Web page classification based on content extraction

FIT 5190 Introduction to IT Research Methods

Lecturer: Chung-Hsing Yeh

Student: Hu Wanling 2819****

Zheng Xuan 2819****

Assignment 2: Critical Review of Published Research

2017-5-11

Faculty of Information Technology

Abstract

Nowadays, network with diverse messages has been penetrated into all aspects of society and people lives. And along with the evolution of the network technology, the information overload of network has been more and more noteworthy as it become difficult for users to conveniently and accurately locate the useful information among the network. The research of web classification can effectively strengthen the result of this process. This report reviews what has been done in the literature to classify web pages into different types. We also review several kinds of skills of web page classification. Conclusions will be drawn by proposing some assessments on the investigated algorithms and analyzing the opportunities and challenges of web page classification.

Introduction

With the rapid development of the Internet in modern society, the amount of web pages also has been increasing very fast. More and more people manage their daily business relying on the Internet, and benefit from the plentiful information contented in these pages.

The contents of web page can be divided into two parts, one part is called main body, which is the theme of the web page content, and another part is nothing to do with the body parts, namely "noise" section (i.e., navigation links, links, advertising, copyright statement, etc.). However, due to several kinds of reasons, noise is embedded in and occupies 40%-50% of the web pages, which has brought difficulties for text analysis task to a great extent. In order to effectively filter the useless pages, and maximumly utilize the information of web pages, it is necessary to classify them.

Web page classification is usually based on two strategies: manual classification and automatic classification. Manual classification relies on domain experts to classify web pages. A typical example of this strategy is Yahoo (www.yahoo.com). Automatic classification is to pre-build a classifier, the training set is input to the classifier, then train the classifier. Classifier to obtain relevant knowledge to prepare the next work of classification. Although the accuracy of manual classification is high, but the efficiency is not. As the number of pages increase, manual classification has been unable to meet the actual needs.

Scope and method

All of the 14 relative articles are searched from academic papers and the database of ACM, PKDD and IEEE with some keywords, such as web classification, content extraction and feature extraction. The criterion of the selections is that the chosen articles are influential and have certain association with this topic. All of these correlative articles were published later than 2000. Because the previous studies proposed plenty of different methods about web page classification, this paper will mainly review the automatic classification. Details of methods and algorithms as well as strength and weakness will be explained in the following sections. At the same time, this paper only focuses on the training of thought in design. The details of how to realize the process and analyze the result of those methods will not be discussed in this paper.

Body of the review

Nowadays, the web classification technology will develop towards three directions [1]: (1) Using text classification algorithms. As the basic parts of web page is obviously texts, we could through the text information extraction to complete the web classification using appropriate text classification algorithms. In this field, the vector space model has become the most extensive representation. It represents documents as a series of vector with disorder key words, and after transferring these documents into vectors with different weights, some traditional classification algorithms, like class center vector algorithm, KNN or support vector machine (SVM), will be used for classification. (2) Classification based on web page features. Due to the complexity and accompanying noise of the web pages, it cannot achieve the best classification effectiveness to utilize all the text content. Therefore, researchers should use some other characteristics of a web page to filter the text content, and in this way to improve the accuracy of classification. (3) Using the adjacent information page. Sometimes those useful characteristics in web pages may not exist or is likely to be unable to identify. The features related to the original page that extracted from neighboring web pages could be used as supplementary information to overcome the problems. Just like mentioned in by Qi et al. [2] the using of neighboring information that obtained from web link graph will evidently enhances the result of the original web pages' classification.

According to the categories of websites, the classification of web page has been found to have three features [3]: (1) Web page could be regarded as a mixture of multi-view data, because the web contains usually consist of two or more varietiesrain of of data (e.g., text, hyperlinks and images), and each variety of data can be called a view. (2) The classification of websites is a semi-supervised application. Compared to unlabeled ones, labeled pages are difficult to be gathered on the Internet, because the processes of labeling are comparatively expensive and usually consume more time. (3) Date in web page is high-dimensional. The data usually contains much miscellaneous information, and it is a challenge for classification tasks to extract low-dimensional information and useful characteristics from the web page data.

Context extraction

Web could be classified into three varieties according to the web content [4]. The first kind called topic-specific web, it can be defined as a page that use paragraphs of words to describe one or more themes. Rarely pictures, videos and links will be displayed in this type of web. The second kind of web page called Hub web, which always provides hyperlinks for relevant websites and seldom does these web pages describe a specific thing. The third kind web page named multimedia web page, where pictures and videos occupy the most spaces of the web page and words are displayed only as a statement for multimedia.

As early as 2001, Rahman and his team has proposed the concept of "content extraction". Content extraction plays an important role in Web mining, natural language processing (NLP), information check Rope, and displaying web pages in a small screen size devices (mobile phone, PAD, etc.) [5]. aiming at topic-specific Web pages, Xiong [6] and the team have put forward a new management to extract the main content of web pages. In the process of pre-filter, the method removed the independent elements in the web page using regular expressions, and divided the web page into linear blocks according to the structure characteristics, then divided each block into link blocks and text blocks according to the text features. At the same time, the method completed body positioning by using the result of the successive appearance of noise blocks to get the main text information of web page.

Content extraction relies on text features and structure analysis. The method with strong versatility of content also combines the text characteristic and of the web page structure. This method use building Chinese topic crawler as the application background, though the extraction of web content to avoid the influence of topic correlation calculation that caused by "noise" parts, at the same time save the storage space of the system.

Visual features

In order to complete the automatic classification and processing of web pages, textual content that consist of the displayed content, which is constructed by color, type fonts, images and videos, and the connotative (HTML) information always be used. Viktor de Boer and Maarten van Someren proposed a new method that can be taken advantage of classifying the web pages into several different types [7].

They had used the visual appearances of web pages to achieve the classification, and their approach was based on using the web pages as they appear to users. They used machine learning and divide the process into two steps. At first, they selected related feature using chi-square standard and then constructed a classifier using the Nave Bayes classifier. In the research, the researchers rendered an image for each web page using web browser. For every page, they saved screen shots and stored them as .PNG files. They had built two binary classification based on aesthetic value and decency. The results showed that these global visual features, like color and edge histograms, Gabor and texture features, already produced good classification results.

Spectral Hashing

Some experts have proposed some web page classification algorithms based on statistics and machine learning, including KNN algorithms, decision tree, support vector machine [8], naive Bayes probability model [9] and neural network [10]. These classification algorithms are effective in dealing with small amount of web pages. However, the efficiency of these methods is not satisfactory when dealing with large-scale web pages. Therefore, a distributed classification method is proposed. This method can meet the requirement of classification efficiency, but it does not improve the classification algorithm. So, Tian [11] propose a hash algorithm which is based on hash and KNN to design a classification algorithm for applying to large scale web pages.

The method of traditional classification is to express the feature information in a web page as a point in a high dimensional space. When this method is used to represent web pages, the dimension of the vector space can reach the world-wide web, and the operation cost of the high dimensional vector makes it difficult for them to cope with the large-scale web page classification. However, simple reduction of feature terms and reduction of feature space dimension will affect the accuracy of classifiers. So, the solution should not affect the accuracy of the classifier on the premise, and the dimension of the original page feature space is reduced. Spectral Hashing [12] is an ideal method, which can be used to represent the characteristics of the web page with short hash code, while retaining the original web page features, after the hash, web pages can significantly reduce the dimension. The basic idea of spectral hashing is to map the high-dimensional space vector to the low dimensional Hamming space [13], and keep the original space vector similarity, so that the Hamming distance of the new space vector reflects the similarity of the original space vector.

Tian uses TF-IDF to assist in dealing with the words in the web page. TF-IDF is a statistical method used to evaluate the importance of a word in a file set or a file in a corpus. The importance of the word is proportional to the number of times it appears in the document, but it is inversely proportional to the frequency it appears in the corpus. The main idea of TF-IDF is: If a word or phrase in an article appears in the frequency TF high, and rarely appear in other articles, think that the word or phrase has a good class distinction, suitable for classification. K-Nearest Neighbor [14] algorithm is one of the most mature methods and one of the simplest machine learning algorithms.

Interpretation and conclusion

Web page classification deal with data according to the information which carried by the web page. Web page can be stored in the corresponding database according to the category. It is convenient for users to quickly and accurately find their own pages. Web page classification can not only help improve the efficiency of search engines and rate of find all the information and accuracy, but also play a virtual role in network security management as one of the key technologies.

The method of hashes and KNN which is based on the design of a classification algorithm is an improve method of KNN. It is feasible to apply on large-scale web classification. Another method of web page classification is visual features.

How to use more specific visual features to significantly boost the accuracy. Form this method, we know that more local feature can also have a positive effect. By identifying different visual elements on a web page (photograph, text block, banner, etc.), more abstract features which can be used to better classify the page can be constructed. If we want to further strengthen the method of visual features which apply on the web page classification, in the future work, we should be focused on the integration of these visual features with other features of web page. This includes the textual context and the underlying HTML, the used technologies and functionalities of a web page. By combining visual and the underlying HTML, we can better identify elements on a web page, which can be used as better features for classification. The analysis of the visual appearance of a web page can be combined with analysis based on textual content, technological implementation, functionalities or usage data. Another possible expansion of the tool's functionality is that users can define their own web site topics. Through this web site we are looking towards gaining much more data and user evaluations of that data. This method can effectively reduce the memory overhead and time overhead. The way about extraction textual elements of multi-types webpages by fusing structure and content of the webpage could help to effectively use the webpage related features in the design structure and text content.

But we still should consider how to find a more appropriate algorithm and its improved method. So, the further work will consider the relationship in the structure and context between the three elements of Deeping into the web page and the other parts of the page. In particular, it is important to find the candidate block of page context for better results. All of these studies have the space for further improvement.

In this review, we focus on the method of classification that the experts used in different web environment. We review some papers about feature extraction, context extraction and KNN algorithm in the web page classification. We introduction those method and algorithm which apply in the classification of web pages, how to use those methods and its improvement in the web page classification. We analyze the strengths and shortcomings of those solutions, and its future in-depth study should take into account the direction. The purpose of reviewing papers in context and content extraction of the web page classification is to apply the method and algorithm proposed in this field to solve the similar problems in web page and improve the efficiency of classification.

References

[1] Qianxi Chen, Lei Fan. Webpage Classification Based on Deep Learning Algorithm. *Microcomputer Applications*, 2016, 32(2): 25-28.

[2] Qi X, Davison B D. Knowing a web page by the company it keeps. Proceedings of the 15th ACM international conference on Information and knowledge management. ACM, 2006: 228-237.

[3] Web Page Classification Based on Uncorrelated Semi-Supervised Intra-View and Inter-View Manifold Discriminant Feature Extraction.

[4] Zhigang Zhang, Jing Chen, Xiaoming Li. An HTML web page purification method. *Information Learned Journal*, 2004, 23(4): 387-393.

[5] Rahman A, Alam H, Hartono R. Content Extraction from HTML Documents. Proc. of the 1st International Workshop on Web Document Analysis. New York, USA: ACM Press, 2001.

[6] Zhongyang Xiong, et al. Content Extraction Method Combining Web Page Structure and Text Feature. *Computer Engineering*, 2013, 39(12): 200-210.

[7] Viktor de Boer et al. Classifying Web Pages with Visual Features. *International Conference on Webist*, 2010 :245-252.

[8] Adankon, M. M., & Cheriet, M. (2002). Support vector machine. *Computer Science,* 1(4), 1-28.

[9] Qin, Z. (2006). Naive Bayes Classification Given Probability Estimation Trees. *International Conference on Machine Learning and Applications* (pp.34-42). IEEE.

[10] Liu, Y., Wei, F., Li, S., Ji, H., Zhou, M., & Wang, H. (2015). A dependency-based neural network for relation classification. *Computer Science*.

[11] Weiss, Y., Torralba, A. and Fergus, R. (2008) Spectral Hashing. *Neural Information Processing Systems*,

282, 1753-1760.

[12] Dandan Chen. Large Scale Web Page Classification Algorithm Based on Spectral Hashing, 2-16, 5(1), 65-74.

[13] Song, Y., Huang, J and Zhou, D., (2007) IKNN: Informative K-Nearest Neighbor Pattern Classification. Knowledge Discovery in Databases: PKDD. Springer Berlin Heidelberg, 248-264.

[14] Charon, I., Cohen, G., *et al.* (2010) New Identifying Codes in the Binary Hamming Space. *European Journal of Combinatorics*, 31(1), 491-501.