Faculty of Science and Technology
Tokushima University

Interactive Text Mining on a Corpus

Lecturer Minoru Yoshida

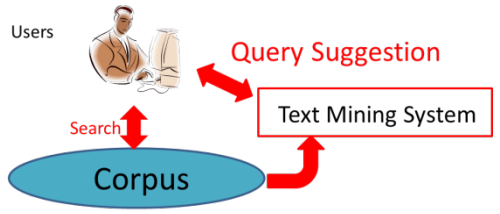

Fig.1: Workflow: Query Suggestion by Text Mining

Fig.2: Our System: Query Suggestion by Text Mining

Fig.3: Mining Numbers on Text

Content:

The amount of electronic texts is rapidly growing, making it difficult to analyze them by humans. We especially focus on ``middle data'' that is not so big but not small data. Such data includes all Wikipedia pages and operation reports in call centers, etc. We propose a ``query suggestion system by real-time text mining.'' Text mining is a task to analyze how given words are used in the given corpus (i.e., set of texts). We use the index structure called ``suffix arrays'' to provide two types of text mining results, namely, usage extraction and synonym extraction, for the given query. (See Figure 2.)

We also propose a system for mining numbers in text. Many numbers are included in text, but most of existing text mining systems treat them as mere strings of digits. We propose a system that provide a function to use ``range of numbers'' as queries. (See Figure 3.)

Keywords: Text mining, Suffix arrays

E-mail: mino@tokushima-u.ac.jp
Tel. +81-88-656-9689
Fax: +81-88-656-9689