The Classification Techniques of Websites for The Case of China-Africa Related Topics

Francois Tchiegue, Rui Li, Shilong Ma
State Key Lab. Of Software Development Environment
School of Computer Science & Engineering, Beihang University
Beijing 100191,China
{FrancoisT, lirui, slma}@nlsde.buaa.edu.cn

Abstract—**Having observed the existing search engines for online information, considering the huge (and unnecessary) amount of search results for each online search, the goal of this research work is to build on an accurate webpage classification technique, by combining feature selection techniques, and pushing the clustering concept to a next step. In this paper, we conducted experiments with various numbers of websites selected by different feature selection algorithms on a well-defined initial set of features and show that by combining some textual classification methods, we do obtain considerable classification accuracy.**

Keywords-search engines, clustering, online search, webpage classification, feature selection

I. Introduction

China and Africa are two great nations that have been sticking together for many decades, helping each other in improving all aspects of life, despite various challenges. One of the biggest challenges would be to overcome the language barrier, and the cultural difference, to create a platform through which average citizens of both nations would easily access information related to China and Africa relationship in various fields (Education, Health, Politics and Cooperation, Infrastructure, Natural resources, Culture etc.).

However, for most of average Africans who live in Africa (where I grew up for over two decades), the Internet access is still a luxury, and the surfing time for each user is very limited; thus there is a real need to get straight to specific search results each time…

Unfortunately, when using the existing search engines to search information on the Internet, the World Wide Web always provides millions of results (web pages) for each and every search term, making it very difficult for users to quickly get to specific results, because they are interested in only tiny part of the results instead of the huge rest of results; therefore the retrieval is very inefficient. Meanwhile, the classification could solve the chaos of the Internet information to a great extent, and make it easier for users to precisely target the required information and divert information.

The website automatic classification has become a key technology of high practical value.

II. The World Wide Web's basic structure

The WWW attracts numerous users with its rich contents.

The Web is a mesh-structured system, which includes lots of information formats such as text, picture, audio and video. Text is a major form of information resources. According to the statistics, 80% information is stored in the form of text in the online storage[1].

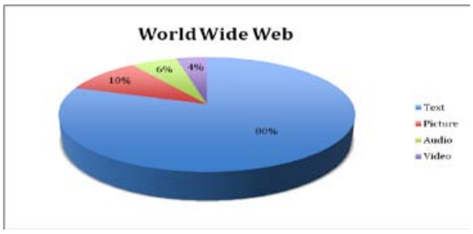

Figure 1. The World Wide Web Information Type Repartition

Information diversity and complexity is a new challenge for the research and development of information classification and retrieval[2].

This research paper is based on the above background, and will take cultural websites as an example to conduct in-depth research on the website classification technology.

III. General Procedures

To make good use of the large-scale mass information, there are two main problems to be solved:

1) Implementation of the search algorithm with high precision and recall ratio;

2) Implementation of classification algorithm with high accuracy.

A. Data Mining

Since the data mining [3][4] processes tremendous data, the data integrity, consistency and accuracy are difficult to guarantee.

Therefore, it is crucial to guarantee the efficiency, validity and expandability of data mining algorithm.

B. *Classification of Web Mining*

Based on the Web Mining Structure Classification proposed by R. Cooley et al, Web mining [5] is divided into three types according to different applications:

1) *Content Mining*

Content Mining is the process of extracting information from text, picture and other contents on the pages. For example: Which pages are related to health, and which pages are related to environment protection.

Search engines, intelligent Agent and some other recommended engines use content mining to help users to find specific targets in the vast network.

2) *Usage Mining*

Usage Mining is the process of extracting information from the ways that website users utilize the website contents. For example: Which pages did they visit, how long did they stay on each page and what are their next visits.

3) *Structure Mining*

Structure Mining is the process of extracting information from the topological structure of the web page links. For example: Which pages are incoming and outgoing pages? Which pages are Authority pages?

TABLE 1 COMPARISON BETWEEN THREE TYPES OF MINING

Mining type	Data source	Data type	Data object	Set
Content Mining	Page content	Text	Index	Page set
Usage Mining	Access method	Click stream	User behavior	Log data
Structure Mining	Topological structure	Hyperlink	Directed graph	Hyperlink set

C. *Web mining process*

The Web mining mainly includes four stages [6]: Data Collection, Data Preparation, Pattern Discovery and Pattern Analysis, as shown in Figure 2.

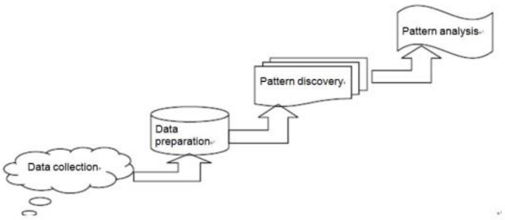

Figure 2. Web mining process

First, collect raw data to be mined from WWW through data collection, but the actual collected data is disordered, redundant and incomplete.

Therefore, prepare the high quality data for the data mining through data preparation.

In the data preparation stage, standardize the data heterogeneity; remove noise data, blank data, irrelevant data and redundant data, use heuristic rules to find the missing data, and use standardization, reduction, switch, rotation and projection operations to reduce the data and improve the knowledge discovery efficiency.

In the pattern discovery stage, select the appropriate mining tools to conduct the actual mining operations.

In the pattern analysis stage, conduct knowledge discovery operations and verify the discovered knowledge.

D. *Web Text Classification Mining*

The WTCM [7] goes through three main steps:

- Web text Preparation: Data Collection, Word Segmentation, Create Textual Feature Database
- Textual Classification
- Classification Performance Evaluation

IV. THE DESIGN AND IMPLEMENTATION OF WEB-BASED CULTURAL WEBSITE CLASSIFICATION SYSTEM

The user's request format is a web page information.

Figure 3 shows the system architecture of Web document classification system we designed. The whole system mainly consists of the following library components and functional modules:

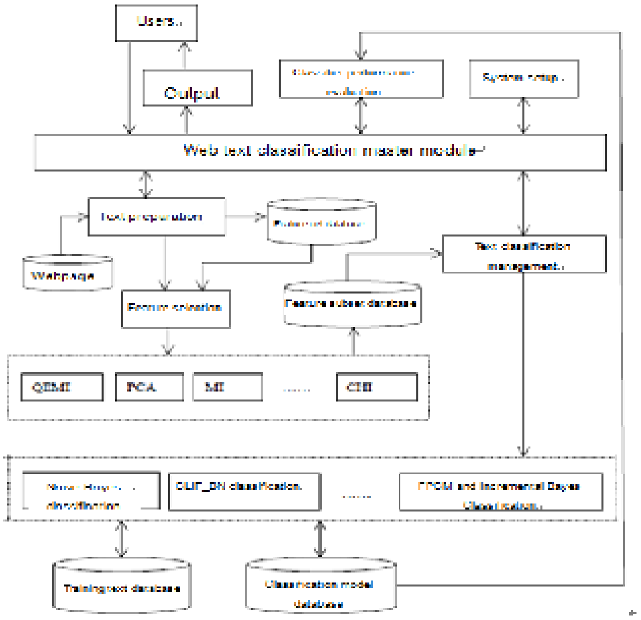
Users
Output
Classification performance evaluation
System setup
Web text classification master module
Text preparation
Feature set database
Webpage
Text classification management
Feature selection
Feature subset database
QEMI
PCA
MI
……
CHI
Naive Bayes classification
CLIT_BN classification
……
FPCM and Incremental Bayes Classification
Training text database
Classification model database

Figure 3. The architecture of the Web Text Automatic Classification prototype system

A. *Text preparation modules*

The web pages downloaded from the Internet are different from plain texts. They contain a lot of format tags, such as <HEAD>&</HEAD>, <TITLE>&</TITLE>, <BODY>& </BODY> etc. These tags represent different sections of the text. We can use the tags to increase weights of important sections, and also consider the importance of paragraphs in different positions. For example, the first and last paragraphs and the first sentence of each paragraph summarize the central topic of the text. Extract these important text paragraphs and sentences to form the reduced texts. Then, segment the words and remove the stop words in a conventional way. Extract the feature words and use the feature vector to represent the text to form the feature set library.

Here's the platform's main page:

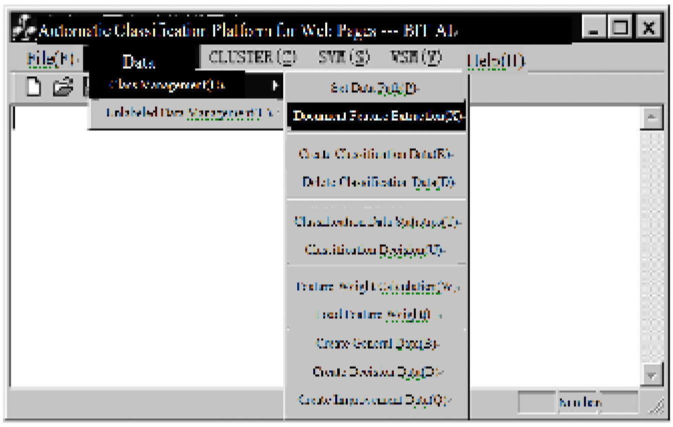
Automatic Classification Platform for Web Pages --- BIT AI
File(F)
Data
CLUSTER (C)
SVM (S)
VSM (V)
Help(H)
Class Management(C)
Unlabeled Data Management(U)
Set Data Path(P)
Document Feature Extraction(X)
Create Classification Data(R)
Delete Classification Data(D)
Classification Data Statistics(A)
Classification Division(V)
Feature Weight Calculation(W)
Load Feature Weight(I)
Create General Data(G)
Create Division Data(D)
Create Improvement Data(Q)

Figure 4. Dialog box that Creates training and test data

In the experiment, we can select the total number of the training samples for each class, and set the proportion between training data and test data. Meanwhile, we can also manually add training and test web pages.

V. WEB TEXT CLASSIFICATION OF UNLABELED TRAINING SAMPLES

This work analyzed the issue in the perspective of clustering, because the major difference between classification and clustering is that classification needs to know the attribute value for the classification in advance, while clustering needs to find out the attribute value for the classification.

As a key technique for text mining, clustering analysis is an unsupervised learning which does not depend on the predefined classes and training samples with class labels, and therefore we can carry out pre-processing of supervised classification, i.e. finding out the classification basis. In addition, since the class itself is a fuzzy concept designed by researchers, and the relationship between various attributes and classes are fuzzy, therefore we can combine the fuzzy clustering theory with automatic classification.

The specific concept is to carry out fuzzy clustering on unlabeled samples with the principle of "Things of one kind come together"

A. *Algorithm*

The traditional fuzzy clustering analysis uses the following algorithm:

Algorithm:

Step1: Generate the text vector matrix D=(d1，d2，…，dn) based on the feature selection

Step2: Conduct singular value decomposition(SVD)onthe text vector matrix D

Step3: Establish the F similarity relation R of Matrix D

Step4: Convert the similarity relation matrix R into F equivalent matrix Q

Step5: For different λ values, acquire the corresponding fuzzy clustering models;Select k singular values from Matrix S0 and set its values to zero, then Matrix D can be approximately expressed by D=TSPT, where T is the n×k matrix with standard orthogonal row, S is the k×k diagonal matrix, and P is also the m×k matrix with standard orthogonal row.

B. *The Web Text Classification Method Combining the Fuzzy Partition Clustering Method (FPCM) and Naive Bayesian Augment Learning*

The simple process of the classification method combining the Fuzzy Partition Clustering Method (FPCM) and Naive Bayesian Augment Learning is shown on the figure 5 below:

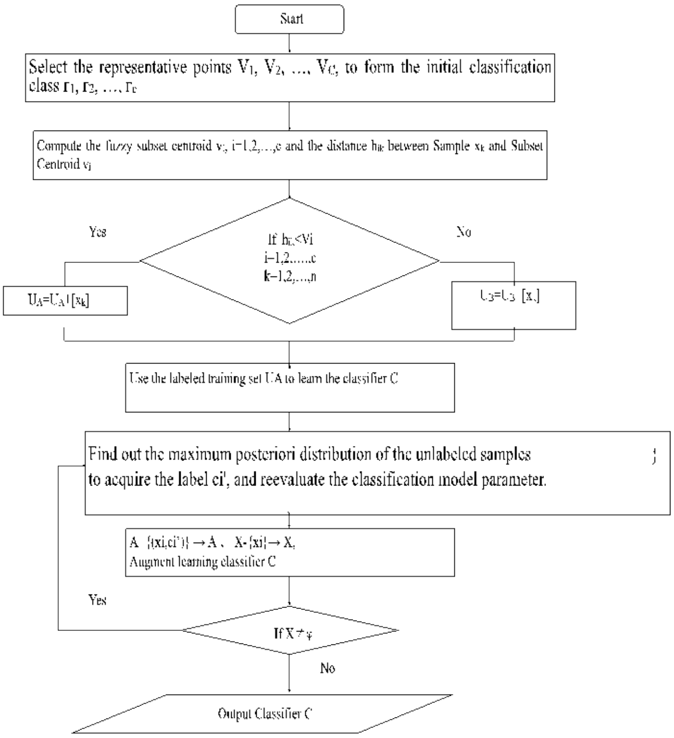
Start
Select the representative points V_1, V_2, …, V_C, to form the initial classification class r_1, r_2, …, r_c
Compute the fuzzy subset centroid v_i, i=1,2,…,c and the distance h_{ik} between Sample x_k and Subset Centroid v_j
Yes
No
If h_{ik}<Vi
i=1,2,…,c
k=1,2,…,n
$U_A=U_A \cup [x_k]$
$U_B=U_B \cup [x_k]$
Use the labeled training set UA to learn the classifier C
Find out the maximum posteriori distribution of the unlabeled samples
to acquire the label ci', and reevaluate the classification model parameter.
A∪{(xi,ci')}→A、X-{xi}→X,
Augment learning classifier C
Yes
If X≠φ
No
Output Classifier C

Figure 5. The classification learning process combining the Fuzzy Partition Clustering Method (FPCM), and Naive Bayesian Augment Learning.

Let X={x_1，x_2，…，x_n} which is the text set, x_k is a sample in X, i=1，2，…n;

Every sample xk has a feature vector p(x_k)=(x_{k1}，x_{k2}，…，x_{ks}), where in x_{kj} (1≤j≤s) is the jth property value, p(x_k) is the feature vector of x_k.

VI. THE EXPERIMENTAL RESULTS

The experimental data are derived from the 4800 Web pages downloaded from the Internet including six classes:

- Education (800 pages),
- Economics and Trade (800 pages),
- Tourism (800 pages),
- Politics and Cooperation (800 pages),
- Natural Resources (800 pages),
- Culture (800 pages).

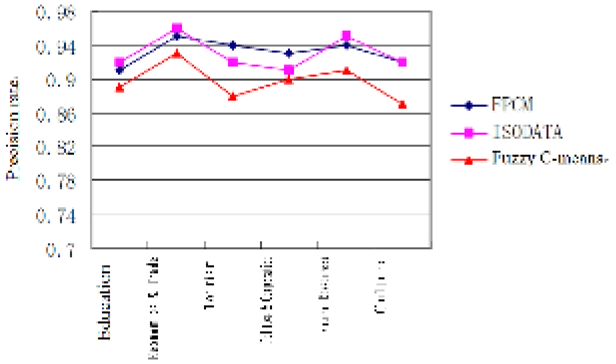

Figure 6. Use FPCM, ISODATA and C-means clustering methods to combine with Naive Bayesian Augment Learning to classify the unlabeled Web text.

From the experiment shown in Figure 6, we conclude that FPCM and ISODATA clustering have similar precision rate when combining with Naive Bayesian Augment, fuzzy C-means clustering has poor results than the above methods when combining with Naive Bayesian Augment. In the experiment, we have also observed that in the whole classification process, the FPCM clustering has the fastest classification speed, fuzzy C-means is slower than FPCM, and ISODATA has the slowest classification speed. In case of low precision requirement in the clustering application in this paper, FPCM is undoubtedly the best choice, because FPCM is more precise than the fuzzy C-means and faster than the ISODATA. Therefore, in the unlabeled text classification, using FPCM clustering to conduct the early-stage labeling work of Naive Bayesian classification can achieve the satisfactory classification result.

Here, we compare the classification result of FPCM and Naive Bayesian Augment Learning with that of fuzzy C-means clustering, ISODATA clustering, as shown in Figure 7.

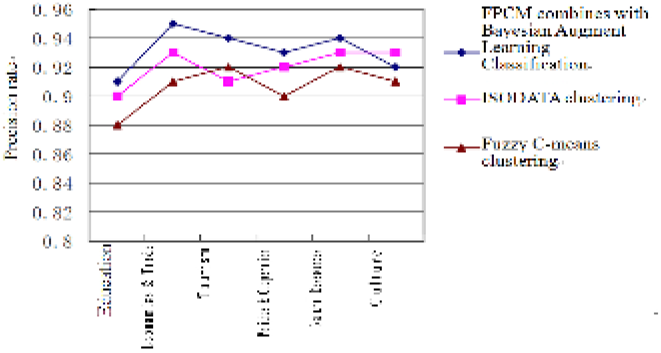

Figure 7: Comparison between the classification results of FPCM and Bayesian Augment Learning and the classification results of ISODATA clustering, fuzzy C-means clustering.

VII. CONCLUSION

The major research findings of this paper include the following aspects:

To solve the high dimensional feature space issue in the Web text classification, using General Theory of Information as the theoretical basis, this paper has proposed QEMI-based feature selection method, removed the information irrelevant to the class in the feature space or feature words with low information content, selected a certain amount of features which are useful for the classification in the high dimensional feature space, reduced the dimensions of the feature space, and provided a solid data basis for designing a highly effective classifier.

This paper has studied conditional independence assumption in the Naive Bayesian Classification Method, proposed to use intermediate variable set combination to replace the CLIF_NB classification learning method with linear inseparability, improved the constraints of the conditional independence by making the intermediate variable space and class relatively independent, and learned a series of classifiers to further improve the performance of CLIF_NB classification model.

This paper has designed and implemented a Web Text Automatic Classification System. The system adopts a modular structural design from the perspective of the actual research and application requirements, which makes it more flexible in the application scope and classification method selection, and facilitates the extension of the system function and the performance improvement.

REFERENCES

[1] Keim, D.A., “Information visualization and visual data mining,” IEEE Transactions, vol. 8, pp. 1–8, Jan 2002.

[2] Manning, Christopher D., Prabhakar Raghavan, and Hinrich Schütze. Introduction to information retrieval. Vol. 1. Cambridge: Cambridge university press, 2008.

[3] Han, Jiawei, Micheline Kamber, and Jian Pei. Data mining: concepts and techniques: concepts and techniques. Elsevier, 2011.

[4] Cabena, Peter, et al. Discovering data mining: from concept to implementation. Prentice-Hall, Inc., 1998.

[5] Cooley, Robert, Bamshad Mobasher, and Jaideep Srivastava. "Web mining: Information and pattern discovery on the world wide web." Tools with Artificial Intelligence, 1997. Proceedings., Ninth IEEE International Conference on. IEEE, 1997.

[6] Kosala, Raymond, and Hendrik Blockeel. "Web mining research: A survey." ACM Sigkdd Explorations Newsletter 2.1 (2000): 1-15.

[7] Berry, Michael W. "Survey of text mining." Computing Reviews 45.9 (2004): 548.