# Research and Implementation of Real-time Automatic Web Page Classification System

Weihong Han [1, a *], Weihui Zhu[1,b] and Yan Jia[1,c]

[1] Computer School, National University of Defense Technology, Changsha, China

[a]hanweihongnudt@139.com, [b] p0zwh@qq.com, [c]jiayanjy@vip.sina.com

**Keywords:** Web Page Classification, security filtering, service discovery, service collection.

**Abstract.** With the development of Internet and communication technology, the Internet data growth rapidly, and the type of network services varied. According to the different properties of the network services, network services classification is the foundation of many network applications, including network service management, green Internet, network bandwidth usage category management, network reputation management, security filtering and so on. Due to the variety of web content and text length, the traditional classification methods can't effectively solve the problem of large-scale web page classification. In this paper, we design and implement a real-time automatic Web page classification system AWCS, including self-feedback system architecture, multi-dimensional network services classification standard, active and passive combining network service discovery and collection technology, automatic self-correction network service classification techniques. Performance tests show that the classification accuracy of AWCS is significantly higher than the traditional algorithms. This framework offers a promising approach for large-scale real-time network data classification system.

## Introduction

Modern network technology develops at an unprecedented rapid rate, and affects all aspects of the country's economic and social life. Diverse varied types of network services, classification of network services according to different attributes is the basis for many network applications, including network service management, green Internet, network bandwidth usage classification management, network reputation management, security filtering . In addition, as types of services on the network get bigger, and quality of service varies greatly, network service classification, in particular the use of network services recommended classification of network services and navigation has become another important application on the network. Therefore, the network service classification research has gradually become a new hotspot.

## Related Research

According to different classification algorithms, research on the network service classification can be divided into four categories: URL feature-based automatic services classification, which extracts corresponding feature in the URL of the service and uses it in services automatic classification [1]; Content-based service classification, which primarily based on the content of web page and uses it to classify the network services[2]; Web site structure-based service classification, which is mainly based on the structure of the Web site to provide characteristics of network services classification[3,4]; multiple technologies integrated service classification, which integrated uses of content-based, structure-based and other service classification methods [5,6,7].

About the products of network service classification, some well-known manufacturers have done a lot of work, more representative include: Web ThreatPak, which is developed by American eSoft company [8], has achieved some success in the service automatic classification. Web ThreatPak can automatic analysis the text and image that service request URL is pointing to, and based on this, it can automatic classify the URL using statistical methods. American companies Websense [9]has created a complete URL classification database, which contains over 3600 million websites, fit into more than 90

 

URL categories, covering 50 languages. Websense combines automatic classification software and human inspection techniques to categorize and maintain the URL.

## Self-feedback system architecture

We designed and implemented a real-time automatic classification system for massive network service AWCS (Automatic Web Page Classification System). System architecture is shown in Figure 1, which is composed by network service discovery subsystem, network services information collecting subsystem, network service automatically classification subsystem, network service classification verification and feedback subsystems and network service classification interface.

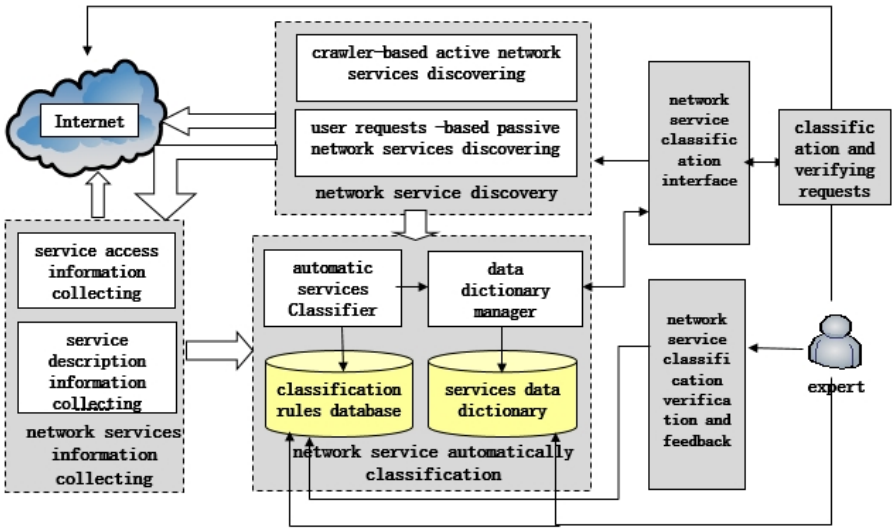


Fig.1 Architecture of real-time automatic classification system

The network service discovery subsystem is used to discover network services, including crawler-based active network services discovering modes and user requests -based passive network services discovering modes. The network services information collecting subsystem is used to collect network service information that is needed by service classification, including service access information collecting and service description information collecting. The network service automatically classification subsystem is used to automatic classify the network service, including automatic network services Classifier, network service classification rules database, network services data dictionary and data dictionary manager. The network service classification verification and feedback subsystems enables network service classification verify of people in the loop, receiving network service classification expert's verifying service classification amendments, and network service classification rule's feedback correction. The network service classification interface is used to receive real-time classification and verifying requests and return the classification results to the user.

## Key Technology

### Multi-dimensional and multi-level network services classification criteria

Web services classification criteria are multi-dimensional service classification standards based on multi-level characteristics on various network services, according to different attributes and different aspects of the network services. Classification criteria can be changed according to the changes of user requirements or the changes of network situation. Web services classification criteria as shown below.

Basic design principles of building a network service classification criteria are practicality, comprehensive and dynamic. This is to say, the network service classification criteria should be comprehensive coverage to meet the various needs of different users, and can be updated and expanded according to the change of network information. Multi-dimensional classification refers to the establishment of service standards from different dimensions of different user requirements, such

as: content dimension classification criteria for the users who care services content; Language dimension of service classification criteria for the users who care about language; Regional dimension of service classification criteria for users who care about the services' area. Multi-level classification refers to a hierarchical multi-level classification criteria according to the services' characteristics.

**service discovering and collecting technology Combined of active and passive mode**

The network services discovering subsystem is used to discovery new network service. The network services information collecting subsystem is used to collect network service information that is needed by service classification.

Network service discovering subsystem combines both active discovery mode and passive discovery mode. The active discovery mode finds new web services by the seed database spreading over, then collects new web services information. The passive discovery mode is based on the user requests. If the service requested by user is not found in the network service data dictionary, then collects new web services information. The service classification subsystem classify the service and storage it in network service data dictionary.

Network service information collecting subsystem is mainly based on web crawler technology. Web crawler is a program that extract web pages automatically, and it is generally used to download web page for search engines. Web crawler is an important component of search engines. According to different ways of crawling, Web crawler can be divided into general crawler and targeted crawler. General crawler begin from one or several initial URL page, and get URL from the initial page. Then it continue to extract new URL from the current page into the queue until the system meets the stop condition. Different with general crawler, targeted crawler will filter the unrelated links based on the analysis algorithms, and retain useful links and put them in the URL queue waiting to be crawled. Then it will select the next web page URL to crawl from the queue by some search strategy, and repeat the process until it reaches a certain stopping condition. In addition, all the crawled pages will be stored in the system, then they will be analyzed, filtered, and automatic classification, so that they can be queried and retrieval.

**Self-correcting network service automatic classification technology**

Web services automatic classification subsystem is used to automatic classify web services on real-time. Subsystem architecture is shown in Figure 2, including offline classification rules learning module and online service classification module.

Offline service classification learning module mainly uses the web service record information and existing network services classification information to learn the network service classification rules, and the generated network service classification rules will be added to network services classification rules base under the direction of experts. Online classification module will classify the network services on real-time when the crawler finds the new network services or the system receives service classification request from the users. First, it will check whether the network service classification information already exists in the data dictionary. If exists, return the classification result. If the data dictionary hasn't the classification information of the network services, the system will classify the network services automatically according to the classification rules. The classification results will be returned to the user while storied in the data dictionary.

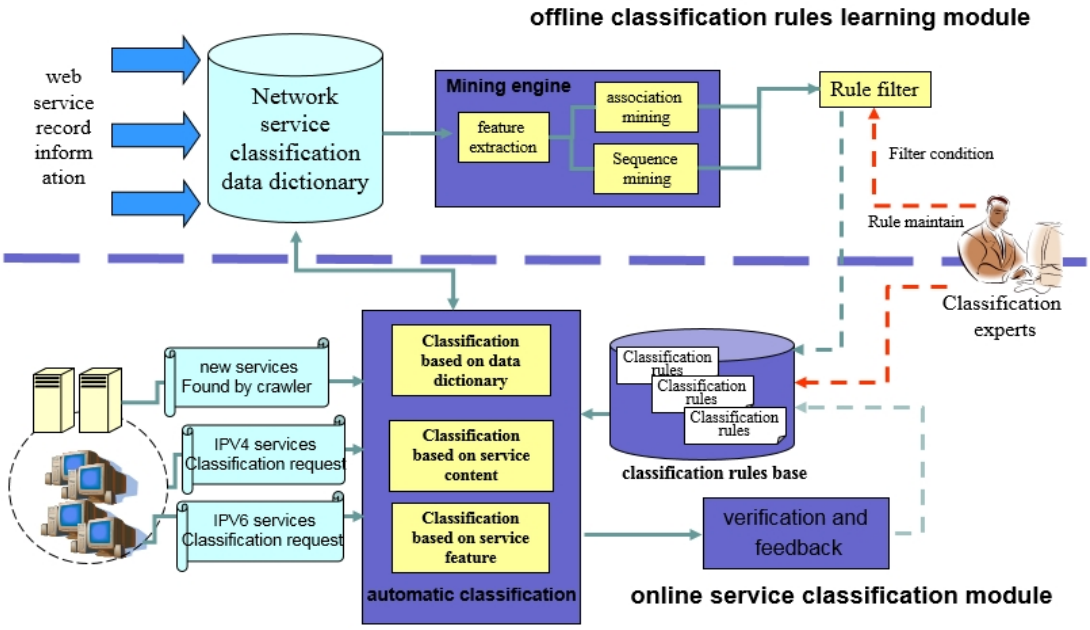


Fig. 2 Architecture of automatic classification subsystem

We tested four kinds of classification algorism for large scale web page, including Flat, top-down, two-stage and class tree Reconstruction, and five kinds of classifier models, including decision trees, naive Bayes, maximum entropy, centroid classifier and SVM. By experimental testing, we adopt the centroid classifier to calculate Candidate category, and then we use maximum entropy to classify the web service. So we realize an automatic classification method based on network services content.

## Performance Analysis

### Experiment data

A lot of research has been made on text corpus which can be used as a training and testing standard. There have been a large number of excellent corpus in international, such as WebKB [10], NewsGroup [11], Reuters, etc. At present, domestic research in this area is still lacked. Sogou is a relatively well-known text corpus, which contains a large number of artificial edited and classified news corpus form Sohu websit, and its page size is about one hundred thousand documents. Another Chinese web page classification training set CCT2006 is collected by Network and Distributed Laboratory Skynet group in Peking University, which includes 960 training pages of and 240 test pages, distributed in eight categories. But there are several drawbacks of these corpus: (1) Sogou corpus contains only news category, it has only one category, and cannot be used for multi-class tests. (2) The corpus of Peking University is too small, it cannot be representative of the general situation in the Internet, so that the error caused by the training and classification will be relatively large.

In the experiments in this paper, we manually collect 35,637 pages for the 42 Category in figure 2, average of 900 pages for each category as the training set. For the test set, we used 9007 pages from hao123 which has been artificial marked.

### Feature selection algorithm experiments

Different feature selection algorithm are good at different application, the effect of different classification algorithm and feature selection algorithm combines is different too. In this paper, in order to find the most suitable classification model and feature selection algorithm, we compares several commonly used feature selection algorithms combined with our classification model, particularly CHI + MI feature selection algorithm has been compared too. In the experiment, the feature item scale is controlled in 9500, the experimental results is shown in Figure 3.

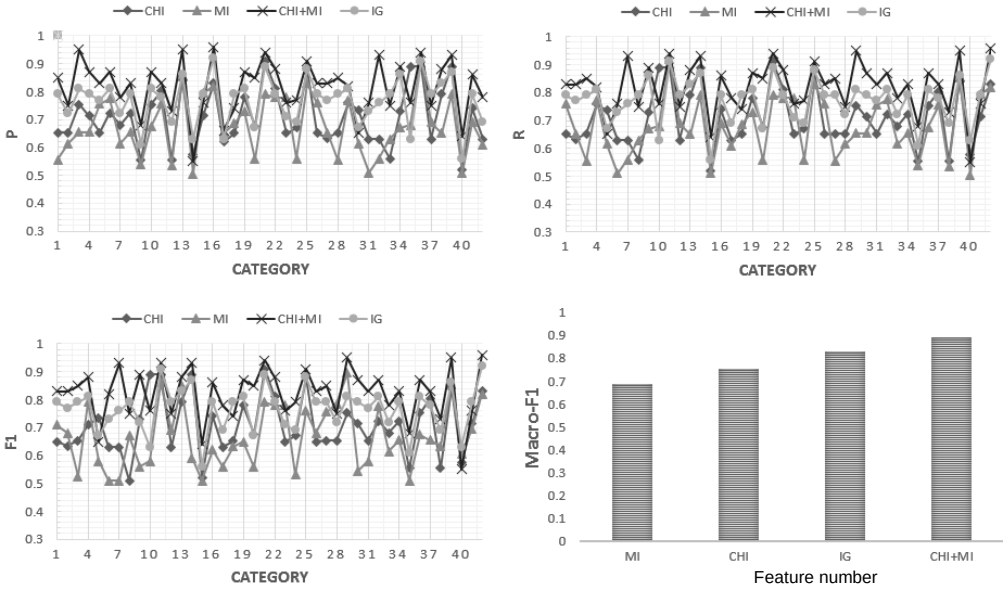


Fig. 3 Feature selection algorithm results contrast

We can get the following conclusions from the data Analysis in Figure 6:

1） In the classification model of this paper, CHI + MI feature selection algorithm achieved the best classification results.

2） From the four feature selection algorithms comparative experiments, we can find that MI is the worst, IG is the better, CHI is between them.

**Classification algorithm comparative experiments**

In order to verify the effect of the classification model proposed in this paper, based on the above two experiments, we make a comparative experiment between our classification system and the common classification algorithms. We use CHI + MI feature selection algorithm, and the feature item scale is 9500. The comparing experiment results are shown in Figure 4 (where x represents the classification system of this paper).

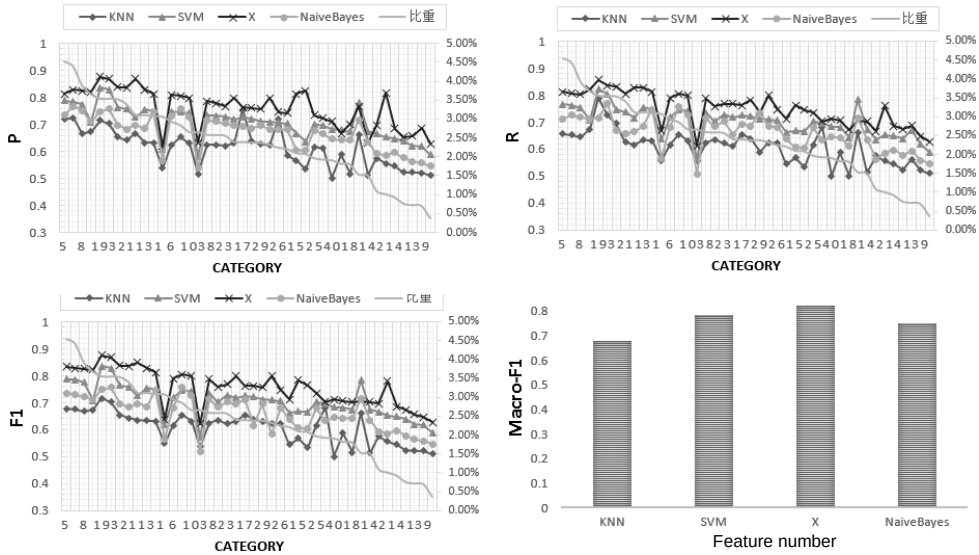


Fig. 4 Classification algorithms compare results

We can get the following conclusions from analysis of these results：

1） The classification system designed in this paper is significantly better than the rest three algorithms when it is used for news, blogs classification. The reason is because that news and blogs web contents are variety, and the feature items are obvious. So commonly used classification algorithms cannot classify them effectively. But classification algorithm based on URL structure pre classification can accurately extract keywords from news and blogs' domain name, and accurately determine its category, making the accuracy dramatically.

2） For photography, video and other categories whose web content is short text, Labled_LDA short text feature space expansion method proposed in this paper, also can make a significant upgrade for the classification results.

3） Regardless of the changes in the proportion of the training corpus, classification results of classification system designed in this paper are better than the other three classification algorithm. So the integration of different classifiers results can often achieve better classification results than a single classifier.

**References**

[1] G Kou, C Lou. Multiple factor hierarchical clustering algorithm for large scale web page and search engine clickstream data. Annals of Operations Research, 2012 - Springer.

[2] WY Dai, Yong Yu, CL Zhang, J Han, and GR Xue. A Novel Web Page Categorization Algorithm Based on. WAIM 2006, LNCS 4016, pp. 435 – 446.

[3] W Lai, R Cai, J Yang, WY Ma. L Zhang. Forum web page clustering based on repetitive regions. US Patent 8,051,083, 2011 - Google Patents.

[4] D Godoy, A Amandi. Exploiting the social capital of folksonomies for web page classification. Software Services for E-World, 2010 - Springer.

[5] S.Gowri Shanthi. WEB PAGE CATEGORIZATION USING WEB MINING. International Journal of Advanced Research in Computer Engineering & Technology (IJARCET) Volume 1, Issue 7, September 2012.

[6] Tong Zhang, Alexandrin Popescul, Byron Dom. Linear Prediction Models with Graph Regularization for Webpage Categorization. KDD'06, August 20–23, 2006, Philadelphia, Pennsylvania, USA.

[7] C. Lindemann and L. Littig, Coarse-grained Classification of Web Sites by Their Structural Properties, Proc. 8th Int. Workshop on Web Information and Data Management, Arlington, VA, 2006

[8] http://www.esoft.com/.

[9] http://www.websense.com/content/Regional/SCH/URLCategories.aspx

[10] Philippe Martin. The WebKB set of tools: A common scheme for shared WWW annotations, shared knowledge bases and information retrieval[J].1997,1257:585-588.

[11] http://kdd.ics.uci.edu/databases/20newsgroups/20newsgroups.html