http://ansinet.com/itj

ITJ

ISSN 1812-5638

INFORMATION TECHNOLOGY JOURNAL

ANSInet
Asian Network for Scientific Information
308 Lasani Town, Sargodha Road, Faisalabad - Pakistan

Information Technology Journal 12 (19): 5044-5051, 2013
ISSN 1812-5638 / DOI: 10.3923/itj.2013.5044.5051

Study of Documents Multi-hierarchy Categorization Based on Topic Label and LSI

[1,3]Li Kairong, [1]Qu Libing, [1,2]Zhu Junwu, [1] Kong Zhaokun and [1]Zhao Dongwei
[1]College of Information Engineering, Yangzhou University, Yangzhou, Jiangsu, 225127, China
[2]Department of Computing and Information Science, University of Guelph, Guelph n1g2k8, Canada
[3]Information Technology Research Base of Civil Aviation Administration of China,
Civil Aviation University of China, Tianjin, 300300, China

Abstract: For hierarchical text classification problem, the existing prototype-based classifiers, such as k-NN, kNN Model and Centroid classifier, have achieved the aim of expected function and performance. However, due to high dimensionality and complex class structures of document data sets, they usually perform less effectively, we proposed a new method for text classification that extracts semantic labels and builds a tree structure for each level of the classification hierarchy. We compare the proposed method with KNN method, using several multi-hierarchical classification datasets. Our experimental analysis shows that our method fully considers the semantic information between the contact hierarchies of the category, as well as enhances the efficiency of text classification.

Key words: Latent semantic index, text categorization, topic label vector space model, entry amplification, gibbs sampling

INTRODUCTION

Currently, the exponential growth of online information of data stimulates a greater demand for text categorization. In recent years, numerous classification techniques have been used widely for text categorization. According to predefined Classification structure composition and the structure of the final output, delimit text categorization algorithm can be divided into two categories: plane Classification (Cohen and Singer, 1996) and Hierarchical Classification (Sun and Lim, 2001; Loh and Shih, 1997; Rastogi and Shim, 1998; Koller and Sahami, 1997).These include decision tree classifiers, Bayes classifiers, Support Vector Machine (SVM), neural network classifiers, boosting methods and prototype-based classifiers.

Classified the nature of the text is through a specific hierarchy of tree structure to represent the full text categories, that is, through the tree structure could be divided into and tree classification task level corresponding to the classification of smaller sub-problems. So, after the division of simplified equivalent to the original classification task, short the time of classification processing operations. Using the hierarchical classification method due to the classification problem resolve into several sub problems, each sub problem with just a few words can say, this also makes in each sub-problem, the meaning of vocabulary is limited, so as to control the polysemy from happening, effectively improve the classification accuracy and simplify the classification problem.Text hierarchical classification in information retrieval and text mining, showed many advantages, has been achieved good classification effect (Bade *et al*., 2006; Liu and Dong, 2001; Wang and Gai, 2004; Yao and Wu, 2004; Xu, 2004; Song, 2004). Silla Jr. and Freitas (2010) took a survey of hierarchical classification across different application domains and summarized the research achievements of recent years. Vens *et al.* (2008) constructed decision trees for hierarchical multi-label classification to settle the hierarchical problems. Cerri and de Carvalho (2010) improved efficiency of multi-hierarchical classification using top-down label combination and artificial neural networks. Mayne and Perry (2009) proposed a local-based method and employed it in the document classification task. (Cerri and de Carvalho, 2010) put forward two new methods following the local approach which is based on multi-label non-hierarchical classification methods.

However, In most cases a single document category actually contains multiple subtopics, indicating that the documents in the same class may comprise multiple subclasses, each associated with its individual term subspace. There is little research on text classification based on topics of category. Semantic information using

Corresponding Author: Li Kairong, College of Information Engineering, Yangzhou University, Yangzhou, Jiangsu, 225127, China

the topic label, not only express the semantic information of a category but also clearly distinguish the categories with a topic label marking which is conducive to the semantic information of document type. These ways of hierarchical organization help us to accept the new knowledge and to maintain the existing knowledge. Using hierarchical subject classification model is better to simulate the organization form of knowledge in the brain, by subject categories in the classification tree structure of the organization more accord with people the way of learning.

In information retrieval, LSI is proposed in order to solve synonyms, synonyms and words noise problems, when it is introduced into the field of text classification, it also has good classification performance and gets widely application. (Zelikovitz and Hirsh, 2001) combine not classified documents to the training set, LSI model is set up on the document in the set. The similarity between test documents and training set is calculated by the model. Liu *et al.* (2004) propose a local latent semantic indexing method (local LSI), by applying the latent semantic indexing method to the entry-document matrix which is related to the topic. Sun and Lim (2001) select basic solution vectors by means of iterative which are used to represent each class of LSI but this method requires a lot of computation. In the study, we first extract a series of topics using Gibbs sampling in a category, take the topics with probabilities of sampling as semantic label and then use the topic label to improve the performance of LSI (Latent Semantic Indexing).

TOPIC EXTRACTION

Expression of topic: As a whole, text is made up of a series of related words in certain forms of permutation and combination.In text categorization, the unstructured natural language use vector model or word frequency matrix to represent, making it become a structured mathematical model.

When carrying on the text preprocessing, we usually hypothesized that terms and the terms are independent of each other. Our purpose is to reduce the complexity of the text information processing, at the same time improve the classification effect and performance, because the actual text vectors are so many and the size of the word frequency matrix is very huge. The bag-of-words model is widely used to represent the text in text classification. It is a simplifying assumption that transforms natural language into statistical descriptions of the word. But The bag-of-words model ignores the text structure information.

Vector space model based on bag-of-words, just get textual preventatives information and ignore to consider

Topic 1 ↩		↩	Topic 2 ↩	
Word ↩	Prob. ↩	↩	Word ↩	Prob. ↩
EDU ↩	0.208↩	↩	SPACE ↩	0.042↩
STATE ↩	0.117↩	↩	ACCESS ↩	0.30↩
OHIO ↩	0.101↩	↩	DIGEX ↩	0.029↩
CMU ↩	0.089↩	↩	SIC ↩	0.029↩
CLUB ↩	0.041↩	↩	MISC ↩	0.021↩
NEW ↩	0.022↩	↩	NET ↩	0.018↩
APR ↩	0.013↩	↩	EDU ↩	0.013↩
TALK ↩	0.019↩	↩	PAT ↩	0.013↩
LINE ↩	0.014↩	↩	TALK ↩	0.012↩
PATH ↩	0.005↩	↩	MISSION ↩	0.012↩
...↩	...↩	↩	...↩↩	...↩↩

Fig. 1: Probability of two topics

the hidden semantic information in classifying. When we face a large number of documents of the same category, the first thought is the focus (namely the topic) of a category. The semantic information of topic is exactly considered to the probability distribution of some special terms, Fig. 1 shows two topic projects. Referenced by the ideas of artificial classification of experts, when manually classified judgment above all is the topic of category and then depending on the topic to find the best match. This study use Gibbs sampling (Griffiths, T.L. Steyvers, M.) to extract topic of category. Gibbs sampling employs the conditional distribution to build markov transferring nuclear. The algorithm can be used to complex probability distribution sampling and doing statistics.

Probability distribution of topic: When using the topic label, we must first determine the probability distribution of topics. And then in each topic, we will extract all the entries of a new document. The probability formula that entry appeared in the new document is expressed as:

$$P(w) = \sum_{i=1}^{T} P(w \mid z_w = i) P(z_w = i) \quad (1)$$

T represents the number of topics. $P(z_w = i)$ is the probability distribution of topics. $P(w \mid z_w = i)$ represents the probability that entry w appears in the i-th topic

In this study, we assume that there are D document sets, contain T topics, W different entries, with d_i and w_i respectively represents document and entry. Gibbs sampling method is this: First of all, we identify the conditional probability that entries appear in a certain topic. And then figure out the probability of the entry corresponding to each topic, namely, the probability

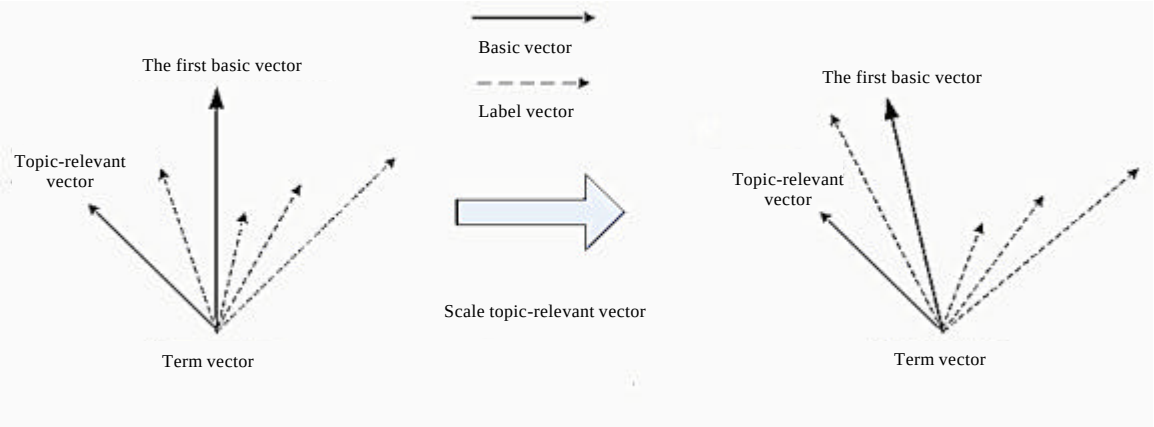

Fig. 2: Algorithm description in this study

distribution of each topic., The conditional distribution is expressed as $P(z_i = j | z_{-i}, w_i, d_i, \bullet)$, $z_i = j$ said entry is assigned to the topic j, z_{-i} is the distribution of other entries and "•" said fixed parameters α and β. Grirrith and Steyvrs decompose $P(z_i = j | z_{-i}, w_i, d_i, \bullet)$ into:

$$P(z_i = j | z_{-i}, w_i, d_i, \bullet) \propto \frac{C_{w_{ij}}^{WT} + \beta}{\sum_{w=1}^{W} C_{w_{ij}}^{WT} + W\beta} \cdot \frac{C_{d_{ij}}^{DT} + \alpha}{\sum_{t=1}^{T} C_{d_{ij}}^{DT} + T\alpha} \quad (2)$$

$C_{w_{ij}}^{WT}$: The number that entry w can be drawn from topic j, is W×T-dimensional matrix; $C_{d_{ij}}^{DT}$: The sum of number that topic j is assigned to all entries of the document d, is D×T-dimensional matrix. The formula is composed of two parts.: the probability that entry w appears in topic j and the probability that topic j appears in document d. The implementation process of Gibbs sampling: First, we randomly assign each entry in a topic, using a markov chain to carry on the multiple iterations, each time by sampling all the topics find a new initial state, after enough times of iteration, the probability reaches a stable state, record the current value. We intercept k entries of the maximum probability as the topic label of category.

MULTI-HIERARCHY CATEGORIZATION COMBINED WITH TOPIC LABEL

LSI: LSI is constructed around the Singular Value Decomposition (SVD) of X, we use the SVD method to obtain a smaller matrix which includes most information of the original semantic space. In LSA, the matrix is approximated by a truncated SVD in which the first k diagonal values are retained but the rest are set to zero. That is:

$$X = U_\lambda \sum_\lambda V_\lambda^T \approx U_k \sum_k V_k^T \quad (3)$$

where, U_λ and V_λ are orthogonal matrices and $\sum_\lambda = \text{diag}(\sigma_1, \sigma_2, \cdots, \sigma_\lambda)$ is diagonal and non-negative. The diagonal values of $\sum_\lambda$ are ordered to be non-increasing. The truncated SVD is the best rank λ approximation to X in the Frobenius norm:

$$s = (XV_k)(qV_k)^T = \left(U_k \sum_k\right)\left(V_k^T q^T\right) \quad (4)$$

LSI topic label of category: When we use the topic labels for the query, If the query entry contains words of topic label, We think that the relationship between the query entry and the topic label is close. In order to make the topic labels get better effect, we amplify the topic label which is a vector consist of many entries:

$$\hat{t} = (1 + q) \cdot t \quad (5)$$

Which t is a entry of the topic label that is close to the query entry. q is a real number and $q \geq 0$. The method also can be used for entry-document matrix.We first use the amplification formula to highlight some rows, which are similar to the query term on the semantic relationsand then we conducted singular value decomposition on obtained new matrix.so, An entry-entry matrix can be obtained by below formula:

$$XX^T = U_K \sum_K^2 U_K^T \quad (6)$$

The matrix describes the degree of similarity between entries.In this study, We selected the k most similar entries combining the topic label, as the final category label. Category label obtained by this method is not only achieve high accuracy but also fully integrated with semantic information.

The proposed method increases the degree of similarity between documents of the same categoryand reduces the degree of similarity between documents of different categories. LSI input is entry-document matrix, conducting singular value decomposition, we get the left singular value vector that represents in the direction of the vector of term, as the basic vector. We amplify the topic labels, significantly enhance the relation between the basic vector and other topic-relevant vectors.We assume that the front label obtained by probability sampling is optimal, after a subsequent LSI treatment, the similarity between documents of the same category has been upgraded and the performance of text categorization could has been enhanced accordingly.

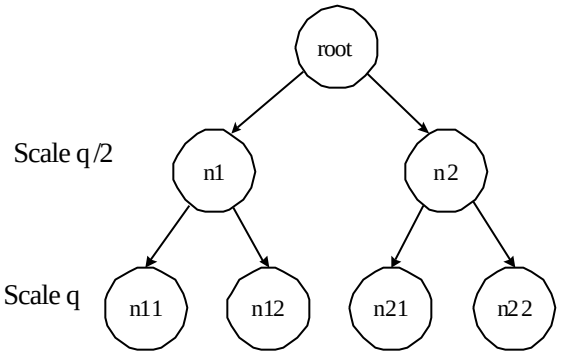

Fig. 3: A simple hierarchy tree

MODEL OF CLASSIFICATION

In this study, We use a tree structure to represent one category. The leaf nodes of the tree are the embodiment of the category, upwards gradually blurred.We assume that amplification parameter of the leaf node is q, then the amplification parameter of a certain node in the hierarchical tree could be expressed as follows:

$$q_i = \frac{q}{c(i)} \tag{7}$$

c(i) stands for the number of leaf nodes that affiliated to current node. If c(i) is the leaf node, then c(i) = 1.

The classification algorithms can be generally described as: First, we use topic label to construct category hierarchy tree, then according to the node amplification formula, obtain amplification parameters of

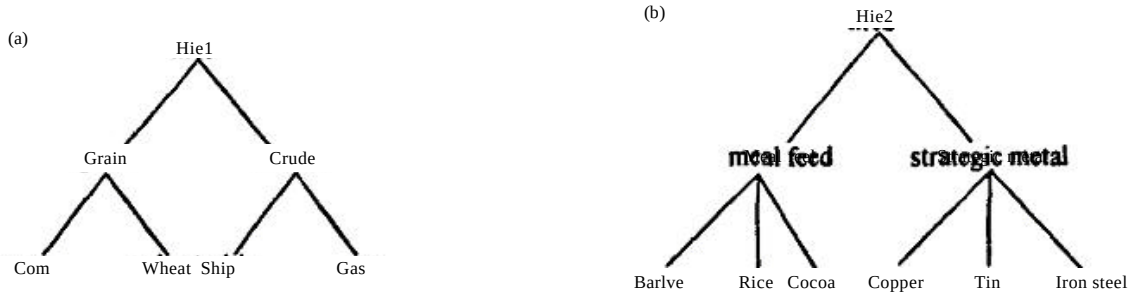

Fig. 4: Reuters-21578 classification tree

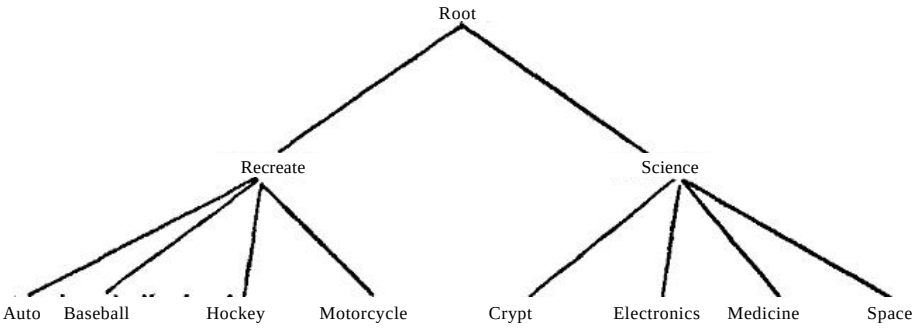

Fig. 5: Merged 20 newsgroups classification tree

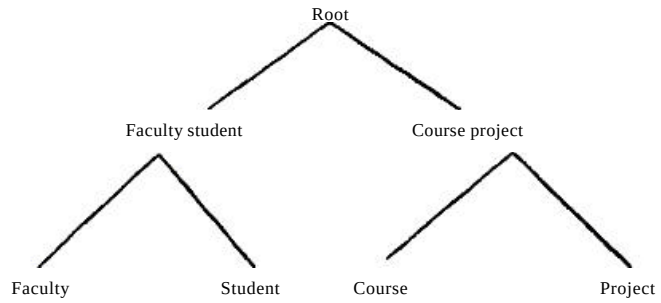

Fig. 6: WebKB classification tree

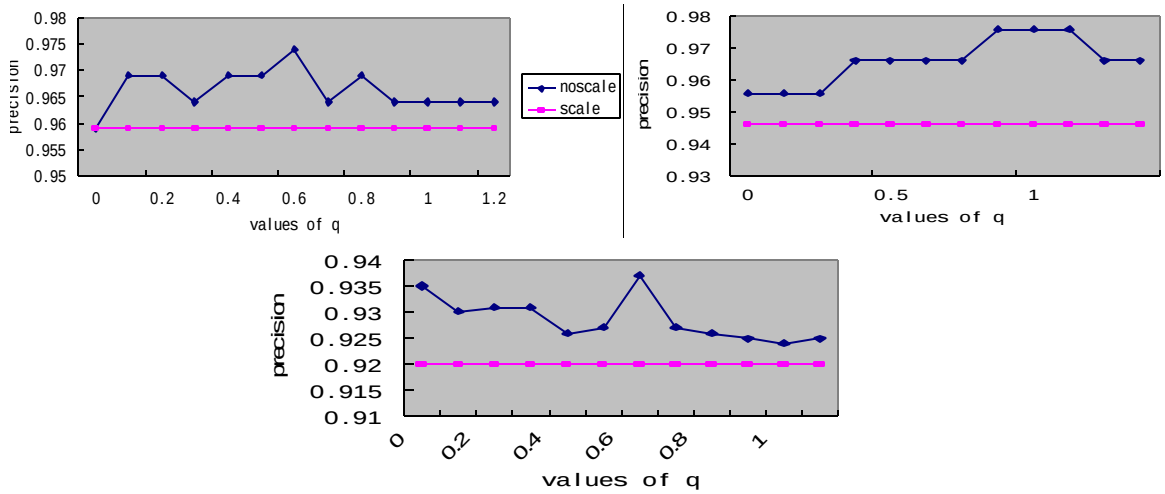

Fig. 7(a-c): Classification accuracy under different amplification parameters, (a) Classification accuracy of Tree hie1 in Reuters-21578, (b) Classification accuracy of Tree hie2 in Reuters-21578 and (c) Classification accuracy of Tree in Newsgroups

every layer of the term-document matrix, so LSI is applied to the new matrix. In this study, We use K-NN method to classify test documents.

EXPERIMENTAL ANALYSIS

Data set: Experiments conducted in this study chose three public databases, there are: Reuters-21578, 20 newsgroups and WebKB. Data set category is divided into hierarchical structure in advance.

- **Reuters-21578:** we have adopted the data set is Sun and Lim (2001) build good category hierarchy tree, selection of the two as the experimental data, we selected in the experiment is a leaf node corresponding categories, two hierarchy tree as shown
- **20 Newsgroups:** 20 newsgroups data set, we select rec and two sci from seven sub trees and combine them an architectural tree as the experimental data
- **WebKB:** We adopt in this data set and (Zelikovitz s. and Hrirsh, 2011) of the same data: student, faculty, course and project. The same as the front two data sets, we built the hierarchy tree:

Experimental parameters setting: We pre-process all data sets, remove stop words, stem each entry and filter out the entries which are composed of two letters or less. In the label extraction formula, there are two important parameters α and β, this study adopts the method of fixed value, set β is 50/T, T as the number of topicsand $\beta = 0.01$. The dimension of the matrix is very important, we fix dimension of 50 day. Through the experiment, the KNN parameter is set to 20. To illustrate the algorithm of this study indeed improves the classification performance, we compare with the traditional SVM classifier, we adopt libSVM tools to implement.

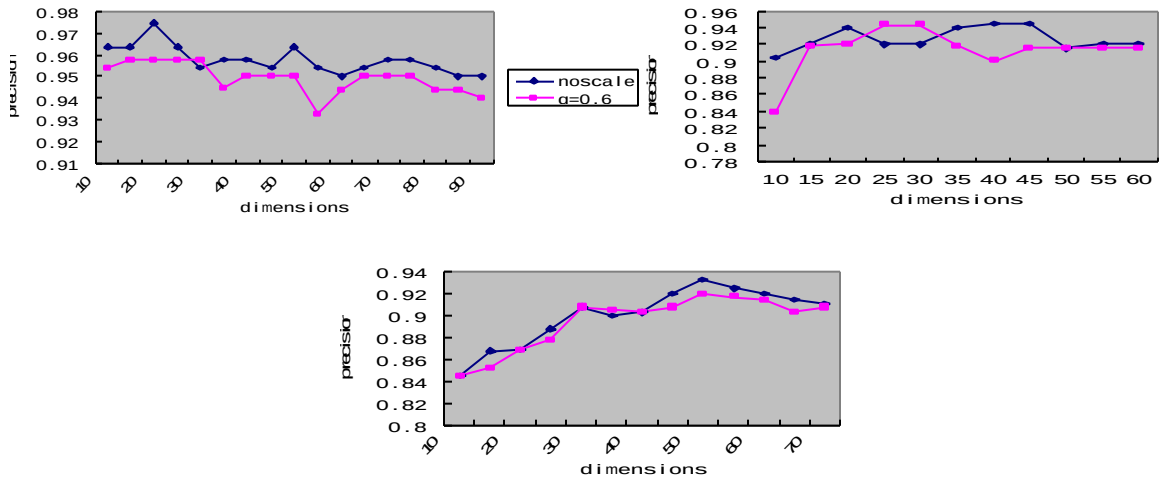

Fig. 8(a-c): Classification accuracy, q = 0.6 Classification accuracy of Tree hie1, (b) q = 0.9 Classification accuracy of Tree hie2 and (c) q = 0.05 Classification accuracy of 20NG

Table 1: Classification results of Retures-21758

	Precision (%)			Recall (%)			F1(%)		
Category	KLSI	SLSI	SVM	KLSI	SLSI	SVM	KLSI	SLSI	SVM
Corn	86.21	90.00	96.88	73.33	79.41	91.18	79.37	84.38	93.94
Wheat	86.96	89.13	93.88	85.11	87.23	97.87	86.02	88.17	95.83
Ship	94.19	97.65	95.29	94.19	96.51	94.19	94.19	97.08	94.74
Gas	80.00	82.86	83.33	96.55	100.00	86.20	87.50	90.63	84.75

Table 2: Classification results of Newsgroup

	Precision (%)			Recall (%)			F1 (%)		
Category	KLSI	SLSI	SVM	KLSI	SLSI	SVM	KLSI	SLSI	SVM
Auto	76.67	78.02	69.07	69.00	71.00	67.00	72.63	74.74	68.02
Baseball	68.81	70.64	74.74	75.00	77.00	71.00	71.77	73.68	72.82
Hockey	77.66	79.12	62.59	73.00	72.00	87.00	75.26	75.39	72.80
Motorcycle	73.39	75.93	76.00	80.00	82.00	76.00	76.56	78.85	76.00
Crypt	79.44	82.69	83.63	85.00	86.00	85.00	82.13	84.31	86.29
Electronics	69.07	66.67	75.73	67.00	68.00	78.00	68.02	67.33	76.85
Medicine	77.55	78.57	78.16	76.00	77.00	68.00	76.77	77.78	72.73
Space	79.17	78.57	83.50	76.00	77.00	70.00	77.55	77.78	77.78

Table 3: Classification results of WebKB

	Precision (%)			Recall (%)			F_1(%)		
Category	KLSI	SLSI	SVM	KLSI	SLSI	SVM	KLSI	SLSI	SVM
Faculty	78.57	82.41	80.21	88.00	89.00	77.00	83.02	85.58	78.57
Student	78.95	86.66	84.00	75.00	78.00	84.00	79.62	82.11	84.00
Course	77.27	78.07	82.18	85.00	89.00	83.00	80.95	83.18	82.59
Project	83.13	81.82	80.58	69.00	72.00	83.00	75.41	76.60	81.77

Result of the experiment: This study selects three trees in Reuters-21578 and 20Newsgroups to study that how will amplification parameters change affect classification?(classification assessment using different amplification parameters based on batch and continuous tests).We use classification accuracy to evaluate results, record as shown in Fig. In the experiment, the traditional KNN method is named noscale, our method is named scale.As can be seen from the data in the chart, the proposed algorithm does improve the precision of the classification but the performance is not stable.The reasons for this result may be: the topic label is not the best, because the extraction is based on Hidden Markov chain that gradually tends to a stable value, thus obtaining semantic label may not unambiguously express the semantic information of category; Next, this study

consider that the documents belong to a certain category, the relationship between actual test documents is more complex, as well as all kinds of categories, so some documents can not be inerrably classified

We have already mentioned in front, dimension of LSI matrix is a extremely important parameter for classification performance but we does not get a perfect method to determine the dimension.

In the following figure, we have designed three different amplifier parameters for comparison. Can be seen from figure, when amplification parameter is 0.6, in the classification tree of hie1, 17 day achieve the best classification results. Similarly, in the trees of 20NG and hie2, when the parameters are 0.9 and 0.05, the relationship between classification accuracy and the dimension can be respectively obtained in the figure below.

In this study, we adopt classification accuracy and recall rate and F1 value to assess classification performance. For convenient to record, our approach is called SLSI, the KNN combining LSI is called KLSI.Through the data in the table below, we' come to the conclusion: the latent semantic indexing method based on topic labels can improve the classification accuracy, it shows that this method is effective.

CONCLUSION

In this article, we first aware of that the label has important influence on text classification, especially when the data set is expressed in the form of hierarchical organization. For text categorization, we prove the label can point to better results. On the basis of previous theoretical research, we explore an approach of combining topic label with LSI. The latent semantic classification Based on topic label Can be summarized as: first we extract the topic with maximum probability as label and then enlarge the label entry (this procedure is simply aimed at increasing the semantic similarity of the same kind of documents, at the same time reducing the correlation of documents of a different category, finally improving the accuracy of document classification). The experiments show that our method can obviously improve performance of text multi-hierarchical Categorization.

ACKNOWLEDGMENTS

This study is supported by the National Nature Science Foundation of China (No.~61170201), (No.~61070133) and Science and Technology Project of Yangzhou City (No. YZ2011098), (No. YZ2012051) and also Supported by Open Project Foundation of Information Technology Research Base of Civil Aviation Administration of China (No.CAAC-ITRB-201307).

REFERENCES

Bade, K., E. Hunenneier and A. Numberger, 2006. Hierarchical classification by expected utility maximization. Proceedings of the Sixth Intemational Conference on Data Mining, December 18-22, 2006, Hong Kong, pp: 43-52.

Cerri, R. and A.C.P.L.F. de Carvalho, 2010. Hierarchical multilabel classification using top-down label combination and artificial neural networks. Proceedings of the 11th Brazilian Symposium on Artificial Neural Networks, October 23-28, 2010, Sao Paulo, pp: 253-258.

Cohen, W.W. and Y. Singer, 1996. Context sensitive learning methods for text categorization. Proceedings of the 19th Annual Internation Conference on Research and Development in Information Retrieval, August 18-22, 1996, Zurich, Switzerland, pp: 307-315.

Koller, D. and M. Sahami, 1997. Hierarchically classifying documents using very few words. Proceedings of the 14th International Conference on Machine Learning, July 8-12, Nashville, Tennessee, pp: 170-178.

Liu, S.H. and M.K. Dong, 2001. A multi-hierarchical text classification method based on vector space model. J. Chinese Inform. Proces., 16: 8-14.

Liu, T., Z. Chen, B. Zhang, W.Y. Ma and G. Wu, 2004. Improving text classification using local latent semantic indexing. Proceedings of the 4th International Conference on Data Mining, November 1-4, 2004, Brighton, UK., pp: 162-169.

Loh, W.Y. and Y.S. Shih, 1997. Split selection methods for classification trees. Statistica Sinica, 7: 815-840.

Mayne, A. and R. Perry, 2009. Hierarchically classifying documents with multiple labels. Proceedings of the Symposium on Computational Intelligence and Data Mining, March 30-April 2, 2009, Nashville, TN., pp: 133-139.

Rastogi, R. and K. Shim, 1998. Public: A decision tree classifier that integrates building and pruning. Proceedings of the 24th International Conference on Very Large Data Bases, August 24-27, 1998, New York, USA.

Silla Jr, C.N. and A.A. Freitas, 2010. A survey of hierarchical classification across different application domains. Data Min. Knowl. Discov., 22: 31-72.

Song, F.X., 2004. Some basic problems of automatic text classification research. Ph.D. Thesis, Nanjing University of Technology.

Sun, A. and E.P. Lim, 2001. Hierarchical text classification and evaluation. Proceedings of the International Conference on Data Mining, November 29-December 2, 2001, San Jose, CA., pp: 521-528.

Vens, C., J. Struyf, L. Schietgat, S. Dzeroski and H. Blockeel, 2008. Decision trees for hierarchical multi-label classification. Mach. Learn., 73: 185-214.

Wang, Y. and J. Gai, 2004. Chinese text classification technology based on latent semantic analysis. Res. Appl. Comput., 8: 151-154.

Xu, F.Y., 2004. Study of multi level Chinese text classification technology. Master's Thesis, Tsinghua University.

Yao, L.Q. and G.W. Wu, 2004. A classification model of scientific papers based on the hierarchical structure. Comput. Eng. Appl., 4: 18-22.

Zelikovitz, S. and H. Hirsh, 2011. Using LSI for text classification in the presence of background text. Proceedings of the 10th International Conference on Information and Knowledge Management, November 5-10-, 2001, ACM Press, New York, pp: 113-118.