International Journal of Advanced Intelligence
Volume 3, Number 2, pp. 173-188, July, 2011.

# A Naive Bayesian Multi-label Classification Algorithm With Application to Visualize Text Search Results

Zhihua Wei, Hongyun Zhang*, Zhifei Zhang, Wen Li, Duoqian Miao

*Department of Computer Science and Engineering, Tongji University, No.4800, Cao'an Road, Shanghai,201804, China*

*Key laboratory "Embedded System and Service Computing" Ministry of Education, No.4800, Cao'an Road,*
*Shanghai,201804, China*

*zhihua_wei, zhanghongyun, dqmiao@tongji.edu.cn, zhifei.zzh@gmail.com, jx_wenli@yahoo.com.cn*

Received (10 January 2011)
Revised (21 April 2011)

Search results visualization has emerged as an important research topic due to its application on search engine amelioration. From the perspective of machine learning, the text search results visualization task fits to the multi-label learning framework that a document is usually related to multiple category labels. In this paper, a Naïve Bayesian (NB) multi-label classification algorithm is proposed by incorporating a two-step feature selection strategy which aims to satisfy the assumption of conditional independency in NB classification theory. The experiments over public data set demonstrate that the proposed method has highly competitive performance with several well-established multi-label classification algorithms. We implement a prototype system named TJ-MLWC based on the proposed algorithm, which acts as an intermediate layer between users and a commercial Internet Search Engine, allowing the search results of a query displaying by one or multiple categories. Testing results indicate that our prototype improves search experience by adding the function of browsing search results by category.

*Keywords*: Multi-label classification; Naïve Bayesian classification; text search results visualization; Feature selection.

## 1. Introduction

Nowadays Web documents have been growing rapidly. For example, everyday a large number of documents are uploaded to Web sites. As one of the most prevalent applications in today's network computing environment, Web search engines are widely used for information seeking and knowledge elaboration. However, too many returned results make us submerge in the sea of information. The current search engines (such as Google, MSN, Yahoo, etc.) return sorted search results according to the keywords input by users. In order to find the interesting information, users should browse the title and digest one by one within the search results.

*Corresponding Author

Focusing on above problems, some researchers begin to explore more advanced information retrieval methods. Generally speaking, there are two ways. One is semantic based information retrieval which uses semantic information to understand the documents and queries.[1] The other is machine learning based method which uses model learned from history data to reorganize Web documents such as classification and clustering. This paper mainly discusses the later, that is, the improvement of information retrieval based on machine learning.

Web search results visualization is a process that layouts search results in a more clear and coherent way according to the content of each result. It aims at improving the search efficiency and accuracy and ameliorating users' browsing experiences. Prevalent technique for this task is text clustering. It regards the visualization task as an unsupervised pattern classification problem. Following the methodology of pattern classification, text features are first extracted from texts to represent the document, and then the document is assigned to a cluster in which documents have high similarity. Generally, the "name" of cluster is given automatically according to the feature words of documents in this cluster. Related researches could be found in literatures.[2,3,4,5]

The search engines based on clustering technology already exist such as Vivisimo[a] and Grokker[b]. However, there is a problem that the extracted "name" of each cluster is far from expressing the main contents of its cluster. In this case, it's difficult for users to locate interesting information in a corresponding cluster. This made the process of visualizing search results less significant.

Unlike the traditional pattern classification problems in which each pattern has a unique label, a document may be associated with more than one label (e.g. a document may include the content about economy and computer). Multi-label learning originated from researches on text categorization problems, which each document may belong to several predefined labels. In this case, each document in training set is associated with a set of labels, and the task is to output a label set for each unseen document through analyzing training documents with known label sets.

In this paper, we borrow the idea of granularity computing [6] to display search results in a more refined granularity level-category. A Naïve Bayesian Multi-label Classification approach is proposed to handle the web text search results visualization problem. Feature selection strategy is concerned to improve classification performance. A prototype system TJ-MLWC is designed as an application example. In TJ-MLWC, users could look for a document within some interesting classes. In the same time, a document could be found in various corresponding classes.

The rest of this paper is organized as follows. Section 2 reviews the related works of visualizing text search results and multi-label learning. Section 3 proposes the Naïve Bayesian Multi-Label (NBML) approach. Section 4 reports experimental results on public data set. Section 5 presents the prototype system TJ-MLWC.

[a] Vivisimo Search Engine. http://www.vivisimo.com.
[b] Grokker Search Engine. http://www.grokker.com.

Finally, Section 6 summarizes and sets up several issues for future work.

## 2. Related Works

### 2.1. *Related works on visualizing text search results based on machine learning*

Visualizing search results by methods based on machine learning is one obvious solution for dealing with information overload. Clustering is one way that allows users to view categorized results without having to deal with the costs and complexities of building taxonomies (see, for example, the Vivisimo search engine). This advantage makes clustering technique prevalent in the field of visualizing search results. D. Roussinov and H. Chen proposed an approach of summarizing query results by automatic text clustering and implemented a prototype.[2] G. Mecca et al. used Latent Semantic Indexing on the whole document content and proposed a strategy, called Dynamic Singular Value Decomposition (SVD) Clustering, to discover the optimal number of singular values to be used for clustering purpose. A tool based on this algorithm has been integrated into the Noodles search engine.[3] Nicole Lang Beebe et al. proposed a kind of post-retrieval clustering of digital forensic text string search results specifically by using Kohonen self-Organizing Maps, a self-organizing neural network approach.[4] Zamir and Etzioni made an empirical comparison of standard ranked-list and clustered presentation systems when designing a search engine interface named Grouper, and reported substantial differences in use patterns between the two.[5] Some researchers have experimented with highly metaphorical visualizations. For example, Cugini et al. presented users with structural overviews of result sets and promoted visualization as the best approach to dealing with broad search tasks.[7] Visualization of this type appears to make it easier for users to locate worthwhile information and to comprehend search results.

Recently, some researches use topic map based on probabilistic graph model theory to improve representation of search results. Newman et al. explored semantic visualizations of Web search results based on topic map.[8] Their topic maps are based on a topic model of the document collection, where the topic model is used to determine the semantic content of each document. The topic model represents documents as a mixture of topics, which is a more flexible representation than k-Means clustering,[9] where a document belongs to just one topic. This richer representation means that one can navigate to related articles via any of the topic facets in a given article (this is not possible for k-Means). Furthermore, since the topic model is based on a generative probabilistic model, it can be flexibly extended to include other meta-data and attributes, such as authors, citation links, and subject headings, and therefore it is also preferred over Nonnegative Matrix Factorization (NMF).[10]

All these researches improved search engine performance to some extent. However, the methodology of clustering suffers from the difficulty of naming the clusters. Users could not comprehend the content of a cluster only from the words automatically given. It is still very difficult for user to find interesting information.

### 2.2. *Related works on multi-label classification approaches*

Traditional single-label classification is concerned with learning from a set of examples that are associated with a single label $l$ from a set of disjoint labels $L$, $|L| > 1$. In multi-label classification, the examples are associated with a set of labels $Y \subseteq L$.[11]

Some multi-label classification algorithms have been successfully designed for the application of text classification. Majority of traditional machine learning algorithms for single label classification problem have been extended to multi-label cases. Representative research works include the AdaBoost.MH[12] which is the extension of AdaBoost, the extensions of several neural network algorithms,[13,14,15] the extension of SVM,[16] the extensions of Naïve Bayesian,[17,18] the extension of Maximum Entropy,[19,20] the extension of decision tree,[21] the extensions of some lazy-style algorithms[22,23,24] etc. Recently, several algorithms have also been proposed to improve the performance of learning systems through explore additional information provided by the correlation of class labels[25,26,27,28,29,30,31] or unlabeled data.[32,33] Naïve Bayesian classification algorithm, as one of the probabilistic generative model, has natural advantage for multi-label learning problem. McCallum[17] proposed a Bayesian approach to multi-label document classification, where a mixture probabilistic model (one mixture component per category) is assumed to generate each document and the EM[34] algorithm is used to learn the mixture weights and the word distributions in each mixture component. Min-ling Zhang et al. adapted traditional Naïve Bayesian algorithm to deal with multi-label instances and incorporated feature selection mechanisms into classifier to improve its performance.[18] Other probabilistic generative models are also used for multi-label text classification such as parametric mixture models (PMM1, PMM2).[35] Above NB multi-label classification algorithms mainly focused on modeling the relations between words and classes and classification performances are improved to some extent. However, a complex model always accompanies highly computational costs. Our work aims at classifying text search results real-time by using multi-label classifiers with low computational complexity. Consequently, a kind of less complex feature selection strategy is more proper for this task.

## 3. Naïve Bayesian Multi-label Classification Approach

Naïve Bayesian (NB) algorithm is efficient. For the real-time system, it has a wide application such as Spam filtering system. For the task of multi-label classification, classic NB classifier should be adapted to multi-label data. In addition, there is a factor that may negatively affect its performance. NB classifier has the assumption of class conditional independence, i.e. the effect of a feature value on a given class is independent of the values of other features. In real world applications, however, this assumption does not usually hold which in turn may deteriorate the learning performance.[18]

In this part, the NB classifier adapted for multi-label text classification is designed. The two-step feature selection strategy is proposed to obtain the features as independent as possible.

### 3.1. *Naïve Bayesian Multi-Label (NBML) classifier for text classification*

For a random document $d \in D$ associated with label set $Y \subseteq L$ , $|L|$ binary NB classifiers $H_l : d \rightarrow \{l, \neg l\}$ are learned, one for each different label $l$ in $L$ . That is, the original data set is transformed into $|L|$ data sets $D_l$ that contain all examples of original data set, labeled as $l$ if the labels of the original example contained $l$ and as $\neg l$ otherwise. In this case, the multi-label classification problem is transformed into the combination of several single-label NB classifiers.

For the classification of a new document $d$ , this method outputs a set of labels which is the union of the labels that are output by the $|L|$ classifiers.

$$H(d) = \bigcup_{l \in L} \{l : H_l(d) = l\} \tag{1}$$

Each single-label NB classifier works according to the classic NB theory which is described as follows.

For a random document $d_j$ , its feature is $(a_1, a_2, ...a_m)$ , here $a_k$ is $k$th feature in document $d_j$ . Document label set is $L = \{l_1, l_2, ..., l_n\}$ . The conditional probability of document $d_j$ with relate to each class label $P(l_i|d_j)$ is defined as follows.

$$P(l_i|d_j) = \frac{P(l_i)P(d_j|l_i)}{P(d_j)} \tag{2}$$

Because $P(l_i)$ does not change the result, it can be ignored. $P(d_j|l_i)$ can be obtained from following formula.

$$P(d_j|l_i) \approx \prod_{k=1}^{m} P(a_k|l_i) \tag{3}$$

Here, $P(l_i)$ and $P(a_k|l_i)$ can be estimated according to following formulas.

$$\overset{\wedge}{P}(L = l_i) = \frac{N_i}{N} \tag{4}$$

$$\overset{\wedge}{P}(a_k|l_i) = \frac{1 + N_{ki}}{m + \sum\limits_{k=1}^{m} N_{ki}} \tag{5}$$

Here, $N_i$ is the amount of texts having the label $l_i$. $N_{ki}$ is the total frequency of word $a_k$ appearing in documents in category $l_i$. For single-label classifier, the predicted category of document $d_j$ is the maximum of probability of these categories. In the case of multi-label classification, we use a parameter $P_{thres}$ to represent the average of posterior probability of document $d_j$ in each class as follows.

$$P_{thres} = \frac{1}{n}\sum_{i=1}^{n} P(l_i|d_j) \tag{6}$$

When $P(l_i|d_j) \geq P_{thres}$, we consider that $d_j$ has label $l_i$. In this strategy, new document $d$ has all labels which satisfy $P(l|d) \geq P_{thres}$. The Eq.(1) could be described as follows.

$$H(d) = \bigcup_{l \in L} \{l : P(l|d) \geq P_{thres}\} \tag{7}$$

The least amount of labels of a document is "1" when the probability of the document belonging to a class is obviously higher than that of the document belonging to other classes. The maximum amount of labels of a document is "$n$" when the probabilities of the document belonging to each class are equal.

### 3.2. *Two-step feature selection strategy*

Feature selection is a space reduction method which attempts to select the more discriminative features from preprocessed documents in order to improve classification quality and reduce computational complexity. As many words are extracted from documents, we remove stop list words and then perform two-step feature selection. Firstly, discriminative features are selected based on DF (Document Frequency) and $\chi^2$ statistical analysis.[36] Then FCBF (Fast Correlation-Based Filter Solution) algorithm proposed by Yu and Liu[37] is conducted to filter relevant and redundant features among classes.

#### 3.2.1. *The first step feature selection*

The number of words in each class is great. However, most of them occur only one or two times. This kind of words is not representative for a class. It is necessary to cancel these words in a class before further feature selection. We adopt DF to filter this kind of features, which is defined as follows.

In a corpus $D$, each text belongs to a class set $Y$. Here, $Y \subseteq L$ , $L = \{l_1, l_2, ..., l_n\}$ is the class set defined before classification. Relative text frequency is:

$$Text\_freq\_relative_{ij} = \frac{Text\_freq_{ij}}{N_i} \tag{8}$$

Here, $N_i$ is the quantity of texts with label $l_i$ in training set. $Text\_freq_{ij}$ is the number of texts with label $l_i$ which include word $j$. Algorithm 1 is designed to filter rare features within a class.

**Algorithm 1.**

```
Begin
For $l_i \in L$, $L = \{l_1, l_2...l_i...l_n\}$,
$Term'_i = \emptyset$, $Term = \emptyset$;
For $word_j \in Term_i$,
If $(Text\_freq\_relative_{ij} < \alpha)$
remove $word_j$;
Else
$word_j \in Term'_i$.
$Term = \{Term'_1, Term'_2...Term'_i...Term'_n\}$.
End.
```

Here, $Term_i$ includes all the words extracted in the documents with label $l_i$ , $Term'_i$ includes all the words selected in documents with label $l_i$ and $Term$ is word set selected in training set by Algorithm 1.

We construct "feature by class label" matrix (noted as $Matrix_{cf}$ ) to select discriminative features. In $Matrix_{cf}$, each feature (word) "$j$" is assigned a numerical score based on its occurrence within the different document labels "$l_i$". The choice of the scoring method in this work is the $\chi^2$ test. There are many other tests available as summarized by Sebastiani.[38] However, the $\chi^2$ is often cited as one of the best methods for feature selection. The score of word "$k$" is:

$$\sum_i \frac{(O_{ik} - E_{ik})^2}{E_{ik}} \tag{9}$$

Where $O_{ik}$ is the frequency of word $k$ in documents with label $l_i$. $E_{ik}$ represent the expectation value in the hypothesis of independence of classes and features.

$$E_{ik} = \frac{O_{i+} \times O_{+k}}{O_{++}} \tag{10}$$

Here, $O_{i+}$ is the total word frequency on label $l_i$ and $O_{+k}$ is the accumulation of occurrence of word $k$ in training set.

### 3.2.2. *The second step feature selection*

FCBF algorithm proposed a fast feature filter method which could identify relevant features as well as redundancy among relevant features without pair wise correlation analysis.[37] In this algorithm, feature dimensionality is reduced dramatically by introducing a novel concept, predominant correlation based on symmetrical uncertainty. Here, FCBF algorithm is conducted on the feature selection results obtained in the first step. The first step of the above algorithm has a linear time complexity in terms of the number of features $M$. The time complexity of second step algorithm FCBF is $O\,(NMlogM)$, where, $N$ is the number of instances.[37]

## 4. Experiments and Discussions

As reviewed in Section 2, there have been several approaches to solving multi-label problems. In this paper, our algorithm NBML is compared with Adaboost.MH,[21] multi-label $K$ nearest neighbor (ML-kNN)[24] and the multi-label kernel method Rank-SVM,[16] which are all multi-label learning algorithms applicable to various multi-label problems including text classification. We would not compare our method to the NB multi-label classification algorithm proposed respectively by A. McCallum[17] and Min-Ling Zhang[18] because they have high computational complexity which is improper for the application of Web search results visualization.

For the sake of well evaluating our classification algorithm, our experiments are conducted on an public English corpus coming from NTT communication science research group, which contains real Web pages linked from the "yahoo.com" domain, where it consists of 14 top-level categories (i.e. "Arts & Humanities", "Business & Economy", etc.) and each category is classified into a number of second-level subcategories [c]. By focusing on the second-level categories, we selected 11 out of the 14 independent text categorization problems. For each problem, the training set contains 2,000 documents while the test set contains 3,000 documents.

We adopt various evaluation metrics such as Hamming Loss, One-error, Coverage and Average Precision. The definitions of these metrics could be found in Ref.12. Tables 1-4 show the comparisons on various metrics between our NBML and AdaBoost.MH, ML-kNN and Rank-SVM algorithms, where the best result on each corpus is in bold face. The experiment results of AdaBoost.MH, ML-kNN and RANK-SVM are cited.[24] Classification algorithm of NBML is performed on the free platform WEKA.[39] Ten-fold cross-validation is carried out on data set. In the first step of feature selection, we choose $\alpha = 0.02$.

Table 1. Comparison among different algorithms on *Hamming loss*

| Data Set | NBML | ML-kNN | AdaBoost.MH | Rank-SVM |
|---|---|---|---|---|
| Arts & Humanities | 0.0605 | 0.0612 | **0.0585** | 0.0615 |
| Business & Economy | **0.0253** | 0.0269 | 0.0279 | 0.0275 |
| Computers & Internet | **0.0373** | 0.0412 | 0.0396 | 0.0392 |
| Education | 0.0455 | **0.0387** | 0.0423 | 0.0398 |
| Entertainment | **0.0544** | 0.0604 | 0.0578 | 0.0630 |
| Health | 0.0433 | 0.0458 | **0.0397** | 0.0423 |
| Recreation & sports | 0.0610 | 0.0620 | **0.0584** | 0.0605 |
| Reference | 0.0314 | 0.0314 | **0.0293** | 0.0300 |
| Science | 0.0336 | **0.0325** | 0.0344 | 0.0340 |
| Social & Science | 0.0241 | **0.0218** | 0.0234 | 0.0242 |
| Society & Culture | 0.0553 | **0.0537** | 0.0575 | 0.0555 |
| Average | 0.0429 | 0.0432 | **0.0426** | 0.0434 |

[c] Data set available at http://www.kecl.ntt.co.jp/as/members/ueda/yahoo.tar.gz.

Table 1 shows the comparison of Hamming Loss between NBML and other algorithms. We could find that AdaBoost.MH gives the best performance, ML-kNN and NBML follow and Rank-SVM gives the worst. Hamming Loss evaluates how many times an instance-label pair is misclassified, i.e. a label not belonging to the instance is predicted or a label belonging to the instance is not predicted. AdaBoost.MH is designed to minimize Hamming Loss, thus, it could give better prediction on instance-label pair.

Table 2. Comparison among different algorithms on *One-error*

| Data Set | NBML | ML-kNN | AdaBoost.MH | Rank-SVM |
|---|---|---|---|---|
| Arts & Humanities | 0.5756 | 0.6330 | **0.5617** | 0.6653 |
| Business & Economy | 0.1420 | **0.1213** | 0.1337 | 0.1237 |
| Computers & Internet | 0.4453 | 0.4357 | 0.4613 | **0.4037** |
| Education | 0.5105 | 0.5207 | 0.5753 | **0.4937** |
| Entertainment | 0.5209 | 0.5300 | 0.4940 | **0.4933** |
| Health | 0.3571 | 0.4190 | 0.3470 | **0.3323** |
| Recreation & sports | **0.5517** | 0.7057 | 0.5547 | 0.5627 |
| Reference | 0.4776 | 0.4730 | 0.4840 | **0.4323** |
| Science | 0.5731 | 0.5810 | 0.6170 | **0.5523** |
| Social & Science | 0.3572 | **0.3270** | 0.3600 | 0.3550 |
| Society & Culture | 0.4502 | 0.4357 | 0.4845 | **0.4270** |
| Average | 0.4510 | 0.4711 | 0.4612 | **0.4401** |

Table 2 shows the comparison of One-error between NBML and other algorithms. We could find that Rank-SVM gives the best performance, NBML gives the second best, AdaBoost.MH follows and ML-kNN gives the worst. One-error evaluates how many times the top-ranked label is not in the set of proper labels of the instance. This indicates that NBML could give good prediction on the top-ranked label of instance.

Table 3. Comparison among different algorithms on *coverage*

| Data Set | NBML | ML-kNN | AdaBoost.MH | Rank-SVM |
|---|---|---|---|---|
| Arts & Humanities | 5.3471 | 5.4313 | **5.1900** | 9.2723 |
| Business & Economy | 2.3210 | **2.1840** | 2.4730 | 3.3637 |
| Computers & Internet | **4.4279** | 4.4117 | 4.4747 | 8.7910 |
| Education | 3.6840 | **3.4973** | 3.9663 | 8.9560 |
| Entertainment | 3.1015 | 3.1467 | **3.0877** | 6.5210 |
| Health | **3.0266** | 3.3043 | 3.0843 | 5.5400 |
| Recreation & sports | 4.3838 | 5.1010 | **4.3380** | 5.6680 |
| Reference | **3.2022** | 3.5420 | 3.2643 | 6.9683 |
| Science | 6.4053 | **6.0470** | 6.6027 | 12.4010 |
| Social & Science | 3.8422 | **3.0340** | 3.4820 | 8.2177 |
| Society & Culture | 5.8794 | 5.3653 | **4.9545** | 6.8837 |
| Average | 4.1474 | 4.0968 | **4.0834** | 7.5075 |

Table 3 shows the comparison of Coverage between NBML and other algorithms. We could find that ML-kNN and AdaBoost.MH give the best performance; Results of NBML are very close to the best performance. Rank-SVM gives the worst. Coverage evaluates how far we need, on the average, to go down the list of labels in order to cover all the proper labels of the instance. It is loosely related to precision at the level of perfect recall. This indicates that NBML perform well on recall rate.

Table 4 shows the comparison of Average Precision between NBML and other algorithms. We could find that NBML gives the best performance, AdaBoost.MH and ML-kNN follows and Rank-SVM gives the worst. Average Precision evaluates the average fraction of labels ranked above a particular label $l \in Y$, which actually is in $Y$. This indicates that NBML could give predictions which have the smallest deviation from the true label set of instance.

Table 4. Comparison among different algorithms on *Average precision*

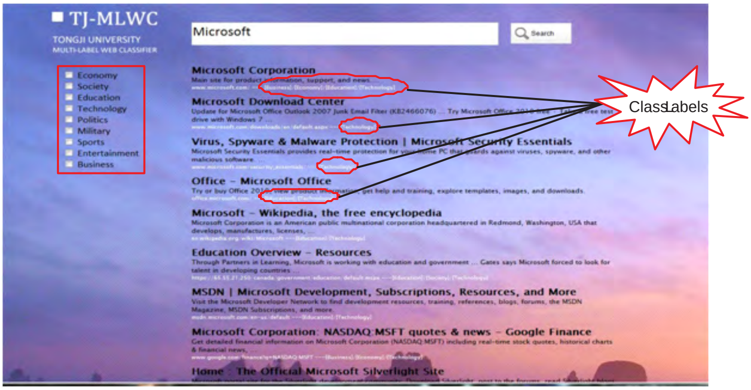
| Data Set | NBML | ML-kNN | AdaBoost.MH | Rank-SVM |
|---|---|---|---|---|
| Arts & Humanities | 0.5391 | 0.5097 | **0.5526** | 0.4170 |
| Business & Economy | **0.8819** | 0.8798 | 0.8702 | 0.8694 |
| Computers & Internet | 0.6205 | **0.6338** | 0.6235 | 0.6123 |
| Education | 0.5882 | **0.5993** | 0.5619 | 0.5702 |
| Entertainment | **0.6503** | 0.6013 | 0.6221 | 0.5637 |
| Health | **0.7355** | 0.6817 | 0.7257 | 0.6839 |
| Recreation & sports | 0.5535 | 0.4552 | **0.5639** | 0.5315 |
| Reference | **0.6309** | 0.6194 | 0.6264 | 0.6176 |
| Science | 0.5234 | **0.5324** | 0.4940 | 0.5007 |
| Social & Science | 0.7309 | **0.7481** | 0.7217 | 0.6788 |
| Society & Culture | 0.6057 | **0.6128** | 0.5881 | 0.5717 |
| Average | **0.6418** | 0.6249 | 0.6318 | 0.6015 |

We summarize above conclusions: Algorithm NBML proposed in this paper has nearly performance comparing to the famous algorithms (i.e. AdaBoost.MH and ML-kNN) in the application of multi-label text classification.

## 5. Prototype System TJ-MLWC

This section presents an example of visualizing text search results by using multi-label classification methodology, named TJ-MLWC2.0 (The version TJ-MLWC1.0 was designed for Chinese text search visualization.). It calls the API of search engine, performs NB multi-label classification on its search results and provides user the visualized search results by system interface. The classifiers are trained off-line based on part of the multi-label corpus referred in our experiments.

A TJ-MLWC session starts with the user entering her query in the query box. For example, the user inputs "Microsoft" and then the system returns the interface as shown in Fig.1. Here, the labels in left box are the categories pre-defined. In the right, there are all entries returned by search engine. We could choose one or a few categories, allowing the system to return the entries in specified categories. Each entry in the right, which different from the displaying form of current prevalent search engine, has one or more class labels added by system.



**Microsoft Corporation**
Main site for product information, support, and news. ...
www.microsoft.com/ ---[Business] [Economy] [Education] [Technology]

**Microsoft Download Center**
Update for Microsoft Office Outlook 2007 Junk Email Filter (KB2466076) ... Try Microsoft Office 2010 free - Take a free test drive with Windows 7 ...
www.microsoft.com/downloads/en/default.aspx ---[Technology]

**Virus, Spyware & Malware Protection | Microsoft Security Essentials**
Microsoft Security Essentials provides real-time protection for your home PC that guards against viruses, spyware, and other malicious software. ...
www.microsoft.com/security_essentials/ ---[Technology]

**Office - Microsoft Office**
Try or buy Office 2010, view product information, get help and training, explore templates, images, and downloads.
office.microsoft.com/ ---[Education] [Technology]

**Microsoft - Wikipedia, the free encyclopedia**
Microsoft Corporation is an American public multinational corporation headquartered in Redmond, Washington, USA that develops, manufactures, licenses, ...
en.wikipedia.org/wiki/Microsoft ---[Education] [Technology]

**Education Overview - Resources**
Through Partners in Learning, Microsoft is working with education and government ... Gates says Microsoft forced to look for talent in developing countries ...
https://65.55.21.250/canada/government/education/default.mspx ---[Education] [Society] [Technology]

**MSDN | Microsoft Development, Subscriptions, Resources, and More**
Visit the Microsoft Developer Network to find development resources, training, references, blogs, forums, the MSDN Magazine, MSDN Subscriptions, and more.
msdn.microsoft.com/en-us/default ---[Education] [Technology]

**Microsoft Corporation: NASDAQ:MSFT quotes & news - Google Finance**
Get detailed financial information on Microsoft Corporation (NASDAQ:MSFT) including real-time stock quotes, historical charts & financial news, ...
www.google.com/finance?q=NASDAQ:MSFT ---[Business] [Economy] [Technology]

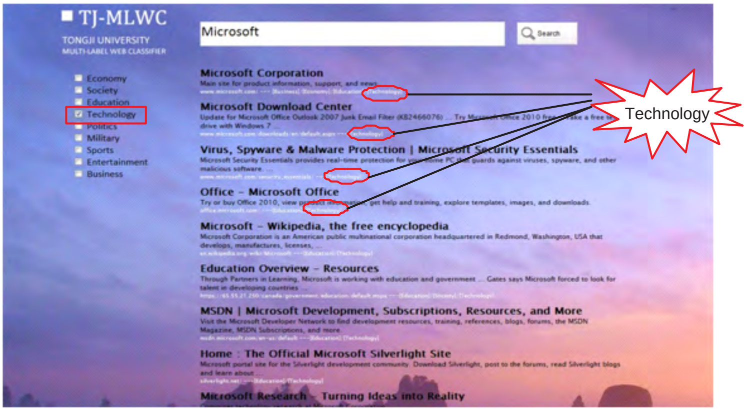
**Home : The Official Microsoft Silverlight Site**
Microsoft portal site for the Silverlight development community. Download Silverlight, post to the forums, read Silverlight blogs

Fig. 1. All search results of “Microsoft”

Fig.2 and Fig.3 show the results of query “Microsoft” by choosing the class label “Technology”, “Education” separately. The system returns all the entries involving class label “Technology” in Fig.2 and involving class label “Education” in Fig.3. Fig.4 shows the results of involving class label “Technology” and “Education” in the same time. From the figures, we could find that there are some entries belong to both “Technology” and “Education”. For example, “Microsoft - Wikipedia, the free encyclopedia”, “Education Overview - Resources” appear in Fig.2 and Fig.3 simultaneously. When we confine the class label as “Technology + Education”, the results will be limited to the scope of entries with the two labels as shown in Fig.4.

From the above figures, we could find that a document could appear in more than one class in our system, which is different from the search engine visualization system based on clustering algorithms. Theoretically, if a document concerns the contents in several classes, we could find it in all these classes. However, search engine based on clustering is more likely to assign a document in a certain cluster. Usually, the label for a cluster obtained by clustering algorithm is hard to reflect the content of documents in its cluster properly. In this case, the cluster labels are not very effective for user’s locating. In TJ-MLWC, class labels are pre-defined to avoid this problem. Moreover, multiple labels ensure that a document could always be found for users having different interests.

Through testing, we found that there are still some problems to be solved in system TJ-MLWC. Firstly, it mixes the text search results and other kind of search results such as image, video and audio. Further works will focus on constructing a hierarchical classifier which classifies files according to their formation first and then



TJ-MLWC

TONGJI UNIVERSITY
MULTI-LABEL WEB CLASSIFIER

☐ Economy
☐ Society
☐ Education
☑ Technology
☐ Politics
☐ Military
☐ Sports
☐ Entertainment
☐ Business

Microsoft

**Microsoft Corporation**
Main site for product information, support, and news.
www.microsoft.com/ ---[Business] [Economy] [Education] [Technology]

**Microsoft Download Center**
Update for Microsoft Office Outlook 2007 Junk Email Filter (KB2466076) ... Try Microsoft Office 2010 free ... Take a free test drive with Windows 7 ...
www.microsoft.com/downloads/en/default.aspx ---[Technology]

**Virus, Spyware & Malware Protection | Microsoft Security Essentials**
Microsoft Security Essentials provides real-time protection for your home PC that guards against viruses, spyware, and other malicious software. ...
www.microsoft.com/security_essentials/ ---[Technology]

**Office – Microsoft Office**
Try or buy Office 2010, view product information, get help and training, explore templates, images, and downloads.
office.microsoft.com/ ---[Education] [Technology]

**Microsoft – Wikipedia, the free encyclopedia**
Microsoft Corporation is an American public multinational corporation headquartered in Redmond, Washington, USA that develops, manufactures, licenses, ...
en.wikipedia.org/wiki/Microsoft ---[Education] [Technology]

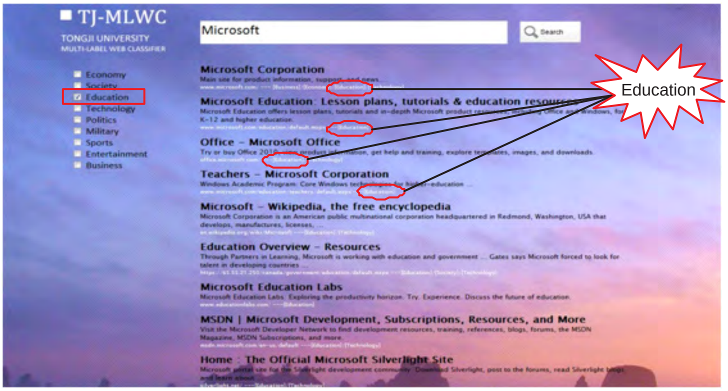
**Education Overview – Resources**
Through Partners in Learning, Microsoft is working with education and government ... Gates says Microsoft forced to look for talent in developing countries ...
https://65.55.21.250/canada/government/education/default.mspx ---[Education] [Society] [Technology]

**MSDN | Microsoft Development, Subscriptions, Resources, and More**
Visit the Microsoft Developer Network to find development resources, training, references, blogs, forums, the MSDN Magazine, MSDN Subscriptions, and more.
msdn.microsoft.com/en-us/default ---[Education] [Technology]

**Home : The Official Microsoft Silverlight Site**
Microsoft portal site for the Silverlight development community. Download Silverlight, post to the forums, read Silverlight blogs and learn about ...
silverlight.net/ ---[Education] [Technology]

**Microsoft Research – Turning Ideas into Reality**

Fig. 2. Search results of "Microsoft" when selecting "Technology" class



TJ-MLWC

TONGJI UNIVERSITY
MULTI-LABEL WEB CLASSIFIER

☐ Economy
☐ Society
☑ Education
☐ Technology
☐ Politics
☐ Military
☐ Sports
☐ Entertainment
☐ Business

Microsoft

**Microsoft Corporation**
Main site for product information, support, and news. ...
www.microsoft.com/ ---[Business] [Economy] [Education] [Technology]

**Microsoft Education: Lesson plans, tutorials & education resources**
Microsoft Education offers lesson plans, tutorials and in-depth Microsoft product resources, including Office and Windows, for K-12 and higher education.
www.microsoft.com/education/default.mspx ---[Education]

**Office – Microsoft Office**
Try or buy Office 2010, view product information, get help and training, explore templates, images, and downloads.
office.microsoft.com/ ---[Education] [Technology]

**Teachers – Microsoft Corporation**
Windows Academic Program: Core Windows technologies for higher-education ...
www.microsoft.com/education/teachers/default.mspx ---[Education]

**Microsoft – Wikipedia, the free encyclopedia**
Microsoft Corporation is an American public multinational corporation headquartered in Redmond, Washington, USA that develops, manufactures, licenses, ...
en.wikipedia.org/wiki/Microsoft ---[Education] [Technology]

**Education Overview – Resources**
Through Partners in Learning, Microsoft is working with education and government ... Gates says Microsoft forced to look for talent in developing countries ...
https://65.55.21.250/canada/government/education/default.mspx ---[Education] [Society] [Technology]

**Microsoft Education Labs**
Microsoft Education Labs: Exploring the productivity horizon. Try. Experience. Discuss the future of education.
www.educationlabs.com/ ---[Education]

**MSDN | Microsoft Development, Subscriptions, Resources, and More**
Visit the Microsoft Developer Network to find development resources, training, references, blogs, forums, the MSDN Magazine, MSDN Subscriptions, and more.
msdn.microsoft.com/en-us/default ---[Education] [Technology]

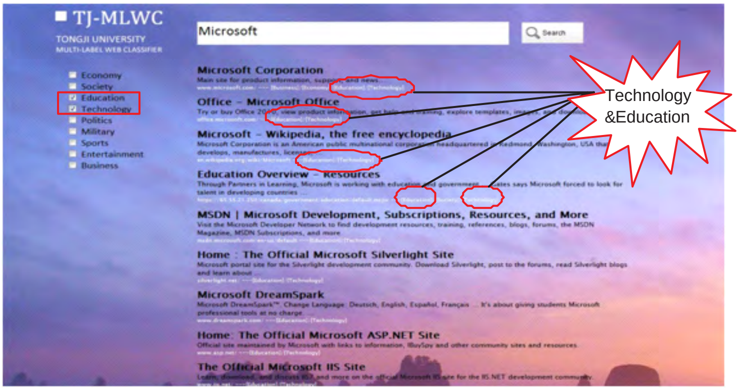
**Home : The Official Microsoft Silverlight Site**
Microsoft portal site for the Silverlight development community. Download Silverlight, post to the forums, read Silverlight blogs and learn about ...
silverlight.net/ ---[Education] [Technology]

Fig. 3. Search results of "Microsoft" when selecting "Education" class

conducts multi-label classification on each kind of files. Secondly, we pre-defined the taxonomy for search results. However, with gradually rectification of Web contents, documents in new classes will always appear. How can we rectify the taxonomy automatically according to real demand? All of these problems are our future research objectives.

## 6. Conclusions

This paper presents a new multi-label classification method based on Naïve Bayesian theory. A two-step feature selection strategy based on DF, $\chi^2$ and FCBF algorithm is incorporated into the method to improve classification performance. Experiments on public multi-label corpus show that our method achieves highly competitive



TJ-MLWC
TONGJI UNIVERSITY
MULTI-LABEL WEB CLASSIFIER

- □ Economy
- □ Society
- ☑ Education
- ☑ Technology
- □ Politics
- □ Military
- □ Sports
- □ Entertainment
- □ Business

Microsoft

Search

**Microsoft Corporation**
Main site for product information, support, and news.
www.microsoft.com ---- [Business] [Economy] [Education] [Technology]

**Office – Microsoft Office**
Try or buy Office 2010, view product information, get help and training, explore templates, images, and downloads
office.microsoft.com ---- [Education] [Technology]

**Microsoft – Wikipedia, the free encyclopedia**
Microsoft Corporation is an American public multinational corporation headquartered in Redmond, Washington, USA that develops, manufactures, licenses ...
en.wikipedia.org/wiki/Microsoft ---- [Education] [Technology]

**Education Overview – Resources**
Through Partners in Learning, Microsoft is working with education and government ... Gates says Microsoft forced to look for talent in developing countries ...
[illegible] ---- [Education] [Society] [Technology]

**MSDN | Microsoft Development, Subscriptions, Resources, and More**
Visit the Microsoft Developer Network to find development resources, training, references, blogs, forums, the MSDN Magazine, MSDN Subscriptions, and more
msdn.microsoft.com/en-us/default ---- [Education] [Technology]

**Home : The Official Microsoft Silverlight Site**
Microsoft portal site for the Silverlight development community. Download Silverlight, post to the forums, read Silverlight blogs and learn about ...
silverlight.net/ ---- [Education] [Technology]

**Microsoft DreamSpark**
Microsoft DreamSpark™. Change Language: Deutsch, English, Español, Français ... It's about giving students Microsoft professional tools at no charge. ...
www.dreamspark.com/ ---- [Education] [Technology]

**Home: The Official Microsoft ASP.NET Site**
Official site maintained by Microsoft with links to information, IBuySpy and other community sites and resources.
www.asp.net/ ---- [Education] [Technology]

**The Official Microsoft IIS Site**
Learn, download, and discuss IIS7 and more on the official Microsoft IIS site for the IIS.NET development community.
www.iis.net/ ---- [Education] [Technology]

Fig. 4. Search results of “Microsoft” when selecting “Technology” and “Education” class

performance with several famous multi-label classification algorithms. Because our method embedded high efficiency feature selection algorithm, it has low computation costs which allows it to be used for the on-line application, especially for handling high-dimensionality text data. However, our NBML algorithm does not concern the correlation between labels while the dependency among labels exists widely in real applications. Consequently, the improvement of NB multi-label classification itself is our future research issue.

A Web search results visualizing prototype system TJ-MLWC is developed in this paper. Testing results indicate that this system has the function of browsing search results by category. It could also avoid the problem of search result visualization system based on clustering methods. However, with the growing of new contents on Web, the pre-defined taxonomy is hard to extend automatically and the classification ability of model trained off-line will be degraded greatly. Semi-supervised multi-label classification method could make use of the unlabeled online data to improve the performance of classifier. Designing semi-supervised MLNB classifier that can handle online learning problem is an interesting problem to be explored.

## Acknowledgments

This paper is sponsored by the National Natural Science Foundation of China (No. 60970061 and No. 61075056) and Shanghai Leading Academic Discipline Project (No.B004).

## References

1. F. Ren, D. B. Bracewell. *Advanced Information Retrieval*, Electr. Notes Theor. Comput. Sci. **225**: pp. 303-317, 2009.

2. D. Roussinov, H. Chen. *Information navigation on the web by clustering and summarizing query results*, Inf. Process. Manage. (IPM) **37**(6), pp. 789-16, 2001.
3. G. Mecca, S. Raunich, A. Pappalardo. *A new algorithm for clustering search results*, Data Knowl. Eng. **62**(3), pp. 504-522, 2007.
4. N. L. Beebe and Jan Guynes Clark. *Digital Forensic Text String Searching: Improving Information Retrieval Effectiveness by Thematically Clustering Search Results*, Digital Investigation. **4**(2), pp. 49-54, 2007.
5. O. Zamir, O. Etzioni. *Grouper: A Dynamic Clustering Interface to Web Search Results*, Computer Networks. **31**(11-16), pp. 1361-1374, 1999.
6. D. Miao, G. Wang and Q. Liu. *Granular Computing: Past, Present and future*, Science Publisher, 2007.
7. J.Cugini, S.Laskowski, & M.Sebrechts. Design of 3D visualization of search results: Evolution and evaluation. In *Proceedings of IST/SPIE's 12th annual international symposium-electronic imaging 2000: Visual data exploration and analysis*, pp. 23-28, 2000.
8. D. Newman, et al.. *Visualizing search results and document collections using topic maps*, Web Semantics: Sci. Serv. Agents World Wide Web, **8**(2-3), pp. 169-175, 2010.
9. I.S. Dhillon, D.S. Modha. *Concept decompositions for large sparse text data using clustering*, Machine Learning, **42**(1/2), pp. 143-175, 2001.
10. T.L. Griffiths, M. Steyvers, J.B.T. Tenenbaum. *Topics in semantic representation*, Psychological Review, **114**(2), pp. 211-244, 2007.
11. G. Tsoumakas, I. Katakis. *Multi-Label Classification: An Overview*, International Journal of Data Warehousing and Mining, **3**(3), pp. 1-13, 2007.
12. R.E. Schapire, Y. Singer. *Boostexter: a boosting-based system for text categorization*, Machine Learning, **39**(2/3), pp. 135-168, 2000.
13. K. Crammer, Y. Singer. A new family of online algorithms for category ranking, In *Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 151-158, 2002.
14. M.-L. Zhang, Z.-H. Zhou. *Multilabel neural networks with applications to functional genomics and text categorization*, IEEE Transactions on Knowledge and Data Engineering, **18**(10), pp. 1338-1351, 2006.
15. M.-L. Zhang. *ML-RBF: RBF neural networks for multi-label learning*, Neural Processing Letters, **29**(2), pp. 61-74, 2009.
16. A. Elisseeff, J. Weston. *A kernel method for multi-labelled classification*, Advances in Neural-Information Processing Systems, **14**, MIT Press, Cambridge, MA, pp. 681-687, 2002.
17. A. McCallum. Multi-label text classification with a mixture model trained by EM. In *Working Notes of the AAAI'99 Workshop on Text Learning*, pp. 1-7, 1999.
18. M.-L. Zhang, J. M. Peña, V. Robles. *Feature selection for multi-label Naïve bayes classification*, Information Sciences, **179**(19), pp. 3218-3229, 2009.
19. K. Nigam, J. Lafferty, A. McCallum. Using maximum entropy for text classification, In *IJCAI-99 Workshop on Machine Learning for Information Filtering*, pp. 61-67, 1999.
20. Y. Wu, F. Ren. A Corpus-based Multi-label Emotion Classification using Maximum Entropy. In *The 6th International Workshop on Natural Language Processing and Cognitive Science*, pp. 103-110, 2009.
21. F.D. Comite, R. Gilleron, M. Tommasi. Learning multi-label alternating decision tree from texts and data, *Lecture Notes in Computer Science:* **2734**, Springer, Berlin, pp. 35-49, 2003.
22. K. Brinker, E. Hllermeier. Case-based multilabel ranking, In *Proceedings of the 20th International Joint Conference on Artificial Intelligence*, pp. 702-707, 2007.
23. F. Kang, R. Jin, R. Sukthankar. Correlated label propagation with application to multi-label learning, In *Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 1719-1726, 2006.
24. M.-L. Zhang, Z.-H. Zhou. *Ml-knn a lazy learning approach to multi-label learning*, Pattern Recognition, **40**(7), pp. 2038-2048, 2007.
25. W. Cheng, E. Hüllermeier. *Combining instance-based learning and logistic regression for multilabel classification*, Machine Learning, **76**(2-3), pp. 211-225, 2009.
26. S. Ji, L. Tang, S. Yu, and J. Ye. Extracting shared subspace for multi-label classification, In

*Proceedings of the 14th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 381-389, 2008.

27. R. Yan, J. Tessšić, and J. R. Smith. Model-shared subspace boosting for multi-label classification, In *Proceedings of the 13th ACM SIGKDD Conference onKnowledge Discovery and Data Mining*, pp. 834-843, 2007.
28. J. Read, B. Pfahringer, and G. Holmes. Multi-label classification using ensembles of pruned sets, In *Proceedings of the 9th IEEE International Conference on Data Mining*, pp. 995-1000, 2008.
29. J. Read, B. Pfahringer, G. Holmes, and E. Frank. Classifier chains for multi-label classification, *Lecture Notes in Artificial Intelligence:* **5782**, pp. 254-269, 2009.
30. G. Tsoumakas, I. Vlahavas. Random k-labelsets: an ensemble method for multilabel classification, *Lecture Notes in Artificial Intelligence:* **4701**, pp. 406-417, 2007.
31. M.-L. Zhang, K. Zhang. Multi-label learning by exploiting label dependency. In *Proceedings of the 16th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD'10)*, pp. 999-1007, 2010.
32. G. Chen, Y. Song, F. Wang, C. Zhang. Semi-supervised multi-label learning by solving a sylvester equation, In *Proceedings of the 2008 SIAM International Conference on Data Mining*, pp.410-419, 2008.
33. Y. Liu, R. Jin, L. Yang. Semi-supervised multi-label learning by constrained non-negative matrix factorization, In *Proceedings of the 21st National Conference on Artificial Intelligence*, pp. 421-426, 2006.
34. A.P. Dempster, N.M. Laird, D.B. Rubin. *Maximum likelihood from incomplete data via the EM algorithm*, Journal of the Royal Statistics Society - B, **39**(1), pp. 1-38, 1977
35. N. Ueda, K. Saito. *Parametric mixture models for multi-label text*, Advance in Neural Information Processing Systems,**15**, MIT Press, Cambridge, MA, pp. 721-728, 2003.
36. Z. Wei, D. Miao, et al.. *N-grams based feature selection and text representation for Chinese Text Classification*, International Journal of Computational Intelligence Systems, **2**(4), pp. 365-374, 2009
37. L. Yu, H. Liu. Feature Selection for High-Dimensional Data: A Fast Correlation-Based Filter Solution. In *Proceedings of the Twentieth International Conference on Machine Learning (ICML-2003)*, pp.856-863, 2003.
38. F. Sebastiani. *Machine learning in automated text categorization*, ACM Computing Surveys, **34**(1), pp. 1-47, 2002.
39. M. Hall et al.. *The WEKA Data Mining Software: An Update*, SIGKDD Explorations, **11**(1), 2009.

## Zhihua Wei

She received the Ph.D degree in 2010 from Department of Computer Science and Technology, Tongji University, China. She is currently a lecturer at Tongji University. Her research interests include Text mining, Machine Learning and Natural Language Processing.

**Hongyun Zhang**

She received the Ph.D degree in 2005 from Department of Computer Science and Technology, Tongji University, China. She is currently a lecturer at Tongji University. Her research interests include Rough set theory, Dimensionality reduction and Data mining, Granular Computing and pattern recognition.

**Zhifei Zhang**

He received the B.E degree in 2008 from Department of Computer Science and Technology, Tongji University. He is currently a Ph.D candidate at Tongji University, China. His research interests include Text Mining, Machine Learing, and Natural Language Processing.

**Wen Li**

She received the master degree in 2005 from Computer Software and Theory, Nanchang University, China. He is currently a lecturer at Nanchang University. Her research interests include Text mining, Machine Learning and Granular Computing.

**Duoqian Miao**

He received the Ph.D degree in 1997 from Institute of Automation, Chinese Academy of Science. He is currently a professor at Tongji University. His research interests include machine learning, Rough set theory, Granular Computing and so on.