *IEEE Transactions on Acoustics Speech and Signal Processing, 2*(4), 4–22.

Lippmann, R. P. (1989). Pattern classification using neural networks. *IEEE Communications, 27*(11), 47–64 [Special Issue on Neural Networks in Communications].

Littlestone, N. (1998). Learning quickly when irrelevant attributes around: A new linear threshold algorithm. *Machine Learning, 2*, 285–318.

Musavi, M. T., Ahmed, W., Chan, K. H., Faris, K. B., & Hummels, D. M. (1992). On the training of radial basis function classifiers. *Neural Networks, 5*, 595–603.

Nadler, M., & Smith, E. P. (1993). *Pattern recognition engineering*. New York, NY, USA: A Wiley-Interscience Publication, John Wiley & Sons.

Niu, Z.-Y., & Ji, D.-H. (2004). Feature selection for Chinese character sense discrimination. *Lecture Notes in Computer Science, 2945*, 201–208.

Openfind, Search Robots, Taiwan. <http://www.openfind.com.tw>.

Porter, M. F. (1980). An algorithm for suffix stripping. *Program, 14*(3), 130–137.

Quinlan, J. R. (1983). Learning efficient classification procedures and their application to chess and games. *Machine Learning: An Artificial Intelligence Approach*, 463–482.

Quinlan, J. R. (1993). *C4.5: Programs for machine learning*. San Mateo, CA: Morgan Kaufmann.

Ribeiro, R. A. (1996). Fuzzy multiple attribute decision making: A review and new preference elicitation techniques. *Fuzzy Sets and Systems, 78*(2), 155–181 [Special issue on fuzzy multiple criteria decision making].

Salton, G. (1983). *Introduction to information retrieval*. McGraw-Hill.

Salton, G. (1989). *Automatic text processing: The transformation, analysis, and retrieval of information by computer*. Addison Wesley.

Yahoo! Classified directory. <http://www.yahoo.com>.

Yam, Classified directory, Taiwan. <http://www.yam.com.tw>.

Yang, Y.-Y. (1993). *Document automatic classification and ranking*. Master Thesis, Hsinchu, Taiwan: Department of Computer Science, National Tsing Hua University.

Yang, Y. (1994). Expert network: Effective and efficient learning from human decisions in text categorization. In *Proceedings of the 17th annual international ACM SIGIR conference on research and development in information retrieval (SIGIR'94)* (pp. 13–22).

Yang, Y. (1999). An evaluation of statistical approaches to text categorization. *Information Retrieval Journal, 1*(1/2), 67–88.

Yang, Y., & Chute, C. G. (1994). An example-based mapping method for text categorization and retrieval. *ACM Transactions on Information Systems (TOIS), 12*(3), 253–277.

Yang, S. H., & Hou, Y. C. (1998). A study on automatic document classification by combine fuzzy theory and genetic algorithms. *Journal of Fuzzy Systems, 4*(1), 45–57.

Yang, Y., & Pedersen, J. O., (1997). A comparative study on feature selection in text categorization. In *Proceedings of the 14th international conference on machine learning (ICML'97)*, Nashville, Tennessee. <http://www.cs.cmu.edu/~yiming/papers.yy/ml97.ps>.

Zimmermann, H.-J. (1987). *Fuzzy sets, decision making, and expert system*. Boston: Kluwer Academic Publishers.

Zurada, J. M. (1992). *Introduction to artificial neural systems*. West Publishing.