Abstract: - Web filtering is an inductive process which automatically builds a filter by learning the description of user interest from a set of pre-assigned web pages, and uses it to assign unprocessed web pages. In Chinese web filtering, content similarity analysis is the core problem, while Chinese words divided syncopation technology, ambiguousness of Chinese words and Chinese context-sensitive expression customs are all disturbing factors. The automatic-learning and relativity-analysis abilities of Machine Learning algorithms help solve the above problems and make ML useful in Chinese web filtering. From evaluation of four general ML algorithms (Decision Tree, Rule Induction, Bayesian algorithm and Support Vector Machines) by analyzing their description and generalization abilities and comparing their robustness and efficiency, SVM indicates most fitful for Chinese web filtering. While in practical applications, filtering results can be divided into three levels of relative, similar and homology based on user demand. A model of biased algorithms can be helpful. BSVM (for Biased Support Vector Machines) imports a stimulant function, uses training examples distribution n+/n- and an user-adaptable parameter k to express the user bias degree of different classes and can help improve the positive examples filtering efficiency.

Key-Words: - Chinese web filtering, Machine Learning algorithms, Biased Support Vector Machines

1 Introduction

The increasing internet resources cause a problem of “Information Overload, Quality Enhancement”. It means that people want to read the most interested messages, and avoid having to read low-quality or uninterested messages. Web filtering is the activity of classifying a stream of incoming web pages dispatched in an asynchronous way by an information producer to an information consumer [1], which helps people find the most interested and valuable information and saves internet users from drowned in information flood.

Web filtering is the task of assigning a boolean value to each web page vector \(d_i \in D\), where \(D\) is a domain of web pages. A value of \(T\) assigned to \(d_i\) indicates a decision to page \(d_i\) relative, while \(F\) indicates irrelative. More formally, the filtering task is to approximate the unknown target function \(\Psi: D \rightarrow \{T, F\}\) which describes how web pages ought to be assigned by means of a function \(\Phi: D \rightarrow \{T, F\}\) called the filter (or hypothesis). How to improve the precision and recall of a filter \(\Phi\) is the core technique of web filtering.

Recent years, the machine learning (ML) paradigm, instead of knowledge engineering and domain experts, becomes more popular in solving the above problem because of its automatic-learning and relativity-analysis abilities. People have done many efforts in comparison of ML algorithms in web categorization on different versions of Reuters [2], but few in web filtering, especially in web filtering of Asian languages. Here, we take Chinese as an example and evaluate the efficiencies of Machine Learning algorithms in Asian language web filtering applications.

According to the bias degree of user demand, web filtering result can be divided into three levels of relative, similar and homology. Thus a model of Biased Support Vector Machines (BSVM) which imports a stimulant function to express and evaluate the bias is introduced and analyzed in Section 4. BSVM can be helpful in web filtering applications of different bias degrees.

This paper forwards several challenging problems in Chinese web filtering applications, evaluates the efficiency of four main Machine Learning algorithms in this area, and introduces an enhancement of biased algorithms (BSVM). Remaining of the paper is organized as follows: In section 2 we forward several problems in Chinese web filtering and emphasize the usage of ML algorithms in web filtering. Section 3 detailedly evaluates the filtering efficiency and robustness of four main ML algorithms, focusing on the influence of feature set size and training set size. Section 4 introduces a model of Biased Support Vector Machines and analyzes its efficiency in web filtering. Section 5 closes the paper with our conclusions.
2 Challenging Problems in Chinese Web Filtering

Web filtering is an application of web categorization. But different from other applications, web filtering concentrates on web content similarity instead of subject relevant by analyzing the expression orientation of the whole web page. Especially in Chinese web filtering applications, Chinese words divided syncopation technology, ambiguousness of Chinese words and Chinese context-sensitive expression customs are all disturbing factors of web content similarity analyzing. So the automatic-learning and relativity-analysis abilities of web filtering algorithms are quite important, ML paradigm is widely adopted as filtering algorithms because of its advantages in it.

**Problem 1:** Web filtering concentrates on web content similarity instead of subject relevant by analyzing the expression orientation of the whole web page. Instead of pattern recognition, effect analysis algorithms are expected for solving binary-class classifying problem. Key sentences, even key paragraphs, are considered as basic analysis units instead of key words. Without predefined class descriptions, filtering algorithms must acquire user interest by itself.

We define a Chinese single character as \( c_i \in D \) where \( D \) is a domain of Chinese single characters, and a Chinese interpunction as \( b_i \in I \) where \( I \) is a Chinese interpunction set. So a Chinese word \( w_i = (a_1, a_2, \ldots, a_k) \) is a sequence of \( k \) single characters with \( k \) usually no more than 5. A sentence \( s_i = \{a_1, a_2, \ldots, a_k, b \} \in D, b \in I \} \) is a combination of characters \( a \) and words \( w \), and an interpunction \( b \) as a final symbol. A Chinese web page is a sequence of sentences. 

**Problem 2** A sentence in Chinese has no blanks between words like that in English, so we must translate the sentence \( s_i = \{a_1, a_2, \ldots, a_k, b \} \in D, b \in I \} \) into a form of words sequence \( s_i = \{a_1, a_2, \ldots, a_k, b \} \in D, b \in I \} \) based on Chinese vocabulary by analyzing the sentence structure, which is called Chinese words divided syncopation technology. But a word \( w_i = (a_1, a_2, \ldots, a_k) \) in different sentence of \( s_m = (a_1, a_2, a_3, a_4, a_5) \) (wo’qu’shang hai’du jia) and \( s_n = (a_1, a_2, a_3, a_4, a_5) \) (wo’shang’ai nan’du jia) may be divided as quite different forms just depending on the sentence meaning. So how to divide the Chinese sentence to best express its real meaning remains a puzzle.

**Problem 3** Ambiguousness of Chinese words: in Chinese, a single word may have more than one meaning, such as the chinese word ‘de’ has 5 pronunciations, 6 part-of-speech and 16 different meanings. Then how to recognize the real meaning of the word in a sentence?

**Problem 4** Chinese context-sensitive expression customs: in Chinese expression custom, person name, famous affairs and poetry may be used as the representation of spirit, behavior and attitude, which suppose readers have the same knowledge background as the writer. But how to comprehend the real meaning of such words in a sentence by computer? The specialties of Chinese expression customs are not limited to the above three. The core idea is that in a Chinese web page, the meaning of a word is judged not by itself, but by the neighbor words, even the neighbor sentences or neighbor paragraphs. Machine Learning paradigm has high relativity-analysis ability of considering different features of a web page vector dependent and combining them into some data structures such as rules or trees which reduce the influence of Chinese complex expression customs.

3 Evaluation of ML Algorithms in Chinese web filtering

The inductive construction of web filters usually consists in the definition of a function \( \Phi : D \rightarrow \{T, F\} \) which gives each web page \( d_j \) a decision value.

3.1 Four General ML Algorithms

We choose four general algorithms to construct Chinese web filter. A brief analysis of the description and generalization abilities for each algorithm is as follows:

- **Decision Trees (C4.5)**
  A decision tree \([3][4][5]\) is a graph of nodes connected by arcs with each internal node corresponding to a feature and each arc to a possible value of that feature. A leaf of the tree specifies the goal label for the records described by the path from root to leaf. The tree assigns a test page \( d_j \) by testing internal nodes along the proper arcs with the value of the internal nodes have in vector \( d_j \) till a leaf node is reached, the label of this leaf node is assigned to \( d_j \). Decision tree method is easily interpretable by humans and has low computational complexity, and it is a quite simple and practical idea in the field of ML.

- **Rules Induction (CN2)**
  Rule induction methods \([6][7]\) try to find a proper set of DNF rules for filtering task such that the error rate on training set is minimal. It attempts to find a compact “covering” rule set which completely assigns each web page with a proper label. By use of local optimization techniques, rule induction methods dynamically evaluate rules and revise the covering rule set.

- **Naïve Bayes Algorithms (NB)**
  Naïve Bayes algorithms \([8]\) view \( \Phi(d_j) \) in terms of \( P(c_i | d_j) \) (the probability that the web page \( d_j \) belongs to the class \( c_i \)) and compute this probability using
Bayes’ theorem. But direct estimation of $P(c_i|d_j)$ is impossible because of the sparseness of training data. Supposing features independent, we can calculate $P(c_i|d_j)$ with Equation (1) and classify $d_j$ into the class with the highest $P(c_i|d_j)$. Naïve Bayes, as a representative probabilistic algorithm, acts well in many applications.

$$P(c_j|d_j) = \frac{p(c_j) \prod_{i=1}^{T} p(w_{ij}|c_j)}{p(d_j)} \quad (1)$$

- Support Vector Machines (SVM)

Support Vector Machines [9] is a process of finding a surface which separates the positives from the negatives with the widest possible margin among all the surfaces in $|T|$-dimensional space, which is strongly supported by the Statistical Learning Theory. In the case that positives and negatives are linearly separable, the decision surfaces are the $|T|-1$-hyperplane decided by the maximum distance between several nearest points of the set. So the hyperplane is only determined by a small set of training vectors, called the support vectors. In the case that positives and negatives are not linearly separable, SVM imports slack variables $\xi_i \geq 0$ to measure misclassification error and uses kernel substitution method to convert linear algorithms into nonlinear ones. SVM acts well in dealing with large scale training set and it has no need of human and machine efforts in parameter tuning.

3.2 Experiments

We evaluate the four algorithms mentioned above of their filtering efficiency and robustness, focusing on the infection of feature set size and training set size on the filtering efficiency.

3.2.1 Data Collection

Standard benchmark collections used as the initial corpora for web filtering are publicly available for experimental purpose. But in Chinese, the only collections can be found are the three provided by Peking University, Chinese Academy of Science and Fudan University. Here we choose Fudan University’s benchmark collections because of their easy acquisition. The collections include 9,804 training examples and 9,833 evaluating web pages, which are assigned to 20 categories mutually. Here, we suppose each category $C_i$ stand for a user interest.

3.2.2 Settings

In evaluating the above four algorithms, we use the same pre-processing techniques to reduce the influence of some disturbing factors. The techniques include:

- Chinese Words Divided Syncopation Technology: multi-pattern matching automata using all the Chinese words in “Modern Chinese Dictionary” as initial feature set.
- Document expression: Vector Space Model with each dimension stands for a Chinese word and each vector stands for a sentence.
- Dimensional reduction: $\chi^2$ minimum Distribution Function method.
- Efficiency validation: repeat each test five times by selecting pre-limited N web pages randomly in the pre-limited category $C_i$, and average the five records as the result.

3.2.3 Experimental Process

We perform many test runs using different algorithms on different interest filtering tasks. In all test runs, none of the examples used in training process were used for testing. During each run for an interest filtering task of $C_i$ (the filter assigns test pages relative to the interest $C_i$ or not), we choose $(N/2)$ web pages from the training (or testing) subset of $C_i$ as positives and $(N/(2*19))$ web pages from other 19 training (or testing) subset as negatives. So each time we choose N web pages to form training (or testing) set. After filtering we compare estimated values of the test web pages with their real values (the benchmarks) to compute the filtering efficiency through the F1 evaluation Equation (2) [1].

$$F_1 = \frac{\text{precision} \cdot \text{recall} + \text{recall}}{\text{precision} + \text{recall}} \quad (2)$$

3.2.4 Filtering Efficiency Evaluation

From Fig. 1 we conclude that the average filtering efficiency of different algorithms is about: SVM>>CN2>>{C4.5, NB}.

NB assigned almost 90–95% positives as positives and 75–90% negatives as positives which reduce its precision much. In other words, NB filter considers different interest filtering tasks quite similar and can not induct their content difference clearly. NB algorithm omits the feature dependence, which does not suit for Chinese expression custom of context-sensitive.

C4.5 does not continually act well in different interest filtering tasks, especially getting quite low score on the filtering tasks of computer, agriculture and history. Take computer as an example, many glossaries that appear in web pages are not included in the initial feature set. C4.5 has to use sub-string of the glossary which appears in the initial feature set as a feature. So many sub-strings that C4.5 used to construct computer interest filter do not express enough semantic information as glossaries do and that leads C4.5 to encounter the over-fitting problem. As a result, C4.5 constructs the interest filtering tree so detailed that the tree does not suit for new web pages of the same interest. The average filtering
efficiency of CN2 is about 80%. It uses local optimal combination of features to express Chinese sensitive-context expression, which partly resolves the over-fitting problem and helps CN2 get relatively good result in Chinese web interest filtering. SVM presents the highest filtering efficiency (around 88.75% when the number of training examples is high). With the performance of quite acceptable training and filtering complexity, SVM is concluded best suited for the task in hand. The firm support of Statistical Learning Theory may avoid the over-fitting problem and play a role in improving the filtering efficiency.

3.2.5 Robustness Evaluation
With the conclusion drawn from Fig. 1, we choose three representative web interest filtering tasks of computer, economy and sports as examples to evaluate the robustness of ML algorithms. As follows are the influences of feature set size and training set size on the filtering efficiency.

Usually the algorithms use feature dependence to describe the filtering model, but different algorithms use different data structures and different feature evaluating functions. So the proper choice of feature set has some effects on the filtering efficiency of different ML algorithms. The left three figures of Fig. 2 illustrate that the filtering efficiency of algorithms improves linearly with the feature set size increasing, but C4.5 fluctuates more obviously than others on filtering tasks of computer and sports by the infection of the over-fitting problem.

The infection of pattern set size on algorithms is: 
\{C4.5\} > \{NB, CN2\} > \{SVM\}. SVM algorithm uses kernel substitution formulas to realize the reversion between low-dimension and high-dimension and selects feature subset itself from the initial feature set. Compared with SVM, C4.5, CN2 and NB use their own feature evaluating functions to choose proper features from the initial feature set during each attempt. Among them three, the feature evaluating formula of CN2 is a little better than that of the other two algorithms.

For different algorithms, the training set size has different influence to filtering efficiency. The right three figures of Fig. 2 illustrate that the filtering efficiency of algorithms improves linearly with the training set size increasing. NB fluctuates more obviously, and the reason is the feature independence suppose of NB. The four algorithms are all much infected by the training set size and the infection of training set size is about: 
\{C4.5\} > \{CN2, NB, SVM\}.

3.3 Result
From the efficiency and robustness evaluation of four ML algorithms on Chinese web filtering, we can conclude:

- SVM acts well in filtering task with strong robustness and acceptable efficiency.
- CN2 uses local optimization techniques to improve the describing ability and gets relatively high score.
- The precondition of NB that omitting the feature dependence reduces its web content analysis ability.
- Over-fitting problem occurring in the procedure of user interest description makes C4.5 not satisfied.

4 Enhancement
In practical web filtering applications, the filtering result is still considerable large. But users may want only several homology pages or all related ones based on the difference of page subject, writer’s viewpoint and expression orientation. So we divided filtering results into three levels:

- Relativity (R): Relative web pages stand for those containing several same key phrases or even several same key sentences. These pages express the same subject, but may be not consistent in viewpoint or orientation. Content relativity filtering is quite useful in user interest filtering tasks such as erotic pages filtering.
- Similarity (S): Similar web pages stand for those of same subject, viewpoint and orientation. Content similarity filtering is widely used in user bias filtering tasks such as harmful information filtering and has most popular applications.
- Homology (H): Homology web pages stand for those having a great many homology sentences, and even several homology paragraphs. Pages citation filtering is a typical application of content homology filtering.

Here we define the filtering result which acquired by ML algorithms as content relativity \((R_1)\), define ML algorithms analyzing result with probability near-to-1 as content homology \((R_2)\), and content similarity result \(R_i \in \{R_j | R_k \subseteq R_i \subset R_j\}\). Most filtering tasks can be described as application of content similarity filtering various in the degree of similarity. So we can adapt the filtering result by use of appropriate algorithms and proper parameters in order to fit the user demand better. A biased model which deals with different classes unequally is necessary.

4.1 Biased Support Vector Machine (BSVM)
In the classical SVM, a penalty function \(F=C \cdot \sum \xi_i\) is introduced as additional capacity control function, where the non-negative variable \(\xi_i\) is a measure of the misclassification errors and the coefficient \(C\) emphasizes the tolerant degree of misclassification.
error. Consequently the width of the margin decreases with C increases.

BSVM introduces a stimulant function, F=C[(k-1)n\sum Y_{i-1}Z_{i}n\sum Y_{i-1}Z_{i}]/n, as the extension of penalty function. In the right figure of Fig. 3, the rectangles mean the examples of \( y_{i}=+1 \), the circles mean the examples of \( y_{i}=-1 \), and those with black dot in them stand for support vectors. Thus we define \( n_+=|\{y_{i}=+1\}| \) and \( n_-|\{y_{i}=-1\}| \). The stimulant function uses both training examples distribution \( n_+/n_- \) and an user-adaptable parameter \( k \) to express the user bias degree of different classes. Together with the effect of penalty function, the bias is described in Equation (3). The width of the margin to the positive side decreases increases with the increase of \( n_+/n_- \) or \( k \).

Thus BSVM can find a proper separating hyperplane with respect to \( w, b, \xi \) that minimizes the functional,

\[
\arg \min_{w,b,\xi} J(w,b,\xi) = \frac{1}{2} ||w||^2 + C \sum_{i=1}^{n} \xi_i \quad \text{subject to} \quad y_i (w \cdot x_i - b) \geq 1 - \xi_i \quad \text{where} \quad \xi_i \geq 0, \forall i
\]

subject to the constraints of Equation

\[
y_i (w \cdot x_i - b) \geq 1 - \xi_i \quad \text{where} \quad \xi_i \geq 0, \forall i
\]

Here \( C_1 \) and \( C_2 \) are the classification errors stimulant coefficients, \( k \geq 0 \) is an adaptable parameter. The solution to the optimization problem of Equation (4) under the constraints of Equation (5) is given by the saddle point of the Lagrangian

\[
\mathcal{L}(w,b,\xi,\alpha,\beta) = \frac{1}{2} ||w||^2 + \sum_{i=1}^{n} \xi_i + \sum_{i=1}^{n} \alpha_i (1 + k_{ij} - y_i y_j) - \sum_{i=1}^{n} \beta_i \xi_i
\]

where \( \alpha, \beta \) are the Lagrange multipliers. The Lagrangian has to be minimized with respect to \( w, b, \xi \) and maximized with respect to \( \alpha, \beta \). The minimum with respect to \( w, b, \xi \) of the Lagrangian \( \mathcal{L} \) is given by

\[
\frac{\partial \mathcal{L}(w,b,\xi,\alpha,\beta)}{\partial w} = w - \sum_{i=1}^{n} \alpha_i y_i x_i = 0
\]

\[
\frac{\partial \mathcal{L}(w,b,\xi,\alpha,\beta)}{\partial b} = - \sum_{i=1}^{n} \alpha_i y_i = 0
\]

\[
\frac{\partial \mathcal{L}(w,b,\xi,\alpha,\beta)}{\partial \xi_i} = \begin{cases} C + C_1 - \alpha_i - \beta_i & \text{if } y_i = 1 \\ C - C_1 - \alpha_i - \beta_i & \text{if } y_i = -1 \end{cases} = 0
\]

And hence the solution to the problem is given by,

\[
\min_{\alpha} \frac{1}{2} \sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j K(x_i, x_j) - \sum_{i=1}^{n} \alpha_i
\]

with constraints,

\[
\sum_{i=1}^{n} y_i \alpha_i = 0
\]

\[
0 \leq \alpha_i \leq C + C_1 \quad \text{if } y_i = 1
\]

\[
0 \leq \alpha_i \leq C - C_1 \quad \text{if } y_i = -1
\]

### 4.2 Experiments and Analysis

For easy comprehension of BSVM, we choose Linear function (C=100) as kernel and simulate the training result of different biases under Matlab (Figure 4).

To show the efficiency of BSVM in practical applications, we choose 50 pages (40 positives, 10 negatives and half as training examples) of same subject from newsgroup, compute the positive sentences filtering precision under different C, and exhibit the influence of \( d=n_+/n_- \) and \( k \) in Fig. 5.

Concluded from the result, the pages filtering precision increases with \( n_+/n_- \) and \( k \) increasing.

### 5 Conclusion

Chinese web filtering is now a major research area which faces two difficulties including intelligent degree of content similarity analysis and complexity of Chinese expression customs. ML methods are imported to Chinese web filtering and make it acquire remarkable improvement with strong theoretical motivations. Among the four general ML algorithms, SVM acts well in filtering task with strong robustness and acceptable efficiency. According to user demand, biased algorithms (such as BSVM) can be helpful in improving filtering efficiency. But problems of user bias description and parameter self-adaptable are still open.

### References

1. Belkin N. J. Information filtering and information retrieval: two sides of the same coin: Commune. ACM 35, 12, 29–38
**Fig. 1.** The interest filtering efficiency of four Machine Learning algorithms.

**Fig. 2.** The infection of feature set size (the left three) and training set size (the right three) on interest filtering efficiency. Solid lines mean the changing tendency and dashed lines mean the actual changes.
Fig. 3. Analysis and demonstration of filtering result estimation. In the left figure, outside the biggest circles means filtering scope U, the smallest circle means user interest U_k, the biggest circle R_1 is the filtering result of general ML algorithms as content relativity, the smaller one R_k is the filtering result as content homology. The middle circle R_i means the biased filtering result according to user demand as content similarity. The right is a corresponding demonstration of Biased Support Vector Machine based on the left figure.

Bias=10  (n+/n-=4, k=6)  Bias=4  (n+/n-=4, k=0)  No bias

Fig. 4. The simulative training result with different biases. Black spots mean examples of y_i=+1, white spots mean those of y_i=-1, and spots with circle in them mean support vectors.

Fig. 5. BSVM filtering efficiency on different k and n+/n-. The left figure shows the influence of parameter d=n+/n- on the positive sentences filtering precision (k=1). The right figure shows the influence of parameter k on the positive sentences filtering precision (n+/n-=1).