Block Extraction and Page Segmentation for Block-Level Web Search Engine

曾, 駿

https://doi.org/10.15017/1398392

出版情報：九州大学, 2013, 博士（情報科学）, 課程博士
バージョン：

（別紙様式２）

氏　　名　：曾　駿

論文題名　：Block Extraction and Page Segmentation for Block-Level Web Search Engine
（ブロックレベルウェブ検索エンジンのためのブロック抽出およびページ分割）

区　　分　：（甲）　乙

論　文　内　容　の　要　旨

Along with the rapid development of web, the quantity of information available on the web today is more than at any point in history. Due to the popularity of web search engines, finding relevant web pages only takes less than one second. This is because before the query terms are submitted, a web search engine has built an index for each word to indicate a list of web pages where the word appears. However, even if web search engines can provide relevant web pages in such a short time, people still need to spend a lot of time reading the pages to find the relevant parts of a web page. An ideal goal of web search engine would be a “block-level” search engine where each web page is segmented into non-overlapped regions of an appropriate size. There have been many researches on extracting relevant parts from web pages for users’ query. However, these approaches are not concerning with covering all the contents of a web page and making an index for the sub-pages of a web page. In other words, it is necessary to formalize the appropriate parts of a web page for a query and to segment a web page into non-overlapped parts in order to realize a “block-level” web search engine. This thesis considers rectangular regions as the index targets and solves these two problems.

The first problem is known as block extraction from a web page with respect to a query. The required blocks may vary according to query. Thus, it is necessary to determine what kind of block is more likely to be a required block with respect to the query. Firstly, we consider the leaf blocks that satisfy the query. Then we analyze the blocks in the path from the leaf blocks to the root of HTML-tree. We analyzed the features of blocks, such as: the text quantity, DOM (Document Object Model) tree depth, number of child blocks, etc. We manually labeled the required blocks from a set of web pages and utilized SVM (Support Vector Machine) to train these features of the labeled blocks in order to determine which feature is most effective in detecting the required blocks. Based on the analysis results, we defined the score of a block as a combination of word weight and block depth. The score is used as a ranking measure of blocks for a query. The experiment results indicated that the proposed method is effective to extract the required block and useful to reduce users’ search time.

The second problem is known as page segmentation. The purpose of page segmentation is to divide a web page into independent segments. We use the visual features of blocks to segment web pages. First, we consider the record blocks where similar kinds of data are displayed adjacent to each other. The

automatically generated outputs from databases are typical examples of these record blocks. We give a formulation of the similarity of blocks and introduce the notion of “layout tree”. For two given blocks, they are first transformed into two layout trees. Then the Tree Edit Distance algorithm is used to calculate the distance of the two layout trees. If the distance is less than the threshold, then the two blocks have similar layout. By using the layout tree, we can recognize the data record blocks in a given web page. We used this similarity to cluster blocks of a web page. We introduce two other measurements of seem degree and content similarity of two blocks. The seam degree describes how neatly the blocks are arranged. The content similarity describes the similarity of contents. The method first recognizes and marks the data record blocks in advance. According to the seam degree and content similarity, it can be determined whether a block should be divided or not. The experiment results show that the proposed method can divide a web page into appropriate suitable semantic segments.

〔作成要領〕

１．用紙はＡ４判上質紙を使用すること。

２．本文の文字サイズは１０．５ポイント（「論文内容の要旨」の文字は１２ポイント）
　　１行の字数４４字，行数４２行、余白（左右２０mm，上下２５mm程度）をあけ，頁数は記入しない。

３．要旨は１頁に２，０００字程度（最大２頁以内を目安）にまとめる。

４．図表・図式等は随意に使用のこと。

５．氏名は外国人の場合，カタカナ表記（漢字圏の学生は漢字）で記入する。

６．論文題名は論文目録と合わせる。（外国語の場合は和訳をカッコ書きで付記する。）

７．区分には甲または乙を明示すること。

この原稿は，「九州大学博士学位論文内容の要旨及び審査結果の要旨」の原稿としてオフセット印刷するので，鮮明な原稿をクリップ止めで提出すること。