A Machine Learning Based Approach for Separating Head from Body in Web-Tables[1]

Sung-Won Jung, Hyuk-Chul Kwon

Korean Language Processing Lab. Department of Computer Science and Engineering
Pusan National University, Busan, Korea
{swjung, hckwon}@pusan.ac.kr

Abstract. This study aims to separate the head from the data in web-tables to extract useful information. To achieve this aim, web-tables must be converted into a machine readable form, an attribute-value pair, the relation of which is similar to that of head-body. We have separated meaningful tables and decorative tables in our previous work, because web-tables are used for the purpose of knowledge structuring as well as document design, and only meaningful tables can be used to extract information. In order to extract the semantic relations existing between language contents in a meaningful table, this study separated the head from the body in meaningful tables using machine learning. We (a) established features observing the editing habit of authors and tables themselves, and (b) established a model using machine learning algorithm, C4.5 in order to separate the head from the body. We obtained 86.2% accuracy in extracting the head from the meaningful tables.

1 Introduction

Information extraction encounters various text types. Generally, editors produce three types of text: free text, structured text, and semi-structured text. Among those, free text, composed of natural language sentences, is the most frequently found. To extract information from free text, a computer must analyze the text using natural-language-processing techniques. However, practical application of natural language understanding is still far from being achieved. On the contrary, authors make structured text for specific aims such as a database or a file. These texts contain restricted and well-formed rules. Computers can easily analyze them even though they do not contain structured information apart from that which is predefined. Semi-structured text falls between structured and free text. We can include tables and charts in this type. These texts are easier to analyze and contain more useful and dense information than free text, because of their structural features. This paper focuses on the table among the semi-structured texts, because the table is usually used in HTML documents and easily extracted from HTML documents.

[1] This work was supported by the National Research Laboratory Program M10400000279-05J0000-27910 of Korea Science and Engineering Foundation.