A Web-based Advertising Content Analysis Platform

Rupeng Dou, Junfeng Wu
School of Automation
Harbin University of Science and Technology
Harbin, 150080, China
bill102@sohu.com

Rupeng Dou, Shuwu Zhang, Wei Liang
High-tech Innovation Center
Institute of Automation, Chinese Academy of Sciences
Beijing, 100190, China
(rpdou, swzhang, wliang)@hitic.ia.ac.cn

***Abstract* - With the booming of online economy, more and more advertisements appear on the network, meanwhile many illegal advertisements emerge. To detect the advertisements automatically, a Web-based advertising content analysis platform is proposed in this paper. This platform consists of the following three parts: web information extraction, advertiser's named entity identification and advertiser's industry identification. The experiments show the effectiveness of the proposed methods.**

Index Terms* - *Web information extraction; Advertiser's named entity identification; Advertiser's industry identification; Support Vector Machines.

I. INTRODUCTION

Nowadays, a large number of people involve in using Internet, hence it creates a prosperous market for advertisements [1]. Nevertheless, more and more illegal advertisements appear. To detect on-line advertisements automatically based on the Advertisement Law of the PRC, an advertising content analysis platform is proposed in this paper. Web information extraction, advertiser's name identification and advertiser's industry identification are three main problems for solution in our platform.

Valter Crescenzi and Alberto H.F.Laender [2, 3] propose techniques for extracting data from HTML automatically. Hai Zhao and Jian Sun [4, 5] propose Chinese named entity recognition models. Simon Tong [6] demonstrates Support vector machine active learning with applications to text classification.

The traditional methods aren't entirely suited to our platform, we chiefly deal with Chinese web pages and Advertiser's name often appears in the web page's title and metadata. The methods on these papers don't consider these special circumstances. We propose effective methods for web information extraction, advertiser's name identification and advertiser's industry identification for web advertisements. The major contribution of our platform is detecting the Web-based advertisements precisely and rapidly.

The remain of this paper is organized as follows: Section 2 describes the system framework; Section 3 gives the detailed design and implement in our platform; The details of the experimental results are presented and discussed in section 4; Finally, Section 5 gives our conclusions.[1]

II. SYSTEM FRAMEWORK

We deal chiefly with three types of advertisements, including text, picture and flash. Text Ads are from search engines' promotion advertisements such as Baidu. The pictures and flashes are collected from portal sites such as sina, sohu, etc.

Every search engine owns a large number of promotion advertisements. If a person enters a keyword in Baidu search engine, Baidu will push advertisements about the keyword to the people on the right of results' pages to attract potential customers to click these links. We built a list of keywords which we want get the advertisements about. We use the following method to acquire Baidu promotion advertisements. First, we download the webpage from a website which consists of a special character string like "http://www.baidu.com/s?wd=" and a keyword to get the web page. Then, we extract the links starting with "http://www.baidu.com/baidu.php" in the webpage. The promotion advertisements' links are gained. Other search engines' promotion advertisements are handled in the same way. Most of flashes in web sites are advertisements, only pictures will require authentication whether it is an advertisement or not.

Based on sufficient statistics, if a hyperlink from another site links to a picture and the picture contains characters, the picture is an advertisement. Therefore, a picture is an advertisement or not is judged as follows. In the first case, (1) the web site containing the picture and the web site linking to the picture are different sites; (2) the picture contains characters. In the second case, if the picture meets only one of the above two conditions, it is a pending advertisement which needs the user of the platform to confirm. In the third case, if none of the above conditions is meted, the picture isn't confirmed as an advertisement. But this method will miss a few of advertisements without characters.

The characters on picture or flash are recognized by optical character recognition (OCR), the trademark and sensitive target (for example, a national flag) on image or flash are also detected. These recognized results and the web pages linking to the picture or flash are used to extract the advertiser's name, identify advertiser's industry. Based on the Advertisement Law of the PRC, we build evaluation criterion to recognize an advertisement's legitimacy. For example, obscene or politically unacceptable content are not allowed in advertisements.

As shown in Figure 1, it's the process of a picture or flash. First, we find out whether it's a known advertisement. Then, according to the type of task, we deal with it differently. Finally, the result is written back to the database.

This work is supported by the National Key Technology R&D Program of China under Grant No.2009BAH48B02 and No.2009BAH43B04

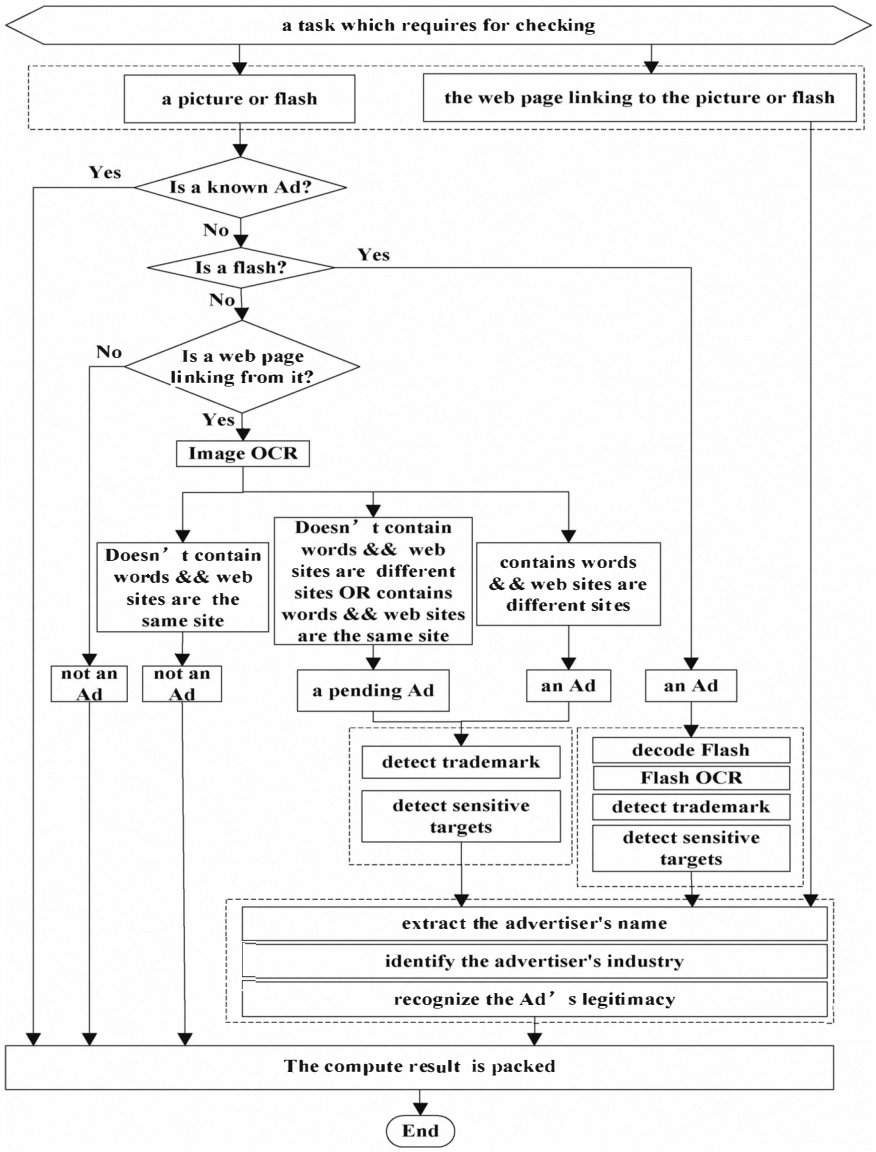

Figure 1: the process of a task

Compared with handling picture and flash, it is relatively easy to cope with the text ads. We only need to complete the following steps: extract advertiser's name, identify advertiser's industry and recognize advertisement's legitimacy.

III. Design and Implement

There is information extraction from web page, advertiser's named entity identification, advertiser's industry identification. These keys are elaborated in the following chapter.

A. Information extraction from HTML

There are many open source software for information extraction from web pages, and we use software whose name is HtmlCleaner. Because Chinese web pages accounts for the majority in our tasks, unrecognizable code is a problem to be handled when we extract information. We take the following steps to extract information in web pages.

- We use HtmlCleaner to read the web page and get coding format from metadata in the platform's default character encoding.
- The achieved coding format is used to read and parse the web page again to extract the useful information of the web page.
- We examine the extracted character string to find unrecognizable code's mark, such as the character "?".
- If the extracted character string doesn't contain the mark, we gain the correct extracted character string. Otherwise, we record the occurrence frequency of unrecognizable code's mark and the extracted character string. We change another coding format to parse it. Repeat this step, until the extracted character string doesn't contain

unrecognizable code's mark or the coding formats are used up.

If the above terminal condition is that the coding formats are used up, we get the correct extracted character string with the least unrecognizable code's mark. In the section 4 of this paper, we note that this method is easy and efficient.

B. Advertiser's name identification

Advertiser's name identification is a kind of Named Entity Recognition of organization [4]. But traditional Named Entity Recognition isn't entirely suit to our tasks. Because an advertiser often appears in a web page's title and repeats in the title and metadata, we can use the following methods to get an advertiser's name more effectively.

An advertiser's name always contains a location name and an organization name's postfix. For example, in the Advertiser's name "中国钮纽服装有限公司", "中国" is a location name and "公司" is an organization name's postfix. We build a rule to identify an advertiser's name when a phrase begins with a location name and ends in organization name's postfix. But, there is a lack of effective recognition strategy for an abbreviated advertiser's name such as "中国重工". Because a web page's characteristic features, we extract the longest repeated character string in a webpage's title and metadata. If the common string doesn't exist in the webpage, we purify the title to eliminate meaningless symbol and phrase such as "最好的", consider this purified character string to be an advertiser's name.

C. Advertiser's industry identification

At present, we chiefly deal with the illegal Ads in the medical Ads, so we need to separate the medical Ads from the others. We get the advertiser's industry in the following ways.

- If we deal with a picture or flash, we will detect the trademark on it. We have built a database which contains trademarks, trademarks' advertisers and trademarks' industries. If we get the trademark, the advertiser's industry is gained.
- If we can't get the trademark from the database, we will use text classification method. The classifier is built as follows. Firstly, we select the training dataset from the web pages in our database and separate them into two types which are medical Ads and the others. Secondly, we extract the character information from every webpage, split these character strings into phrase and use TF-IDF method [7] to compute each phrase's weighing to establish vector space model of term weight. Thirdly, we use support vector machine method trains the vector space model for building a classifier. We can use the classifier to separate the medical Ads from the others.

IV. EXPERIMENTS

The dataset is gained from Baidu promotion and contains 20152 web pages. There are 10412 distinct web pages in the dataset, because the same advertisement can be gained from distinct keywords. For example, we enter peanut and nut separately, we gain the same site whose name is "中粮我买网". After we remove duplicate web pages, we extract the title and metadata from the remaining web pages. Firstly, we use the method in this paper to only extract the web pages' title and metadata; compared with the former, we extract the web pages in the coding format extracted from the metadata of the web page. Table 1 shows the comparison of the results obtained by two methods. Our computer's CPU is Core2 CPU 2.10GHz and our computer's memory is 3GB.

TABLE I

	time(s)	web pages in unrecognizable code
this paper's method	109.815	0
contrasting method	101.873	117

As shown in Table 1, although the method in this paper spends a little more time in parsing web pages, it's much more efficient to avoid unrecognizable code. The reason for information extraction from web pages in unrecognizable code is that editing the web pages is lack of strict standardization. Sometimes web pages' editors' mistake character encoding 'gb2312' for 'utf-8', or mistake traditional Chinese character for simplified Chinese character. Losing a little efficiency make the unrecognizable code's phenomenon dropped substantially, this method is important to our platform.

Then we use the above extracted information to extract advertiser's name. We make a contrast with the proposed method and the approach proposed by Hai Zhao [4]. We conduct evaluations in terms of precision (P) and recall (R).

$$P = \frac{number\ of\ correctly\ identified\ Advertisers'\ name}{number\ of\ identified\ Advertisers'\ name} \quad (1)$$

$$R = \frac{number\ of\ correctly\ identified\ Advertisers'\ name}{number\ of\ all\ Advertisers'\ name} \quad (2)$$

TABLE II

method	time(s)	precision	recall
the proposed method	35.416	95.1%	97.3%
Hai Zhao's method	62.406	88.6%	90.5%

As shown in Table 2, our method outperforms Hai Zhao's method. Although the method proposed by Hai Zhao is outstanding in SIGHAN-6 [4], it isn't entirely suited to extract advertisers' names in network environment. Extracting advertisers' names in web pages can base on the following assumptions. First, we assume that all of the advertisers are in web pages' title and metadata. It is showed from the experiment that there are eleven exceptional web pages in five thousand. Second, to achieve a better search engine rank, web sites enable the owners' names repeat in the title and metadata of the web pages. Third, the whole title sometimes is an advertiser's name. This paper's method is directed towards these special circumstances, so it gets a better result.

V. CONCLUSION

In this paper, we propose a Web-based advertising content analysis platform and introduce the process of handling web

advertisements. There are three main parts in our platform: information extraction from HTML, advertiser's name identification and advertiser's industry identification. Experimental results demonstrate that the proposed method is practical and efficient.

ACKNOWLEDGMENT

This work has been supported by the National Key Technology R&D Program of China under Grant No.2009BAH48B02 and No.2009BAH43B04.

REFERENCES

[1] Shuguang Liu, Hao Duan, "Web-based Advertising in China: a Brief Review," 2008 International Conference on Digital Object Identifier, pp. 1-6, 2008.

[2] Valter Crescenzi, Giansalvatore Mecca, Paolo Merialdo, "RoadRunner: Towards Automatic Data Extraction from Large Web Sites," Proceedings of the 27th International Conference on Very Large Data Bases, pp. 109-118, September 2001.

[3] Alberto H. F. Laender, Berthier A. Ribeiro-Neto, Altigran S. da Silva, Juliana S. Teixeira, "A brief survey of web data extraction tools," ACM SIGMOD Record, v. 31, n. 2, June 2002.

[4] Hai Zhao and Chunyu Kit, "Unsupervised Segmentation Helps Supervised Learning of Character Tagging for Word Segmentation and Named Entity Recognition," The Sixth SIGHAN Workshop on Chinese Language Processing, Hyderabad, India, pp. 106-111, January 2008.

[5] Jian Sun, Jianfeng Gao, Lei Zhang, Ming Zhou, Changning Huang, "Chinese named entity identification using class-based language model," Proceedings of the 19th international conference on Computational linguistics, Taipei, Taiwan, pp. 1-7, August 2002.

[6] Simon Tong, Daphne Koller, "Support vector machine active learning with applications to text classification," The Journal of Machine Learning Research, Volume 2, pp. 45-66, March 2002.

[7] Raymond Kosala, Hendrik Blockeel, "Web mining research: a survey," ACM SIGKDD Explorations Newsletter, Volume 2, Issue 1, pp. 1-15, June 2000.