A Literature Review on Web Page Segmentation

Yang Yan 2759****

Xianqiang Gao 2731****

SOUTHEAST UNIVERSITY - MONASH UNIVERSITY JOINT GRADUATE SCHOOL

13 May 2016

Abstract

Web page segmentation aims to divide a web page into segments that reveal the information presentation logic of the page designer and show coherent structure to the readers. This report reviews what has been done in the literature to automatically identify such segments in a web page. We also review state of art of the algorithms that measure the importance of different segments in web information archiving and mining. Conclusions will be drawn by proposing some assessments on the investigated algorithms and analyzing the opportunities and challenges of web page segmentation in the fields of Web information archiving and mining.

Introduction

Web pages are typically designed for visual interaction. Web page designer normally organizes 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. Therefore web page segmentation aims at dividing a web page into visually and semantically coherent segments called blocks.

Scope and method

Web page segmentation has been proposed to address problems in different fields including mobile web, duplicate detection, information retrieval, web page clustering, etc. Among these, we mainly focus on the fields of Web information archiving and mining which all need to measure the importance of different segments in a web page. Different algorithms are designed for different purposes based on what they want to achieve in different fields.

Firstly, we summarize the problems that web page segmentation has addressed in the fields of Web information archiving and mining. Secondly, a notable technique Vision-based Page Segmentation (VIPS) algorithm and the other two algorithms which segment web pages on the basis of the VIPS are investigated. Thirdly, we choose several existing algorithms which are related to Web information archiving and mining and research what they do with the identified segments. Finally we briefly conclude the pros and cons of the chosen algorithms.

Body of review

Web page segmentation means the process of dividing a web page into visually and semantically coherent segments called blocks. Detecting these different blocks is an important step for many applications. Here are some fields of Web information archiving and mining in which web page segmentation can be used to address the problems.

Web information archiving and mining

Archiving: Archiving the web has become important for keeping useful information source. Crawlers harvest the web by periodically downloading the latest versions of web content to guarantee a web archive up-to-date. However, the situation that crawlers download a new page version with unimportant changes frequently occurs. Furthermore, due to large scale useless data stored, querying such archive can be expensive (Saad et al., 2010). Under this circumstance, web page segmentation can be used to extract interesting parts to be stored. By giving relative weights to blocks according to their importance, it allows the detection of relevant changes (changes in important blocks) from distinct versions of a page. This is useful for crawling optimization, as it allows tuning of crawlers so that they will revisit pages with important changes more often (Pehlivan et al., 2010).

Information Extraction: Some researches have focused on trying to build wrappers to structure web data to break the limitations of web browsing and keyword searching. But for these wrappers, block information is missing (Cai et al., 2003). Segments in a web page that contains noise information which is irrelevant to the main content of the page can easily harm the outcome of data mining. Therefore identifying the importance and role of web page components plays a significant role in web information mining. Web page segmentation can identify informative and unimportant content blocks, and support information extraction by making use of the features in the informative content blocks (Lin et al., 2002).

Information Retrieval: Traditionally, link analysis assumes that a link represents a relationship between two web pages (A relates to B), however a link might represent a relationship between parts of web pages (part of A relates to part of B). Furthermore, noise in a web page can cause topic drift problem (Cai et al., 2003). There are two major issues: if a web page is considered as a single semantic unit then it does not consider multiple topics in a web page, which means topics can be scattered at various regions of the page, it can cause low retrieval precision; if the web page contains multiple unrelated topics, then the calculation of correlations among terms in a web page may be inappropriate (Cai et al., 2004). Web page segmentation can be used to improve the precision of information retrieval. This is mainly because during a search the most important segments which are discovered after segmenting the web page can be treated more important which would give better search performance.

Vision-based web page segmentation algorithms

Since the earlier 2000's, web page segmentation has been a very active research area. Several works are published, showing that the visual aspect of the page is a key for segmenting Web pages.

VIPS algorithm extracts semantic content structure of a web page by utilizing heuristic rules based on the DOM (Document Object Model) representation as well as visual features. Semantic content structure is hierarchical which can also be called block. Every block gets a DoC (Degree of Coherence) value which indicates the degree that the content of the block coheres based on view perception. VIPS algorithm detects web content structure following the automatic top-down and tag-tree independent principles. It has three main steps: block extraction, separator detection and content structure construction. It first extracts all suitable blocks from the HTML DOM tree, then tries to find the separators between those blocks, finally the semantic structure of the web page is constructed based on these separator. This process will be implemented recursively until the DoC value of the block is greater than pre-defined value. VIPS algorithm takes perfect advantages of page structure feature and obtain a better segmentation of a web page at semantic level (Cai et al., 2003).

Due to the increasing complicated structure of web pages and ever-changing of web design, the rules in VIPS become numerous and are no longer fully applicable (Wei et al., 2015). We choose two vision-based web page segmentation algorithms which may alleviate the shortcoming of VIPS.

Integrating VIPS algorithm with the Progressive Probabilistic Hough Transform (PPHT) in image processing for detecting lines can compensate the shortcoming of VIPS. On the basis of image processing technology and characteristics of web page, this new approach introduces Hough transform in image processing and takes advantage from DOM tree and visual cues. Hough transform is used to detect separator in web page and can perform well in real-time applications with a fixed amount of available processing time. The overall steps include investigating the characteristic of separator of web pages, extracting the visual separator in web pages according to the perceptive of web designers, and adjusting the segmented information blocks by Hough transformation in the hope of enhancing the VIPS algorithm compatibility. The new approach can not only effectively extract the separator which uses a very thin image indicating a line, but also segment the large blocks into many semantic independent sub-blocks (Wei et al., 2015).

Block-o-Matic (BoM) framework is a hybrid method which integrates visual-based segmentation approach and automated document processing technique. To make full use of web page content, user’s visual understanding such as flow order and block classification should be taken into account. The algorithm consists of three phases: analysis, understanding and reconstruction. Three corresponding structures are

involved: DOM tree, content structure and logical structure. The content structure indicates the position of objects in the web page and the classification of the objects. The logic structure indicates the associations between blocks from user's perspective. BoM defines a Web Page Segmentation Model which firstly builds the content structure from the DOM tree with the d2c algorithm, then maps the content structure into a logical structure with the c2l algorithm, finally creates the segmented web page by merging the three structures, matching the order flow. Experimentation shows the result of BoM algorithm is better than VIPS (Sanoja et al., 2014).

Algorithms for detecting important segments or interesting areas

An accurate detection of interesting areas or important information we want from a web page will surely improve the performance of Web information archiving and mining. To measure the importance of different segments, machine learning algorithm and heuristic rules are proposed in the literature.

Several methods has been proposed to segment a web page into different blocks. Though these methods do not treat a web page as a unit, the blocks in a web page are treated uniformly. A user study shows that people hold coincident opinions about the semantic value of the same block in a web page which validates that there exists a block importance model. A block importance model will assign importance values to blocks in a web page according to their property. To get general block importance models, firstly, the method use VIPS algorithm to do segmentation work on a web page by comparing the visual structure of the web page. VIPS can assemble relevant contents together because it can availably distinguish semantically blocks from each other. Then the block attributes (including content attributes and spatial attributes) are extracted to denote the features of the blocks. Finally, according to these features, Support Vector Machines (SVM) and neural network methods are implemented using manually labeled blocks as training set (Song et al., 2004).

DVPS algorithm can filter out the redundant and uncorrelated information and gain valuable information in a web page with heuristic rules. Based on the most frequently used DIV + CSS page layout, the algorithm proposes a DIV_FOREST web representation model for this current used web design standards. Firstly, by analyzing the DOM tree page representation, the page is presented as a tree structure. DIV_FOREST model can reserve the structure and semantic correlation of nodes in the web page according to the DIV tag. Secondly, after DIV_FOREST tree structure is produced, this algorithm divides DIV_FOREST model into several basic DIV data unit, mapping to the visual block on the page. Finally, according to the division of the page, DVPS algorithm respectively extracts and quantifies block DIV semantic feature, spatial characteristics, and visual features which are the standards to determine the type of the block and to assign importance degree to the block. Combining the semantic information and visual representation of block, DVPS algorithm can effectively improve the block partition of the page and the effect of data filtering (Shen et al., 2014).

Interpretation and conclusion

Web page segmentation is regarded as an essential task in Web information archiving and mining which all need to measure the importance of different segments in a web page. VIPS algorithm extracts semantic blocks of a web page by utilizing heuristic rules based on the DOM tree as well as visual features, showing that the visual aspect of the page is a key for segmenting Web pages. It is a notable technique which improves the situation that the visual effect is not consistent with the corresponding DOM tree in a web page (Wei et al., 2015). However rules in VIPS become numerous and not fully applicable due to complicated web page structure. To address this problem, several vision-based algorithms are inspired to go beyond heuristic methods and they are proved to have better performance. Furthermore, with the increasing number of web pages, neither training the block importance model with manually labeled blocks nor heuristic rules can handle large scale web pages in the fields of Web information archiving and mining. Future work can be done by integrating a new vision-based web page segmentation algorithm and a semi-supervised learning model to identify important blocks or interesting areas in web pages.

References

Saad, M. B., & Gançarski, S. (2010). Using visual pages analysis for optimizing web archiving. In *Proceedings of the 2010 EDBT/ICDT Workshops* (p. 43). ACM.

Pehlivan, Z., Ben-Saad, M., & Gançarski, S. (2010). Vi-DIFF: Understanding web pages changes. In *Database and Expert Systems Applications* (pp. 1-15). Springer Berlin Heidelberg.

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.

Lin, S. H., & Ho, J. M. (2002). Discovering informative content blocks from Web documents. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 588-593). ACM.

Cai, D., Yu, S., Wen, J. R., & Ma, W. Y. (2004). Block-based web search. In *Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval* (pp. 456-463). 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.

Sanoja, A., & Gancarski, S. (2014). Block-o-matic: A web page segmentation framework. In *Multimedia Computing and Systems (ICMCS), 2014 International Conference on* (pp. 595-600). IEEE.

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.

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.