Data Preprocessing Based on Partially Supervised Learning

Na Liu[1,2, a], Guanglai Gao[1,b], Guiping Liu[2,c]

[1]College of Computer Science Inner Mongolia University Hohhot, China

[2]Department of Science Hetao University Bayannur, China

[a]csna_liu@163.com ,[b]csggl@imu.edu.cn, [c]csguiping_liu@163.com

Keywords: data preprocessing; Web Log Mining; rule; partially supervised learning

Abstract. Data preprocessing is the foundation to improve the quality of data mining and determines the effect of Web mining. Currently, data for mining is typically collected from the server, but data set from the client is more accurate. In order to better deal with these data, we propose a data preprocessing method based on partially supervised learning. The paper discusses in detail data cleaning process based on partially supervised learning, and conducts experiments to verify the validity of the method employed, and ultimately determines the optimal number of training documents.

Introduction

With the advent of the information age, Internet has become an important way for people to obtain information. Companies perform e-commerce activities on the Internet, deploy and develop Internet marketing, and this has become an important part of marketing business development. Therefore, web mining, which aims to find the users access law, has become a hot topic for all enterprises and organizations under the network environment.

The goal of Web mining is to explore the useful value from the hyperlinks structure, content and usage logs. According to the use of data categories in the mining process, Web mining is divided into three types: Web structure mining, Web content mining and Web log mining [1]. The main purpose of Web log mining is to find Web page user access mode. Mining is generally carried out in three steps: preprocessing, data mining and subsequent processing. Due to immature extraction technology, there are some incomplete data, isolated points and noise data in Web logs and these logs can not be directly used for mining[2][3]. So, pretreatment determines the effect of Web mining and is the basis of this process. Improving the quality of preprocessing of data mining to meet the specific requirements of the data can improve the efficiency of decision-making.

This paper proposes a data preprocessing method based on partially supervised learning and discusses in detail the data cleaning process.

Related Works

In 1999 Pyle [4] for the first time used data preprocessing in Web log mining. In the same year, Cooley [5] pointed out that fixing data errors and handling missing data are key tasks for data preprocessing. Because users access to multiple Webs or use application servers, Tanasa [6] in 2005 presented in his article that the log files from multiple Webs and application servers be merged. [7] It also proposed a preprocessing algorithm based on collaborative filtering.

Currently, because most data for Web log mining is collected directly from the Web server's log files, there is a lot of junk data and incomplete browsing records, inaccurate browsing time and other problems. Client data based on user behavior sources can provide comprehensive and accurate information on user browsing behavior. In earlier years, Sitellelper’s [8] team used the client to extract user information. Due to violation of user privacy, the system failed to fully go into market operations. P3P (Platform for Privacy Preferences) technology provides a privacy protection strategy, and allows client data acquisition to become a reality. But in the Web log mining field, client data mining research is still relatively rare; there is much room for improvement in data preprocessing techniques.

Data preprocessing includes data cleaning, transformation, integration, reduction and so on. This paper intends to use partially supervised learning methods to clean up web browsing log files, and to validate the method through experiments.

Data Cleaning Based on Partially Supervised Learning

For various reasons, there are always dirty data in the collected data set e.g. incomplete data, illegal value, the null value, inconsistent values, isolated points etc. To solve the above problems in the data set is to clean up dirty data.

A. Principle of data cleanup

The core idea of data cleaning is to find out characteristics of data and to extract, according to these features, to design and implement effective algorithm, rules and strategies, and ultimately complete data cleaning.

The main task of data cleaning includes:

- Predict missing values and complement incomplete data.
- Identify and remove illegal and null values.
- Convert or delete inconsistent data.
- Apply suitable algorithms, rules and strategies to amend or delete the abnormal data.

Based on the above analysis of data cleansing principles and tasks, the basic process of data cleaning includes analyzing characteristics of data, defining data cleaning rules, performing data cleansing, validating data and reflowing clean data. The basic process of data cleaning is as follows:

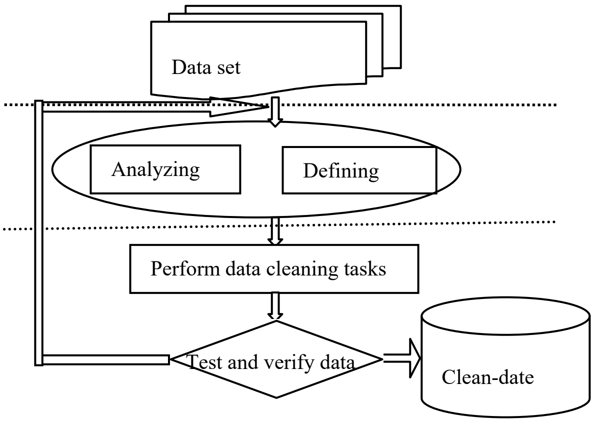


Fig.1 Process of data cleaning

Data mining analysis is to analyze the collected data and extract the laws. The discovery process is to find out abnormal data and define the preliminary cleanup rules and procedures.

Defining data cleanup rules means classifying of data on the basis of further data analysis, and defining of a detailed cleanup rules for different categories.

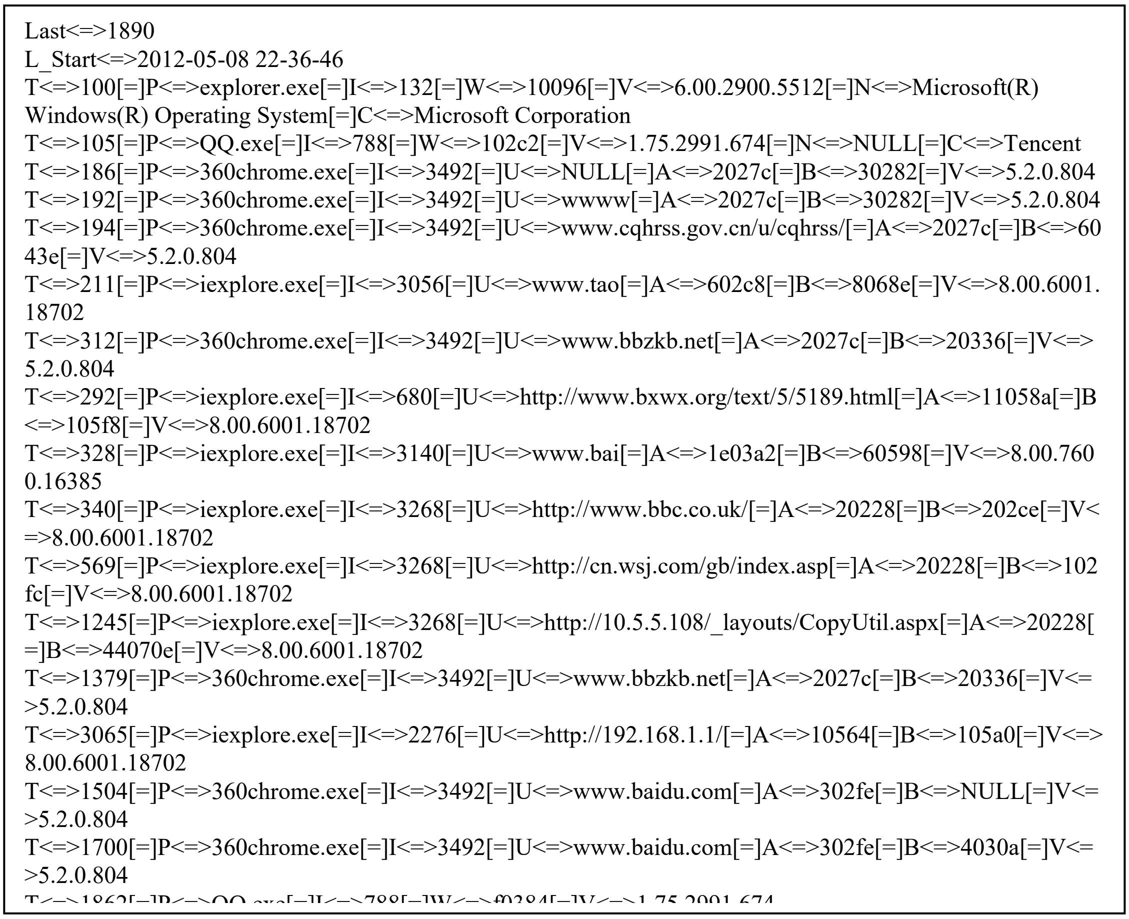
Performing data cleaning refers to defining good data cleanup rules and applying appropriate algorithms and strategies to implement and execute on the data source.

Data validation refers to verifying the correctness of the data cleanup rules and evaluation of efficiency through analysis of multiple iteration, definition, implementation and verification until satisfactory data cleaning rules are found.

Reflow means that when the data is cleansed, clean data should replace the source data.

B.Marking Positive Examples for Rule-Based Learning

Dirty data (negative example) is inherently uncertain, and causes the difficulties in the identification and definition of cleanup rules. And in our case, most complete data (positive example) from source data have significant features, and conform to business rules. In this study, the original data samples are as shown in Fig.2.



```
Last<=>1890
L_Start<=>2012-05-08 22-36-46
T<=>100[=]P<=>explorer.exe[=]I<=>132[=]W<=>10096[=]V<=>6.00.2900.5512[=]N<=>Microsoft(R)
Windows(R) Operating System[=]C<=>Microsoft Corporation
T<=>105[=]P<=>QQ.exe[=]I<=>788[=]W<=>102c2[=]V<=>1.75.2991.674[=]N<=>NULL[=]C<=>Tencent
T<=>186[=]P<=>360chrome.exe[=]I<=>3492[=]U<=>NULL[=]A<=>2027c[=]B<=>30282[=]V<=>5.2.0.804
T<=>192[=]P<=>360chrome.exe[=]I<=>3492[=]U<=>wwww[=]A<=>2027c[=]B<=>30282[=]V<=>5.2.0.804
T<=>194[=]P<=>360chrome.exe[=]I<=>3492[=]U<=>www.cqhrss.gov.cn/u/cqhrss/[=]A<=>2027c[=]B<=>60
43e[=]V<=>5.2.0.804
T<=>211[=]P<=>iexplore.exe[=]I<=>3056[=]U<=>www.tao[=]A<=>602c8[=]B<=>8068e[=]V<=>8.00.6001.
18702
T<=>312[=]P<=>360chrome.exe[=]I<=>3492[=]U<=>www.bbzkb.net[=]A<=>2027c[=]B<=>20336[=]V<=>
5.2.0.804
T<=>292[=]P<=>iexplore.exe[=]I<=>680[=]U<=>http://www.bxwx.org/text/5/5189.html[=]A<=>11058a[=]B
<=>105f8[=]V<=>8.00.6001.18702
T<=>328[=]P<=>iexplore.exe[=]I<=>3140[=]U<=>www.bai[=]A<=>1e03a2[=]B<=>60598[=]V<=>8.00.760
0.16385
T<=>340[=]P<=>iexplore.exe[=]I<=>3268[=]U<=>http://www.bbc.co.uk/[=]A<=>20228[=]B<=>202ce[=]V<
=>8.00.6001.18702
T<=>569[=]P<=>iexplore.exe[=]I<=>3268[=]U<=>http://cn.wsj.com/gb/index.asp[=]A<=>20228[=]B<=>102
fc[=]V<=>8.00.6001.18702
T<=>1245[=]P<=>iexplore.exe[=]I<=>3268[=]U<=>http://10.5.5.108/_layouts/CopyUtil.aspx[=]A<=>20228[
=]B<=>44070e[=]V<=>8.00.6001.18702
T<=>1379[=]P<=>360chrome.exe[=]I<=>3492[=]U<=>www.bbzkb.net[=]A<=>2027c[=]B<=>20336[=]V<=
>5.2.0.804
T<=>3065[=]P<=>iexplore.exe[=]I<=>2276[=]U<=>http://192.168.1.1/[=]A<=>10564[=]B<=>105a0[=]V<=>
8.00.6001.18702
T<=>1504[=]P<=>360chrome.exe[=]I<=>3492[=]U<=>www.baidu.com[=]A<=>302fe[=]B<=>NULL[=]V<=
>5.2.0.804
T<=>1700[=]P<=>360chrome.exe[=]I<=>3492[=]U<=>www.baidu.com[=]A<=>302fe[=]B<=>4030a[=]V<=
>5.2.0.804
T<=>1862[=]P<=>QQ.exe[=]I<=>788[=]W<=>f0384[=]V<=>1.75.2991.674
```

Fig.2 Raw data

(1)PU leaning

Partially supervised learning is divided into LU learning (learning from Labeled and Unlabeled examples) and PU learning (Learning from Positive and Unlabeled examples).

PU learning is to put data into positive and negative examples. However, there are no labeled negative examples for learning. In this study, attempt to define a set of data in line with business norms (P) and to identify non-labeled data set (U), that contains all kinds of data. PU learning is to use and build a classifier， the positive examples will be marked out.

According to the literature [9], PU study is divided into two steps in this article:

The first step: use the rules to extract positive examples and obtain P.

The second step: establish a SVM classifier and mark the positive examples.

(2)Extraction of rule-based positive examples

Web address, or Uniform Resource Locator (URL), is the standard resource addresses on the Internet. Constitute of the basic is: “scheme://domain:port/path?query_string#fragment_ id” .

In the data collected, due to different browsers and protection of personal privacy, web address is different from a standard or full format, as shown in Fig.2. Through preliminary observation and analysis of the training data, in this case, the format of the data in the server name = [host name]. domain.[Top level domain]. Valid web address is defined as {%% top-level domain %}, as shown in Fig.2, where% represents any string. In order to improve the operating efficiency of the program and the analysis objects are from Chinese Internet users logs, so the top-level domains are all international top-level domains and China’ s national domain is known as.cn. These 20 rules are shown in Table.

Table.1 The Rule list

Rule No.	Formalized rules
1	%.%.com%
2	%.%.net%
3	%.%.cn%
……	……
20	%.%.asia%

According to the above rules, the extracted sample set of positive examples is shown in Fig. 3.

```
T<=>194[=]P<=>360chrome.exe[=]I<=>3492[=]U<=>www.cqhrss.gov.cn/u/cqhrss/[=]A<=>2027c[=]B<=>6043
e[=]V<=>5.2.0.804
T<=>312[=]P<=>360chrome.exe[=]I<=>3492[=]U<=>www.bbzkb.net[=]A<=>2027c[=]B<=>20336[=]V<=>5.2.
0.804
T<=>292[=]P<=>iexplore.exe[=]I<=>680[=]U<=>http://www.bxwx.org/text/5/5189.html[=]A<=>11058a[=]B<=
>105f8[=]V<=>8.00.6001.18702
T<=>569[=]P<=>iexplore.exe[=]I<=>3268[=]U<=>http://cn.wsj.com/gb/index.asp[=]A<=>20228[=]B<=>102fc[
=]V<=>8.00.6001.18702
T<=>1379[=]P<=>360chrome.exe[=]I<=>3492[=]U<=>www.bbzkb.net[=]A<=>2027c[=]B<=>20336[=]V<=>5.
2.0.804
T<=>1504[=]P<=>360chrome.exe[=]I<=>3492[=]U<=>www.baidu.com[=]A<=>302fe[=]B<=>NULL[=]V<=>5.
2.0.804
T<=>1700[=]P<=>360chrome.exe[=]I<=>3492[=]U<=>www.baidu.com[=]A<=>302fe[=]B<=>4030a[=]V<=>5.
2.0.804
```

Fig.3 Positive examples

(3)Building a SVM classifier

Set the training set as, where $\mathbf{x}_i$ is the input vector, y_i is its class ID. Assume that the first $t-1$ examples are positive examples P (typed 1), the rest of the data set are unmarked examples U. $\{(\mathbf{x}_1,y_1),(\mathbf{x}_2,y_2),\cdots,(\mathbf{x}_n,y_n)\}$

According to the related theory, we have this following formulation:

Minimize:

$$\frac{<\mathbf{w}\cdot\mathbf{w}>}{2}+C\sum_{i=t}^{n}\xi_i \tag{1}$$

Conditions:

$$\begin{gathered}<\mathbf{w}\cdot\mathbf{x}_i>+b\geq 1, i=1,2,\cdots,t-1\\ -1(<\mathbf{w}\cdot\mathbf{x}_i>+b)\geq 1-\xi_i, i=t,t+1,\cdots,n\\ \xi_i\geq 0, i=t,t+1,\cdots,n\end{gathered} \tag{2}$$

Where $\mathbf{W}$ is the parameter, the greater the value, the more obvious the border; ξ_i stands for slack variables; $b\subseteq R$ for bias; C is penalty coefficient, if a point belongs to a certain class, but deviates from the class and goes to other places on he border. The greater the , which shows that it does not want to give up the point, the more the boundary will shrink.

In PU learning, the positive examples are marked out and unlabeled data is defined as incomplete data.

C.Cleanup of incomplete data

Through analysis, incomplete data is divided into the following categories:

- (1) Partially-deleted, but auto-complete incomplete data.
- (2) The data is compliant with business rules, but has not been marked as positive examples, such as (www.bbc.co.uk, 10.5.5.108).
- (3) Partially deleted, and need human intervention to complement incomplete data.
- (4) Null values and other error conditions not mentioned above.

To clean up the above data, follow the (1) - (4) in order. Clean-up steps are as follows:

Step1: Compare the incomplete data and data labeled as positive examples. If the key substrings match exactly, then use the data in positive examples to complement the incomplete data.

Step2: Analyze the characteristics of the data, redefine filtering rules, and turn the legal data into a complete data by manual intervention.

Step3: Treat incomplete data in Step3 by human intervention.

Step4: Mark the results of the data acquired from the three steps as labeled positive examples.

Step5: Remove all unlabeled data from the data set.

Experiment Results and Analysis

Based on partial data cleanup, supervised learning experiment is conducted under Windows 7 operating system environment, with MyEclipse development tools and in Java language. Use LibSVM package, parameters involved in the regulation of the software is relatively small, in this study; all experiments were performed using the default parameters.

The training set has 1000 browse log files and the testing has 3000 browse log files. Each file has about 0-2021 records and the size about 1k-3000k. To reduce the impact of noise on the system, before the start of LibSVM experiment, all log records are processed to remove stop words. In this study, a total of 22 stop words are set including www, http, .com.

In order to verify the effectiveness of extraction rules and partially supervised learning, this article evaluated data cleaning effect in the experiment. For validation, comparisons were conducted 10 times during the experiment with randomly selected training set of log files (the number of files is {100,200, ... , 1000})and two performance metrics (p, r), where p is precision, r is recall. Table Ⅱ compares the performance based on different training set. Experiment results show that with the increase of training documents, the impact of increasing the number of files on the results is low, and may be abnormal. For this experiment, the number of training files between 500 and 700 is most preferred.

Table.2 Performance Comparison List

Training document number	p	r
100	**0.9579**	0.7100
300	0.9289	0.7854
500	0.9253	0.8195
600	0.9131	0.8344
700	0.9023	**0.8415**
800	0.8982	0.8242
1000	0.8945	0.8373

Conclusions

The results of this experiment show that the rule-based learning and supervised learning can both improve the efficiency of data preprocessing. This paper discussed in detail the data cleaning process based on partially supervised learning, and through the experiment, the validity of the method employed is verified, and the optimal number of training documents is ultimately determined. In this study, although Web log mining data preprocessing related problems are studied, yet there is still need for further research and improvement: First, optimization of LibSVM parameters. Second, further research of data transformation and reduction methods is needed to complete data preprocessing.

Acknowledgment

This work is supported by the Natural Science Foundation of China (NO.61563040), Natural Science Foundation of Inner Mongolia, China (NO. 2016ZD06), Natural Science Foundation in Universities of Inner Mongolia Autonomous Region, China (NO.NJZY14334).

References

[1]WANG Shi,GAO Wen, LI Jin-Tao.Path clustering: discovering the knowledgein the web site. Journal of Computer Research and Development,vol.04,2001.pp.482-486.

[2]Lenzerini, M.. Data integration: A theoretical perspective[C].//In: Popa, L. (ed.). Proceedings of the Twenty-first ACM SIGACT SIGMOD-SIGART Symposium on Principles of Database Systems (PODS 2002), 2002: 233-246.

[3]Ciszak, L.. Applications of clustering and association methods in data cleaning[C].//In: Proceedings of the International Multiconference on Computer Science and InformationTechnology, 2008,03: 97-103.

[4]Pyle, D. Data Preparation for Data Mining[M]. Morgan Kaufmann Publishers Inc., San Francisco, CA. 1999: 540.

[5]R. Cooley, B. Mobasher and J. Srivastava. Data preparation for mining World Wide Web browsing patterns[J]. Journal of Knowledge and Information Systems, 1999 , 1 (1):5-32.

[6]Tanasa, D.and B.Trousse.Advanced data preprocessing for intersites web usage mining[J]. Intelligent Systems, IEEE, 2005,19 (2): 59-65.

[7]ING Chang-bin and Chen Li. Web Log Data Preprocessing Based On Collaborative Filtering[C].// In:Proceedings of the IEEE 2nd International Workshop On Education Technology and Computer Science, 2010:118-121.

[8]DS Ngu, X. Wu. Sitehelper: A localized agent that helps incremental exploration of the World Wide Web[C]. //In: Proceedings of the 6th International World Wide Web Conference, Santa Clara, 1997:691-700.

[9]B.Liu, Y.Dai, X.Li, W.lee, and P.Yu.Building text classifiers using positive and unlabeled examples[C].//In: Proceedings of the Third IEEE International Conference on Data Mining (ICDM-2003), 2003:19-22.