FIT5190 Introduction to IT Research Methods
Assignment 3

Web Page Segmentation

Group 1

Yang Yan 2759****
Xianqiang Gao 2731****

30 May 2016

Southeast University – Monash University

Joint Graduate School

Web page segmentation

Abstract

Web pages are typically designed for visual interaction which include various visual segments. However, these segments are not explicitly declared in the source code. Therefore, web page segmentation aims at dividing a web page into visually and semantically coherent segments called blocks. In the scenario of Web information archiving, detecting important blocks can facilitate the crawling optimization. We propose a new web page segmentation algorithm to better adapt to the ever-more complicated web page structures and obtain a proper segmentation granularity. To identify important blocks, we apply a semi-supervised model using limited labeled blocks and large amount of unlabeled blocks as training set to classify block importance. It can have a good classification performance and handle large scale web pages efficiently.

Keywords–Web page segmentation; Web archiving; block importance; semi-supervised model

Introduction

Web pages are typically designed to facilitate visual interaction with the human readers. Web page designers normally organize the web page information into different units or functional types, which are arranged in coherent visual segments in the page, such as header, footer, navigation menu, major content, etc. However, these visual segments are not explicitly declared in the source code. What we can find from the source code are the list items, paragraphs, instead of clear visual segmentation or pattern. But automatically identifying different segments from web pages can be very useful for different fields (Bing et al., 2014). Vision-based Page Segmentation (VIPS) has been widely used to divide a web page into visually and semantically coherent segments called blocks. It extracts semantic blocks of a web page by utilizing heuristic rules based on the DOM (Document Object Model) representation as well as visual features (Cai et al., 2003).

Obviously, the information in a web page is not equally important. Different blocks inside a web page have different importance weights according to their location, occupied area, content (Song et al., 2004). In the context of Web information archiving, segmentation can be used to extract interesting parts (Sanoja et al., 2015). It is useful for crawling optimization to detect changes in interesting areas from distinct visions of a page. Thus identifying important blocks is of great advantage. Given a web page partition, Song et al. (2004) propose Support Vector Machines (SVM) and neural network methods to learn general block importance models, while Zhang et al. (2013) and Shen et al. (2014) use some defined heuristics to detect informative blocks.

Objectives

Segmenting a web page may be one of initial steps of Web information archiving performed on that page. Due to the increasing complicated structure of web pages and ever-changing web design, the rules in VIPS become numerous and are no longer fully applicable (Wei et al., 2015). To handle this problem, we propose a new vision-based web page segmentation algorithm based on VIPS to segment a web page into different blocks.

In our research, we define headlines and news link lists as important blocks or interesting areas. A semi-supervised learning model can be applied to detect the important blocks in a web page using limited labeled blocks and a large amount of unlabeled blocks as training set. It can achieve a high learning accuracy and also can be time-saving and efficient with limited labeled blocks. Therefore, to maintain a web archive up-to-date, crawlers can harvest the web by iteratively downloading new versions of documents in which important blocks have changed without wasting time and space for storing useless page versions.

Methodology

VIPS requires that the page layout information must be included in the page tags. But CSS (Cascading Style Sheets) which has been widely used in the design of most web pages brings the problem of separation of page content and page layout. In our algorithm, we first preprocess the web pages to combine the web page content with layout information to fully use style information, and then reconstruct DOM tree structure considering the DIV and CSS page layout. Furthermore, VIPS segments the web pages based on heuristic rules. The segmentation granularity of blocks is controlled by a pre-defined DoC (Degree of Coherence) value. The smaller the pDoC value is, the coarser the segmented content structure would be. How to define a proper pDoC value remains to be solved. Also, VIPS measures the block coherence based on visual characteristics and ignores the semantic association in blocks which makes the segmentation granularity of the same subject too small. Our algorithm can firstly get the text content based on the fine-grained blocks segmented by VIPS and then combine the blocks which have high content similarity. The final segmented blocks can have a better semantic consistency and particle size to facilitate the identification of important blocks.

We propose a tool to do the web page segmentation and label of segmented blocks by JAVA. We collect 300 URLs of different news websites which cover finance, travel, culture, etc. Then these URL links are loaded into the tool to obtain the segmented blocks of the web pages by invoking segmentation algorithm. After rendering the web pages, our proposed tool allows us to see the visual location of the segmented blocks. What’s more, the spatial, visual and content features of each block are automatically extracted by utilizing CSSBox (X)HTML/CSS rendering engine. We can manually

label the importance of each block. These features and labels will be used in the classification.

Many machine-learning researchers have found that unlabeled data, when used in conjunction with a small amount of labeled data, can produce considerable improvement in learning accuracy. The cost associated with the labeling process thus may render a fully labeled training set infeasible, whereas acquisition of unlabeled data is relatively inexpensive. Based on the segmented blocks of the web pages, we define limited labeled blocks and large amount of unlabeled blocks to be the input of our semi-supervised model, while the output are the classification of block importance. Spatial, visual and content features are extracted to represent the segmented input blocks. What's more, we apply label propagation algorithm to train the semi-supervised model. Label propagation denotes a few variations of semi-supervised graph inference algorithms (Zhu et al., 2002). It can be used for classification of our block importance due to the actual graph structure of web pages.

Novelty

Web pages are getting more complex than ever. Web page designers tend to separate the web page content and page layout which definitely challenges VIPS algorithm. Based on VIPS, our segmentation algorithm can preprocess the web pages to extract the layout information and integrate them with the page content. Also we will reconstruct the DOM tree considering DIV tag and CSS layout without only focusing on Table tag. A pre-defined DoC value makes VIPS hard to control the segmentation granularity facing different kinds of web page contents. We try to calculate the text similarity between fine-grained blocks segmented by VIPS, and combine those high similar blocks to avoid the situation that the contents of same subjects have too small segmentation granularity. Therefore, compared with VIPS, our segmentation algorithm can better handle web pages with complex structures and have a proper segmentation granularity.

Identifying important blocks is significant for Web information archiving. Based on the blocks segmented by VIPS, Support Vector Machines (SVM) and neural network methods are applied to learn general block importance models using manually labeled blocks as training set (Song et al., 2004). In the scenario of large scale web pages, the cost associated with the labeling process would render a fully labeled training set infeasible. Another trend of detecting informative blocks is heuristic-based (Zhang et al., 2013 and Shen et al., 2014). A finite set of rules also cannot handle ever-increasing web pages. In our research, we adopt a semi-supervised model implemented with label propagation algorithm. Limited number of labeled blocks in the training set are provided by our proposed tool, while large amount of unlabeled blocks can be acquired relative inexpensively. We do not need to cost much time and effort to do the labelling of training set or come up with new rules to cover all the situation. In conclusion, combining our segmentation algorithm and semi-supervised model can not only handle complex web pages but also achieve high accuracy in identifying important blocks.

Conclusion and Significance

In our research, in order to segment ever-more complicated web pages, we propose a new segmentation algorithm based on VIPS. Our algorithm can overcome the limitations of VIPS and provide more precise segmentation considering new techniques used in the design of web pages. Based on our segmentation algorithm and proposed tool, a semi-supervised model using limited number of labeled blocks and a large number of unlabeled blocks as training set is applied to identify important blocks. The classification accuracy level of block importance in our model reaches 97%. To achieve the same performance, a supervised learning model would need numerous labeled blocks which will cost much time and effort to acquire under the circumstance of large scale web pages. In the context of Web information archiving, our method of combining proposed segmentation algorithm with semi-supervised model will definitely help the crawler download “the most important information” to efficiently store and index data.

References

Bing, L., Guo, R., Lam, W., Niu, Z. Y., & Wang, H. (2014). Web page segmentation with structured prediction and its application in web page classification. In *Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval* (pp. 767-776). ACM.

Cai, D., Yu, S., Wen, J. R., & Ma, W. Y. (2003). *VIPS: a visionbased page segmentation algorithm* (p. 28). Microsoft technical report, MSR-TR-2003-79.

Sanoja, A., & Gançarski, S. (2015). Web page segmentation evaluation. In *Proceedings of the 30th Annual ACM Symposium on Applied Computing* (pp. 753-760). ACM.

Shen, B., Li, L., & Wang, N. W. (2014). Application on Web Page Filtering Technology. *International Journal of Multimedia and Ubiquitous Engineering*,*9*(12), 405-420.

Song, R., Liu, H., Wen, J. R., & Ma, W. Y. (2004). Learning block importance models for web pages. In *Proceedings of the 13th international conference on World Wide Web* (pp. 203-211). ACM.

Wei, T., Lu, Y., Li, X., & Liu, J. (2015). Web page segmentation based on the hough transform and vision cues. In *2015 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA)* (pp. 865-872). IEEE.

Zhang, X., Zhang, Y., He, J., & Cobia, F. (2013). Vision-Based Web Page Block Segmentation and Informative Block Detection. In *Web Intelligence (WI) and Intelligent Agent Technologies (IAT), 2013 IEEE/WIC/ACM International Joint Conferences on* (Vol. 3, pp. 265-269). IEEE.

Zhu, X., & Ghahramani, Z. (2002). *Learning from labeled and unlabeled data with label propagation*. Technical Report CMU-CALD-02-107, Carnegie Mellon University.