Studies on Detecting Uninformative Blocks from Web Pages

Tatsuya Nakamura (510073)

School of Information Science,
Japan Advanced Institute of Science and Technology

February 8, 2007

Keywords: WWW, uninformative block, chunking, learning, information retrieval.

By the recent growth of WWW, various information can be obtained from webpages. There are many studies on Web such as information retrieval or Web mining. But both informative and uninformative blocks exist in Web pages. Uninformative blocks are fragments of Web pages without any useful information, e.g., advertisements, table of contents of pages, search form, etc. It is possible to expect negative effect of uninformative blocks on various applications such as information retrieval or web mining. The processing time of indexing and retrieval is shortened if uninformative blocks are detected and words in such blocks are not used as index terms. Furthermore, retrieving unsuitable pages is prevented by ignoreing keywords in uninformative blocks, so that improvement of information retrieval can be expected. Therefore automatically detecting uninformative blocks is useful for many Web applications.

In this research, we propose a method which detect uninformative blocks of Web pages as the preprocess of various Web applications. The definition of uninformative blocks may be different for Web application. In this research we assume information retrieval is a target application. Our method classifies text which are divided by HTML tag into two classes, informative or not. Uninformative blocks consist of multiple text. So uninformative blocks are detected by chunking of IOB2 model. The chunking model is learned from Web pages annotated with correct uninformative blocks. Our system used YamCha as a chunker. YamCha is chunking tool in general purpose.

Next we explain features used for learning. We searched 21 Web pages different from experiment data. And we find effective features for detecting uninformative blocks. As a result, we set eight features: (1)presence or absence of keywords frequently appearing in uninformative blocks (2)length of text, (3)presence or absence of verb or adjective in text, (4)text is a interior link or exterior link or the other, (5)HTML tag on DOM tree, (6)change of depth of path on DOM when current text was compared to previous text, (7)an average text length in `<table>` tag, (8)a rate of link in `<table>` tag. Furthermore keywords frequently appearing keywords in uninformative blocks are automatically selected from training data. The criterion for selecting keyword is that frequency of occurrence is high, probability that keyword appear in informative blocks is high, and a keyword appears in uninformative blocks in pages of different domains. Keywords that satisfy these conditions are selected. The number of different domain was considered not to select keyword which often appear in uninformative blocks only a certain Web site.

To examine the effectiveness of detecting method of uninformative blocks, we got 781 pages from Web directory by random sampling. Uninformative blocks were marked into Web pages by hand. Our proposed method is tested by 5-fold cross validation.

Accuracy of labeling of our proposed method for text is 0.769. On the other hand, accuracy of labeling of baseline which give label O for all text is 0.698. Thus accuracy of our proposed method is better than the baseline. Precision of detecting uninformative blocks is about 0.3 for a region unit, and is about 0.7 for a text unit. This result means that it is difficult to perfectly detect uninformative blocks, but it is possible to detect partially. It is necessary to improve our system because precision is not so good. 7% of text in informative blocks are misclassified an uninformative blocks. This means that not so much useful information may discarded in error by removing automatically detected uninformative blocks from Web pages. We also examined the effectiveness of features used for chunking. As a result, we found that effective features are (1)presence or absence of keywords frequently appearing in uninformative blocks, and (2)length of text. On the other hand, (6)change of depth of path on DOM when current text was compared to previous text is not effective.