Table styled information extraction from text-based web pages

Yoshimasa Sogabe (110067)

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

February 14, 2003

Keywords: Information extraction, Ontology, the WWW, Tables, Named entities.

In this thesis, we propose a information extraction method that extracts information from texts on the World Wide Web (the WWW), and that produces tables. The tables can be regarded as a concise summary of the texts, and we can grasp the major contents of the texts in a short time by looking at the tables. The conversion of texts to tables allows us to reduce the amount of the energy and the time for finding the texts or information we want from the WWW.

Nowadays, the WWW allows us to access a great amount of texts describing a wide variety of entities. The problem is how to find the texts or the information written in the texts that users want to read or know. We usually use search engines to pick up the texts that we want to read. But the search engines provide only the list of the indexes to the texts that contains the key words we specified, and we are forced to *read* many individual websites for checking if the sites contains the texts we want. This often requires much energy and long time.

One method to reduce the costs of finding the texts is to use conventional automatic summarization systems that reduce amount of the texts with preserving major contents of the texts. If we apply the systems to the texts that the search engines find, we can reduce the costs of reading the texts. However, the output of the summarization systems is still texts, and

we think reading the summary still requires considerable time and energy. We aim at reducing the costs of reading the summaries by providing the summary in the form of tables, which is more readable than normal texts.

Another type of systems that may help us to reduce the costs of the text search are information extraction systems, which are the systems to extract nothing but necessary information out of a huge amount of texts. However, existing information extractions systems require a large amount of hand-crafted extraction patterns to identify the place in the texts that the required information is written. This limits the extents of the texts and the task domains that an information extraction system can be applied to. In other words, there are no domain dependent information extraction systems, and if one wants to port a system to a new domain, time consuming task of writing down extraction patterns is required.

On the other hand, our method does not require such extraction patterns. We generate extraction patterns from the tables found in the WWW and the normal the WWW texts in an automatic and unsupervised manner. Generally, when we describe an object in table form, only information important for the object is expressed by pair of "Attribute" and its "Value" in brief forms. Assume that there is a table about self-introduction. Then, the table is likely to have attributes such as "name", "sex", "hobby" and their values are likely to be "David", "male" ,"ice hockey." Our method convert such attribute value pairs into analogues of extraction patterns and extract the inmportant information from texts by using the results of the conversion. The extracted important information is presented in the form of tables.

In order to apply our method, we need 1) to collect many tables that contain information about the same type of objects and 2) to recognize the logical structures of tables, i.e., which parts of the tables express their attributes and values. For this pre-processing, we apply an existing method [1] to recognize the logical structures of tables in the WWW and to integrate the tables describing similar objects into a single and large table, by using the recognized table structures.

Our method for information extraction uses ontologies from the integrated tables, extracts words that may serve as the values and the attributes in the tables from usual texts, and produces tables expressing the

major contents of the texts. We call the words that can be values and the attributes *table elements*. Our method can be summarized as follows. First, we apply morphological analysis to all table elements in the table as a preprocessing step and to segment the table elements into a sequence of morphemes. Next we assign tags to each table element. More precisely, the tag ⟨ATTRn⟩ is given attribute and ⟨VALn⟩ is given to value. (n : name of class) Next, we collect much text scattered over the the WWW and perform morphological analysis. And then, we assign the same tag to the table elements appeared in the texts. We regard the texts with tags as a training data and apply an existing supervised learning method to learn the appearance pattern of a table element of a table.

We assume that the supervised learning of the table element of a table in this research is a kind of tagging task, which is recognized as one of the most fundamental processing in natural language processing. We use YamCha[2] for doing our task. YamCha is a generic, customizable, and open source text tagger that has already applied to many Natural Language Processing tasks. YamCha is using a state-of-the-art machine learning algorithm called Support Vector Machines.

As features for learning by YamCha, we use "vocabulary", "reading", "standard form", "part-of-speech subdivision", "head 4 bytes of a word", and "final 4 bytes of a word" were used. Furthermore, we add a semantic word class as a feature. In the experiment, we try to apply our method two domains of texts, PC specification and self-introduction. The evaluation was done according to the following criteria.

- Special symbols such as "=" were removed from the evaluation object in evaluating.
- When a table element includes morpheme which consist only of numeric character, we removed from the evaluation object in evaluating.

The experiment results were given as follows.

- According to PC specification, the accuracy ($F_{\beta=1}measure$) is 52.6.
- In case of self-introduction, the accuracy ($F_{\beta=1}measure$) is 53.8.
- In case of self-introduction, without word class as a feature, the accuracy ($F_{\beta=1}measure$) is 52.7.

References

[1] Minoru Yoshida, Kentaro Torisawa and Jun'ichi Tsujii. (2001). A method to integrate tables of the World Wide Web. In the Proceedings of the first International Workshop on Web Document Analysis (WDA 2001). pp. 31-34.

[2] Taku Kudou. YamCha: Yet Another Multipurpose CHunk Annotator . http://cl.aist-nara.ac.jp/~taku-ku/software/yamcha/