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：

氏名・(本籍・国籍)	ソウ シュン 曾 駿 (中 国)
学位の種類	博士 (情報科学)
学位記番号	シ情博甲第５１２号
学位授与の日付	平成25年9月24日
学位授与の要件	学位規則第４条第１項該当 システム情報科学府 情報知能工学専攻
学位論文題目	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)
論文調査委員	(主 査) 教 授 廣 川 佐千男 (副 査) 教 授 荒 木 啓二郎 准教授 峯 恒 憲

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.

論 文 審 査 の 結 果 の 要 旨

本論文は，ブラウザに表示されたときの視覚情報とHTMLの構造を融合した構造化手法としてのレイアウトツリー，それに基づきページ中の重要部分を抽出する手法，重要部分以外の部分についても隣接するブロックの類似度を評価する二つの尺度とそれに基づく分割アルゴリズムをそれぞれ提案し，これらの有効性を評価実験により示したものであり，情報科学について重要な知見を得たものとして価値ある業績であると認める。