Abstract

Most of the data extraction method on web pages using the wrapper induction and automatic data extraction. Automatic data extraction method occurs because the previous method was considered too complicated. In the process of data extraction, automatic data extraction form a pattern that will be fitted with the HTML tags on web pages..

This Final project implemented automatic data extraction method with use algorithm is called IDE (*Instance-based Data Extraction*). This technique involves the user in forming pattern by labels the web page. In the process of instance-based data extraction are three main steps, i.e, page labeling, similarity measure and data extraction.

Accuracy in shaping the pattern extraction can be done by filling the *range node* as much as nodes contained in a template of the *target item*.

IDE algorithm performance is affected by the given value of *range node*. If the node is drawn closer and closer to the *pattern target items* then his performance will improve. Moreover, type of website that is extracted also affect performance. Website which has simple *pattern target item* (the HTML structure of the data to be extracted) is easier to be extracted.

When the extracted data does not have a unique pattern, the IDE algorithm's will be difficult to extract data according to user desires. There are irrelevant data as a result of the extraction results.

Phase implementation and analysis with testing parameters such as *recall ratio* and *precision ratio* shown that built IDE algorithm is proved to obtain information in accordance with the user desires though it is noise.

Keywords: *Automatic Data Extraction, pattern, pattern target item, page labeling, similarity measure, data extraction*