A Novel Feature Selection Method Based on Category Information Analysis for Class Prejudging in Text Classification

Qiang Wang, Yi Guan, XiaoLong Wang, Zhiming Xu

School of Computer Science and Technology, Harbin Institute of Technology, Harbin, China

Summary
This paper presents a new feature selection algorithm with the category information analysis in text classification. The algorithm obscure or reduce the noises of text features by computing the feature contribution with word and document frequency and introducing variance mechanism to mine the latent category information. The algorithm is distinguished from others by providing a pre-fetching technique for classifier while it is compatible with efficient feature selection, which means that the classifier can actively prejudge the candidate class labels to unseen documents using the category information linked to features and classify them in the candidate class space to retrench time expenses. The experimental results on Chinese and English corpus show that the algorithm achieves a high performance. The F measure is 0.73 and 0.93 respectively and the run efficiency of classifier is improved greatly.

Key words:
Category Information Analysis, Pre-Fetching technique, Candidate class label, feature contribution

1. Introduction

As the volume of information available on the Internet and corporate intranets continues to increase, there is a growing need for tools helping people better find, filter and manage these resources. Text classification, the assignment of free text documents to one or more predefined categories based on their content, is an important component in many information management tasks. However, with the explosive growth of the web data set, algorithms that can improve the classification efficiency while maintaining accuracy are highly desired. Dimension Reduction techniques have attracted much attention recently since effective dimension reduction make the learning task such as classification more efficient and save more storage space.

Feature Selection (FS) algorithms are the most popular method for real life text data dimension reduction problems, which aims to remove non-informative features according to corpus statistics. Many novel FS approaches, such as filter and wrapper[1, 2] based algorithm were proposed in the past decades. In the text domain, the most popular used FS algorithms are still the traditional ones such as Information Gain (IG)[3], $\chi^2 - test$ (Chi)[4], Document Frequency (DF) and Mutual Information (MI)[5], etc. Information gain measures the number of bits of information obtained for category prediction by knowing the presence or absence of a term in a document. Given a corpus of training text, we compute the information gain of each term, and then remove those features whose information gain was less than some pre-determined threshold. The computation of CHI, DF, and MI are similar to that of IG. The differences are the approaches to rank features. However, MI is not comparable with IG, DF, and CHI on text categorization[6]. We use two of them, IG and CHI, as our baselines in this paper.

Though IG and CHI are popular in text categorization, they are greedy in nature and thus their solutions are not optimal according to some criterion. In this paper, we propose a novel feature selection algorithm with the category information analysis. We call this algorithm as *Category analysis based Feature Selection* (CAFS). The main advantages of this algorithm are: 1), it uses both of word and document frequency, not only the document frequency, to computes the feature contribution. Moreover it introduces variance mechanism to mine the latent category information. 2),The algorithm is distinguished from others by providing a pre-fetching technique for classifier while it is compatible with efficient feature selection, which means that the classifier can actively prejudge the candidate class labels to unseen documents using the category information linked to features and classify them in the candidate class space to retrench time expenses; 3), it is more efficient than the popular IG and CHI. Experiments on Chinese Library Classification (CLC) dataset and Reuters-21578 Corpus show the efficiency and effectiveness of our proposed approach.

This rest of the paper is organized as follows. In section 2, we introduce some related work on text data feature selection. In section 3, we give the mathematical notations used in this paper and our problem definition. In section 4, we describe our proposed CAFS algorithm. In

section 5, the experimental results on large scale data sets are given. Section 6 concludes our paper.

2. RELATED WORKS

A major characteristic, or difficulty, of text classification problems is the high dimensionality of the feature space. It is highly desirable to reduce the native space without sacrificing classification accuracy. Feature selection attempts to remove non-informative words from documents in order to improve classification effectiveness and reduce computational complexity. In this paper, we involve two popular used feature selection algorithms, Information Gain (IG) and $\chi^2 - test$ (CHI), which have been proved to be effective in the text domain as our baselines.We next give a brief introduction on IG and CHI in this section.

2.1 Information Gain

Following the notations above, Let (c_1,…,c_k) denote the set of possible categories, information gain of a selected group of terms T could be calculated by:

$$IG(t) = -\sum_{j=1}^{K} P(c_j)\log P(c_j) + p(t)\sum_{j=1}^{K} P(c_j|t)\log P(c_j|t) + p(\bar{t})\sum_{j=1}^{K} P(c_j|\bar{t})\log P(c_j|\bar{t}) \quad (1)$$

Where t is used to denote a unique term, IG (t) is the information gain of a term t, P (c_j) is the probability of class c_j , $P(t)$ is the probability of term t and $P(c_j|t)$ is the corresponding conditional probability. Here P (c_j) can be estimated from the fraction of documents in the total collection that belongs to class c_j and $P(t)$ from the fraction of documents in which the word t occurs. Moreover, P ($c_j|t$) can be computed as the fraction of documents from class c_j that have at least one occurrence of word t and $p(c_j|\bar{t})$ as the fraction of documents from class c_j that does not contain word t. 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.

2.2 $\chi^2 - test$

The $\chi^2 - test$ measures the lack of independence between word t and class c_j . It is given by:

$$\chi^2(t,c_j) = \frac{N\times(AD-CB)^2}{(A+C)\times(B+D)\times(A+B)\times(C+D)} \quad (2)$$

Here A is the number of documents from class c_j that contains word t, B is the number of documents that contains t but does not belong to class c_j, C is the number of documents from class c_j that does not contain word t and D is the number of documents that belongs to neither class c_j nor contains word t. N is still the total number of documents.We can compute the χ^2 statistics between each unique term and each category in a training corpus, and then combine the category specific scores of each term into:

$$\chi^2(t) = \sum_{j=1}^{K} p(c_j)\chi^2(t,c_j) \quad (3)$$

3. NOTATIONS AND PROBLEM DEFININITION

In this paper, a corpus of documents are mathematically represented by a $d\times n$ term by document matrix $X \in R^{d\times n}$, which is generated by a variation of the Okapi term weighting formula[7], where n is the number of documents, and d is the number of features (terms). Each document is denoted by a column vector x_i, $i=1,2,\ldots,n$, and the k^{th} entry of x_i is denoted by x_i^k , $k=1,\ldots, d$. X^T is used to denote the transpose of matrix X. Assume that these feature vectors are belonging to c different classes and the class size of the j^{th} class is n_j .

The dimension reduction problem could be defined as the finding of a function $f: R^d \to R^p$, where p is the dimension of data after dimension reduction ($p= d$), so that a document $x_i \in R^d$ is transformed into $y_i = f(x_i) \in R^p$. Thus the Feature Selection problem aims to find an subset of features indexed by k_l, l=1,2,…,p such that the low dimensional representation of original data x_i is denoted by $y_i = f(x_i) = (x_i^{k_1}, x_i^{k_2}, \ldots, x_i^{k_p})^T$.

Notice that usually the selective features are only weighted and regarded as one dimension of document vector, which may not reflect the informativeness of features. Some additional information (e.g., category information linked to the feature) can serve as "guidance" in determining which classes are more informative than the others. In fact, we can find such information along with documents in training corpus. So we incorporate the category information linked to features to feature selection stage and define the category information space of each feature as two tuples <*category, power*>. Namely, *category(t)* refers to classificatory label linked to term t and *power(t)* indicates the inference powers of term t to category identities whose value is confined to 0 (insignificance) or 1 (significance).

Following the notations and discussions above, we define the feature selection problem for text data categorization as:

Given a set of labeled training documents X, learn a transformation of feature space to obscure or reduce the noises of text features and devise intelligent pre-fetching mechanisms that allow for efficient prejudging the candidate class labels to unseen documents using the category information linked to features

4. Category Analysis based Feature Selection

The Category Information Analysis based Feature Selection (CAFS) selects features optimally according to the feature contribution to each class, which is the foundation of our approach. Thus in this section we introduce the definition of term-class contribution firstly. After that, we transform the feature contribution to feature score by using the variance mechanism. At the end of this section, we analyze category information linked to features, choose the appropriate class label and apply it to class prejudging in classifier.

4.1 Term-Class Contribution Criterion

The preprocessing of CAFS includes the stoplist* filtering, Chinese words-segment or English word-stemming. The strategy used to rank feature subsets is according to a correlation based heuristic evaluation function, which is toward subsets that contain features that are highly correlated with certain classes. The evaluation function of feature subset is based on the following hypothesis:

A good feature subset is the one that contains features highly correlated with (predictive of) the class, which can be evaluated by the document frequency (DF) and term within-document frequency (TF) in a class. The larger the DF and TF values in a class, the stronger relation between the feature and the class.

Here we define the term-class contribution criterion ($Sw_{ij}, j, i = 1\mathrm{L}\ n,\ j = 1\mathrm{L}\ m$) as follows:

$$Sw_{ij} = \frac{\log(fw_{ij} + \delta) * \log(dw_{ij} + \delta)}{\sqrt{\sum_{t=1}^{T} [\log(fw_{ij} + \delta) * \log(dw_{ij} + \delta)]^2}} \qquad (4)$$

Where $fw_{ij} = T_{ij} / L_j$, T_{ij} is the *TF* of feature w_i in class j, and L_j is the total number of terms in class j; $dw_{ij} = d_{ij} / D_j$, d_{ij} refers to the *DF* of feature w_i in class j, and D_j is the number of documents in class j, δ is a smooth factor and is set to 1.0 as default.

* Stop words are functional or connective words that are assumed to have no information content

4.2 Computing Feature Goodness via Variance Mechanism

This step filters out the errors among the candidate feature sets (CF_k) using statistical measures. Because some common features that are unworthiness to classifier may have high Sw_{ij} values to most classes, so we introduce variance mechanism to remove these features. We define the term-goodness criterion ($Imp(w_i)$) as follows:

$$Imp(w_i) = \sqrt{\sqrt{\sum_j (Sw_{ij} - \bar{S}_i)^2} \Big/ \sum_j Sw_{ij}} \qquad (5)$$

In equation (5), the larger the $Imp(w_i)$ values, the bigger the difference of feature w_i among classes and the more contributions the feature to classifier. The variable $\bar{S}_i$ in equation is defined as $\bar{S}_i = \sum_j Sw_{ij} / m$.

Then the accepted feature set is obtained by setting a heuristic value φ for threshold.

4.3 Class Prejudging using the Category Information Provide by CASF

A question regarding CAFS is how to determine the category information of the features selected. We use the average term-Class Contribution function approach to solve this problem. The average term-Class Contribution function is defined as:

$$avg\_score(t_j) = \frac{1}{|d|} \sum_{i}^{|d|} sw_{ij} \qquad (6)$$

Where d is the num of class that the sw_{ij} returns non-zero value. Thus if the term-class contribution sw_{ij} of t_j is larger than $avg\_score(t_j)$ and a pre-defined threshold γ, the category information, such as class label and inference power (e.g. 1) is set to term t_j.

In order to applying category information to TC system, the proposed TC model takes the significant reference of class label linked to each feature into consideration and employs class prejudging to cut off noise features, which means that only the classifier that shared class label with the document features can be used to perform classification. But perhaps some features have low inference power to class identities, this case we will back to the traditional method of no prejudging. Suppose that an unseen document D is represented with function R and is classified into class C with function T. This TC procedure that incorporated with CASF can be denoted as a four tuples like $M = <D, C, R, T>$, which the mathematical expectation is $C = (T \cdot R)(D)$. The function T is defined as:

$$T=\begin{cases} T_{prejudge}(D) & if \quad \forall t \ \{c_i | t\in T \wedge c_i \in category(t) \wedge power(t)=1\} \neq \phi \\ T_{No-prejudge}(D) & if \quad \forall t \ \{c_i | t\in T \wedge c_i \in category(t) \wedge power(t)=1\} = \phi \end{cases} \quad (7)$$

In which $T_{prejudge}(D)$ and $T_{No-prejudge}(D)$ refer to classify the unseen document based on class prejudging or not respectively, T means the feature set and c_i denotes the prejudging class label set obtained.

5. Experiments

In this section, we conduct our experiments on two real large scale text data sets to show the performance of CAFS. We first describe the experiments setup, then give the experimental results, and finally discuss the results.

5.1 Experiment Setting

The experiment operates Chinese and English text as processing object. Chinese text evaluation uses the Chinese Library Classification (*CLC*) 4[8] (Simplified Version) as criteria (T and Z are taken no account of) and the training corpus (3,600 documents) and the test data sets (3,600 documents) are the data used in TC evaluating of the High Technology Research and Development Program (863) in 2003 and 2004 respectively. The English text evaluation adopts Reuters-21578 which is Modapte split and training (6,574 documents) and test data sets (2,315 documents) are performed on the 10 largest categories[9].

This paper selects the IG and CHI, which have been proved to be very effective and efficient, as our baseline algorithms. And the one-versus-others SVM algorithm is used as classifier to perform text classification.For SVM, we used the linear models offered by *SVMlight* and sigmoid train is processed by mode-trust algorithm[10].

As performance measures, we followed the standard definition of recall, precision, and F1 measure. For evaluation performance average across categories, we used macro- and micro- averaging method [9]. To verify our method on prejudging candidate class, the Correct Ratio (*CR*) is defined to observe the resulting values.

$$CR=\frac{\sum N_t}{\sum N_t+\sum N_f} \quad (8)$$

Where N_t *and* N_f refer to the prejudge results on one unseen document respectively. If the true class label of the unseen documents is included in the prejudge results, N_t is set to 1, otherwise N_f is set to 1.

We apply the CAFS on all the training data to select various features to compare the effectiveness with baselines. In all our experiments, we use a single computer with Pentium(R) 4 2.80GHz CPU, and 512MB of RAM, to conduct the experiments. The experiment consists of the following steps:

- Apply the feature selection algorithm on a specific size of the training data to select a group of features with various numbers;
- Train the SVM classifier by *SVMlight* (linear kernel is used and the parameters used are all defaulted ones);
- Represent all the testing data to the selected low dimensional space;
- Evaluate the classification performance, using Micro F1, on the transformed testing data;
- Rerun this procedure on different training and testing data and record the average Micro F1 of all the algorithms involved.

5.2 Results

In this section we'll describe a series of experiments conducted on the proposed system. The results achieved below allow us to claim that the CAFS effectively provides us with a more efficient TC algorithm.

5.2.1 CLC

The classification performance on CLC data is summarized in Figure 1. The x-axis is the number of selected features and the y-axis is the Micro F1. From this figure, we can infer that the CAFS is constantly better than its counterpart selected by IG and CHI. In other words, CAFS algorithm can achieve better performance for text classification than the widely used traditional algorithms. And then we can draw the conclusion that, CAFS has better performance with the traditional feature selection algorithms especially in extremely low dimension space and it is more efficient. For instance the improvements by Micro F1 are about 0.14 and 0.09 respectively in contrast to CHI and IG. Moreover, it is proved that using the category information linked to each feature to perform class prejudging is effective and promising.

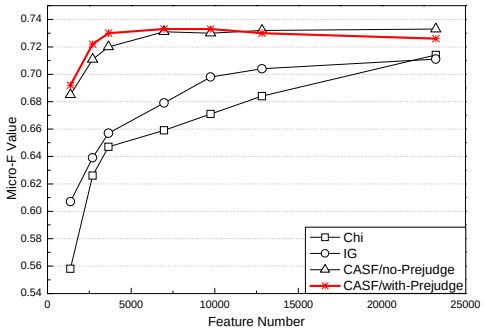

Fig 1. Micro F1 of classification on CLC data dimension reduced by CAFS, IG and CHI

The classifier runs frequency by each algorithm in feature selection for classification is reported in Figure 2. We can see that the classifier's run frequency can be much less than the others by introducing the class prejudging.

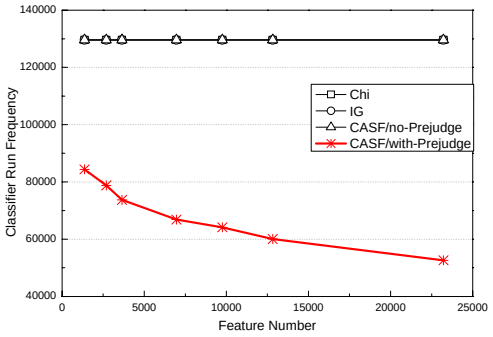

Fig 2. Classifier runs frequency on CLC data dimension reduced by CAFS, IG and CHI

Besides the Micro F1 and classifier runs frequency, we also give the evaluation of class prejudging results by Eq. 6. The line of table 1 shows the number of features selected by CASF. From the table 1 we can see that the CASF algorithm can effectively mine the category information linked to each feature and so the class prejudging can be very precise, which is the base of the classifier combined with class prejudging to achieve high performance.

Table 1: The evaluation results on Correct Ratio which uses the CASF with class prejudging on CLC corpus

	Text Classification based on CAFS		
	Prejudge True (N_T)	***Prejudge False (N_F)***	***Correct Ratio***
10,000	2410	38	98.4%

5.2.2 Reuters-21578

The classification performance on Reuters-21578 data is summarized in Figure 3. From this figure, we can infer that the low dimensional space selected by CAFS have better performance than its counterpart selected by the popular used IG and CHI. The classifier runs frequency with class prejudging or not is showed in Figure 4. The results are comparable same with CLC dataset. These indicate that in practice on web scale data, the performance of CAFS is outstanding. Though the efficiency improvement is only about half, to a real large scale problem, save half time is significant improvement.

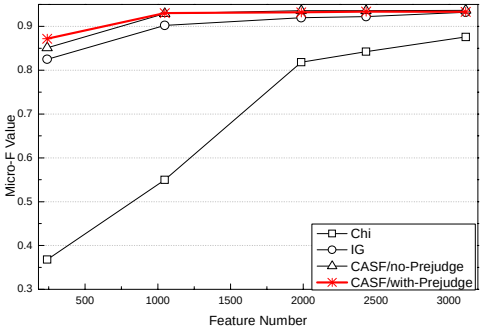

Fig 3. Micro F1 of classification on Reuters-21578 data dimension reduced by CAFS, IG and CHI

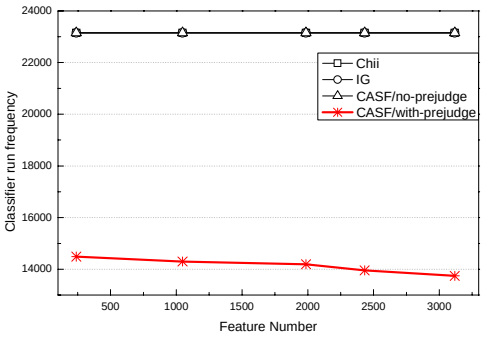

Fig 4. Classifier runs frequency on Reuters-21578 data dimension reduced by CAFS, IG and CHI

Table 2 further proved that the CASF is also effective on Reuters-21578 corpus and we can make use of the category information to perform class prejudging in text classification task.

Table 2: The evaluation results on Correct Ratio which uses the CASF with class prejudging on Reuters corpus

	Text Classification based on CAFS		
	Prejudge True (N_T)	***Prejudge False (N_F)***	***Correct Ratio***
2,500	2105	19	99.1%

5.3 Discussion of Results

From the experiments we can see that the proposed CAFS is consistently better than IG and CHI especially when the reduced dimension is small for text categorization problems. On the other hand, it is more efficient than the others by using only about half of the classifier run frequency to perform unseen document classification. To very large scale data such as the rapid growth web data, saving about half of the computation time is valuable and exciting. From the dimension by Micro F1 figures (Figure 1, Figure 3) we can draw the conclusion that CAFS can get significant improvements than baselines when the selected subspace dimension is small while get a little better performance when the selected subspace dimension is relative large. This phenomenon occurs due to the reason that when the selected feature dimension is small, the proposed CAFS, which applies the category information to the features and recur it to prejudge the candidate class, can outperform the ones that neglect the good inference powers of feature to category identities. With the increasing number of selected features, the saturation of features makes additional category information linked to features of less value. Then when the number of selected features is large enough, all feature selection algorithms involved can achieve comparable performance no matter they are exploring the category information or not.

6. Conclusion and Future work

In this paper, we proposed a novel efficient and effective feature selection algorithm, Category Analysis based Feature Selection (CAFS), for text categorization. With the growing number of text documents on the Web, many traditional text categorization techniques fail to produce a satisfactory result in handling this scale of data due to their time complexity and storage requirements. The CAFS can help save both data storage space and computation time by feature selection. The CAFS can obscure or reduce the noises of text features by computing the feature contribution with word and document frequency and introducing variance mechanism to mine the latent category information. The algorithm is distinguished from others by providing a pre-fetching technique for classifier while it is compatible with efficient feature selection, which means that the classifier can actively prejudge the candidate class labels to unseen documents using the category information linked to features and classify them in the candidate class space to retrench time expenses.

Still further research remains. Firstly, this study uses single word as candidate features, thus leading many valuable domain-specific multi-word terms for TC classifier are lost. So in future works multi-word recognition should be investigated. Secondly, using variance to evaluate the contribution of features among classes may introduce some noise features when lower *DF* and *TF* terms appears in one class by chance, so the solution to the more efficient feature evaluating criterion will continue to be studied.

Acknowledgments

This research was supported by National Natural Science Foundation of China (60435020, 60504021) and Key Project of Chinese Ministry of Education & Microsoft Asia Research Centre (01307620).

References

[1] Kohavi, R. and G. John. *Wrappers for feature subset selection.* Artificial Intelligence, special issue on relevance, 1997. **97(1-2)**: p. 273-324.

[2] Tsamardinos, I. and C.F. Aliferis. *Towards principled feature selection: Relevancy, filters and wrappers.* Ninth International Workshop on Artificial Intelligence and Statistics, 2003.

[3] Quinlan, J. R. *Induction of Decision Trees.* Machine Learning, 1986. **l(1)**: p. 81-106.

[4] Tom, Mitchell. *Machine Learning.* McGraw Hill, 1996.

[5] Kenneth Ward Church and Patrick Hanks. *Word association norms, mutual information and lexicography*. in *In Proceedings of ACL 27*. p. 76-83.Vancouver, Canada: 1989.

[6] Yang, Y, Pedersen and Jo. *A Comparative Study on Feature Selection in Text Categorization.* Proc. of the 14th International Conference on Machine Learning ICML97, 1997: p. 412-420.

[7] Tom, Ault and Yang Yiming. *kNN at TREC-9.* In: Voorhees EM and Harman DK, Eds., Proceedings of the Ninth Text. REtrieval Conference (TREC-9). Department of Commerce, National Institute of Standards and Technology, 1999: p. 127–134.

[8] *Introduction of Chinese Library Classification (CLC).* http://lib.nju.edu.cn/lib_class.html, 2003.

[9] K. Aas and L. Eikvil. *Text Categorisation: a Survey.* Technical Report, Norwegian Computing Center, 1999.

[10] John, C. Platt. *Probabilistic Outputs for Support Vector Machines and Comparisons to Regularized Likelihood Methods.* In : Advances in Large Margin Classifiers, MIT Press, 1999: p. 61-73.

Qiang Wang reived the B.S.degrees in School of Computer Science and Technology from Harbin Institute of Technology in 2000 and now is a PH.D candidate. In 2004, he participates in The First International Joint Conference on Natural Language Processing （IJCNLP2004） and give an oral presentation. He is interested in the

Theories and Methods for Question answering, machine learning and Text mining.

Yi Guan holds a B.Sc. degree in Computer Science and Technology from Tianjin University in 1992, and a Ph.D. degree in Computer Science and Technology from Harbin Institute of Technology in 1999. In 1996, Dr. GUAN was an invited visiting scholar in Canotec Co.,Japan. In 2000, Dr. GUAN was research associate in Human Language Technology Center at Hong Kong University of Science and Technology, and he was a research scientist in Weniwen.com (Hong Kong) limited in 2001. In October 2001, he became an associate professor in School of Computer Science and Technology in Harbin Institute of Technology. Dr. GUAN's research interests include: question answering, statistical language processing, parsing, text mining.

Xiaolong Wang received the B.E. degree in computer science from Harbin Institute of Electrical Technology, China, and the M.E. degree in Computer Architecture from Tianjin University, China, in 1982, and 1984, respectively, and the Ph.D. degree in Computer Science and Engineering from Harbin Institute of Technology, China, in 1989. He was a senior research fellow at the polytechnic University from 1998 to 2000. Currently, he is a Professor of computer Science at Harbin Institute of Technology, China. His research interest includes artificial intelligence, machine learning, computational linguistics, and Chinese information processing.

Zhiming Xu, born in October1967, received Ph.D from Harbin Institute of Technology, China, in 2001. Now he is working for Center of Language Technology, School of Computer Science and Technology, Harbin Institute of Technology as an associate professor. From 2001 to 2004, he worked for Department of Chinese Translation and Linguistics as Senior Research Associate. His research interests include Image Recognition, Biometrics Verification, Image Similarity Computation, Content-based Image search engine, Multilingual Website Data Mining, etc.