Large Scale Hierarchical Classification Framework for Network Big Data

Weihong Han [1, a *], Zizhong Huang[1,b] and Yan Jia[1,c]

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

[a]hanweihongnudt@139.com, [b] 13807319539@139.com, [c]jiayanjy@vip.sina.com

Keywords: big data, hierarchical classification, data collection

Abstract. With the development of Internet technology, Internet data growth rapidly and become big data. According to the different properties of the network big data, network big data classification is the foundation of many network applications, including network data management, green Internet, network bandwidth usage category management, network reputation management, security filtering and so on. Due to the variety and the large scale of network data, the traditional classification methods can't effectively solve the problem of network big data classification. In this paper, we design and implement a large scale hierarchical classification framework (LSHC) for network big data, including self-feedback system architecture, multi-dimensional network big data classification standard, active and passive combining network big data collection technology, automatic self-correction network big data classification techniques. This framework offers a promising approach for large-scale real-time network big 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 big data, classification of network big data according to different attributes is the basis for many network applications, including network big data management, green Internet, network bandwidth usage classification management, network reputation management, security filtering. Therefore, the network big data classification research has gradually become a new hotspot.

Related Research

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

About the products of network big data 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 network data automatic classification. Web ThreatPak can automatic analysis the text and image in network data, and based on this, it can automatic classify the network data using statistical methods. 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 network big data.

Large Scale Hierarchical Classification Framework LSHC

we design and implement a large scale hierarchical classification framework (LSHC) for network big data, including self-feedback system architecture, multi-dimensional network big data classification standard, active and passive combining network big data collection technology,

automatic self-correction network big data classification techniques.

large scale hierarchical classification system architecture

The architecture of LSHC is shown in Figure 1, which is composed of 5 subsystems: network big data discovery subsystem, network big data collecting subsystem, network big data automatically classification subsystem, network big data classification verification and feedback subsystems and network big data classification interface.

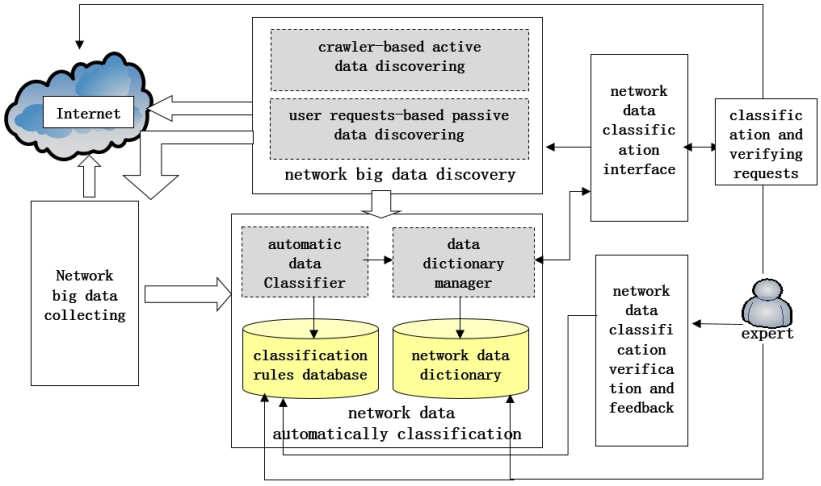

Fig.1 Architecture of real-time automatic classification system

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

Multi-dimensional and multi-level network big data classification criteria

Network big data classification criteria are multi-dimensional classification standards according to different attributes and different aspects of the network big data. Classification criteria can be changed according to the changes of user's requirements or the changes of network situation.

Basic design principles of building a network big data classification criteria are practicality, comprehensive and dynamic. This is to say, the network big data 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 data standards from different dimensions of different user requirements, such as: content dimension classification criteria for the users who care data content; Language dimension classification criteria for the users who care about the data's language; Regional dimension criteria for users who care about the data's area.

Multi-level classification refers to a hierarchical multi-level classification criteria according to the data's characteristics. In fact, most of network big data is hierarchical and the classification criteria can be constructed as a tree.

data discovering and collecting technology Combined of active and passive mode

The network big data discovering subsystem is used to discovery new network big data. The network big data information collecting subsystem is used to collect network big data information which is needed by classification.

Network big data discovering subsystem combines both active discovery mode and passive discovery mode. The active discovery mode finds new data by the seed database spreading over, then collects new network data's information. The passive discovery mode is based on the user requests. If the user's data request is not found in the network big data dictionary, then the system will collect new data's information. The data classification subsystem classify the data and storage it in network big data dictionary.

Network big data 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.

Network big data discovery and collecting subsystem's architecture is shown in Figure 2.

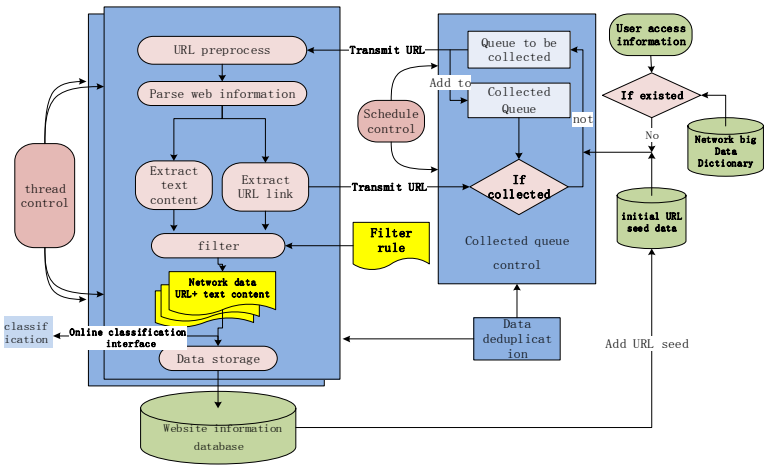

Fig. 2 Network big data discovery and collecting subsystem's architecture

In the figure, we can see that network discovery subsystem and network big data collecting subsystem is integrated together, and when the system find the network big data which meet the requirements, it will collect the data and deliver them to the classification subsystem. First, the system will transfer the data which is in the seed bank or got from user access requirement that does not exist in the dictionary to the subsystem, the subsystem will add them to crawling queue by scheduling control algorithm, and set crawling depth, then add them to the already collected queue. Then start the thread by thread control module, crawling tasks, preprocessing the URL. The URL seed preprocessing module will preprocess URL first, and decide whether the URL is crawling. After parsing the URL page, the extracted URL links will be transfer to the acquisition queue control module, to determine whether the link has been crawled although. If the link has been collected, the collection will not be performed. If it has not been collected, it will be added to the collection queue. The crawler spreading uses finite depth strategy, when the depth of crawler achieve setting depth of crawling system, it will stop the spread of the seeds. The system can find new network big data through the spread of crawler. After parsing the web pages we can also extract the text content. Text content contain description information and content information, such as <title>, <description>, <keywords>, <body>, and other useful information. We filter text content by filtering rules, and then storage the web site URL and text information which meet the requirements to the site database.

Self-correcting network big data automatic classification technology

Network data automatic classification subsystem is used to automatic classify network data on real-time. Subsystem architecture is shown in Figure 3, including offline classification rules learning module and online data classification module.

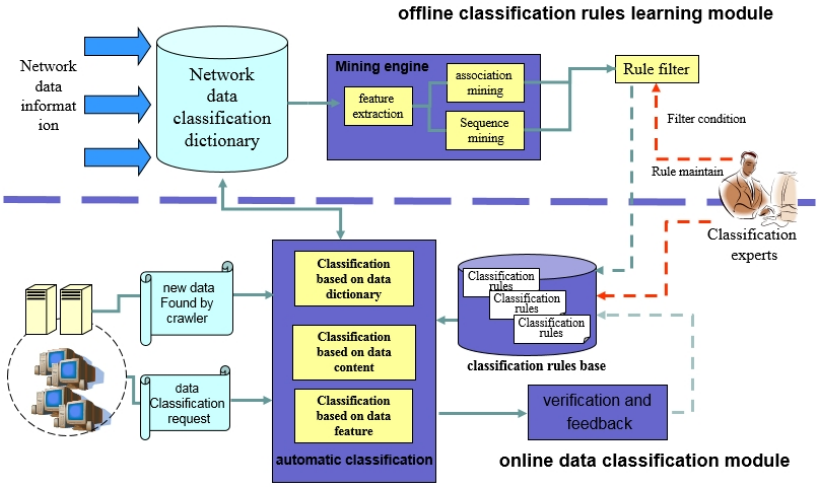

Fig. 3 Architecture of automatic classification subsystem

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

Performance Analysis

experiment data and evaluation standard

We use the ODP Simplified Chinese directory as the test data set. Simplified Chinese ODP directory is a hierarchy tree category whose depth is 6, including 1763 category, 24,570 sites. Therefore, ODP data set is very suitable for verification of the classification framework of this paper. We use ODP categories as a training set of data. For the test set, we used the network data collected by Hao123, which has been artificial annotated.

For the evaluation index of Classification results, we mainly use recall(R), Precision(P), F1 and Micro Averaging (MacroF1). F1 and MacroF1 are calculated as follows:

$$F1=(2*R*P)/(R+P)\times 100\%$$

$$MacroF_1 = \frac{2 * MacroP * MarcoR}{MacroR + MacroP} \times 100\%$$

Classification algorithm comparative experiments

In order to verify the effect of the classification framework proposed in this paper, 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).

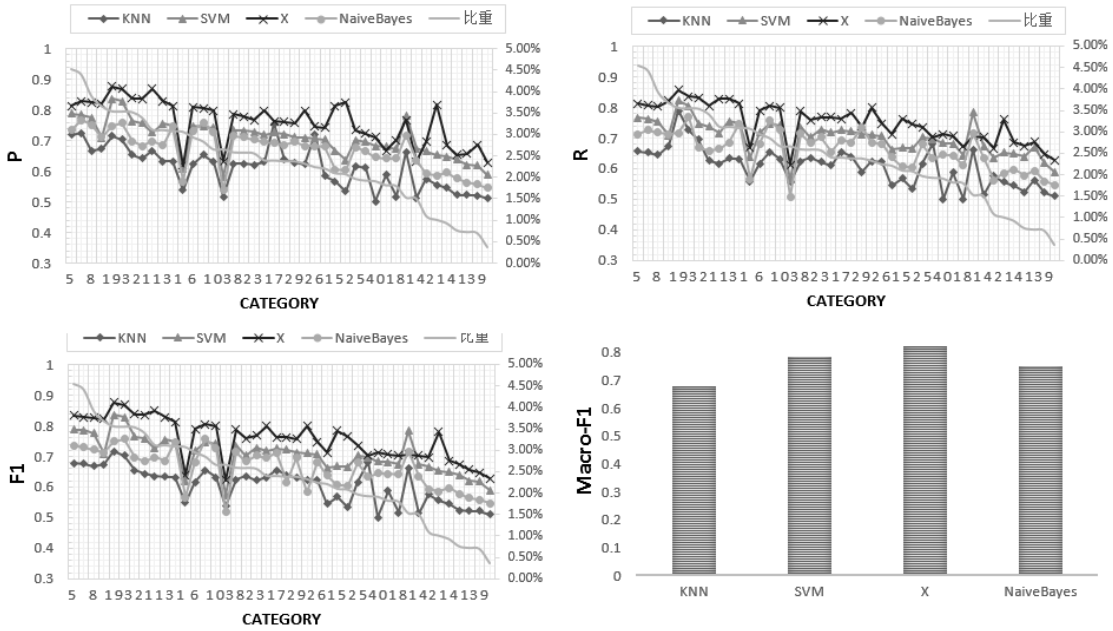

Fig. 4 Classification algorithms compare results

From analysis of these results We can see that the classification framework designed in this paper is significantly better than the rest four algorithms when it is used for Webpage classification.

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