A Study of String Matching System Based on Database Set Operation

Ming-Hsiung Ying and Chien-Yu Lin

Chung Hua University, Department of Information Management, HsinChu, Taiwan
mhying@chu.edu.tw, zw1018242000@gmail.com

Abstract. In recent years, people can easily get various data and information through Internet. People can copy the entire downloaded data, digitized information into own paper work, and form some plagiarism problems. Previous studies use of statistics, vectors matrices to compare string among documents. When someone change the location of words, and add some superfluous words or sentences between strings, it will be greatly reduced the accurate rate of matching system; Moreover, it may cause students keep plagiarism if matching system cannot find the alignments correctly. This study uses Chinese Word Segmentation and Database Set Operation as a base to construct a string matching system to solve the excessive superfluous words and order problems. Database Set Operation may be more efficient than the program with lots of words inside its memory. This study creates a prototype system, and the result of the prototype shows that the accuracy performance is performed well.

Keywords: String Matching, Database Set Operation, Chinese Word Segmentation, Matching System.

1 Introduction

In recent years, the Internet and information technology is very popular. People can easily get all kinds of data, information and knowledge through the Internet and information technology. Therefore, more and more people access to information and knowledge ways, gradually changed from traditional books, documents, etc., into the way through information technology. Many students copy the entire downloaded data, digitized information into own paper work, and form some plagiarism problems. In recent years, many schools appeared plagiarism issues. Many students' homework have copied and revised from other students’ homework, and the teachers must to spend a lot of energy to match the homework content among students.

Previous studies use of statistics, vectors matrices to compare string among documents. When someone change the location of words, and add some superfluous words or sentences into the documents, it will be greatly reduced the accurate rate of matching system.

If matching system cannot find the alignments correctly, students will keep the plagiarism behavior. To solve the problem that adding some superfluous words and change the word order in documents, the string matching system cannot find the
plagiarism words correctly. This study uses Chinese Word Segmentation and Database Set Operation as a base to construct a string matching system. This study will creates a prototype system, which will compare with methods of matching words, and T brand’s matching system, to verify its efficiency and accuracy.

In particular, this study has the following objectives:

- The past string matching method cannot find some plagiarism words in the documents. Especially when the superfluous words have added into sentence or the words order have changed. This study uses Chinese Word Segmentation and Database Set Operation as a base to construct a string matching system to solve the excessive superfluous words and order problems.
- This study will compare with methods of matching words to verify the efficiency and accuracy of prototype system.
- The results of this study provide a reference for researchers in related fields, and hope to contribute in the field of string matching.

2 Literature Review

2.1 Extensible Markup Language

Extensible Markup Language (XML) was published in 1988 by the WWW Development Association (World Wide Web Consortium, W3C). XML and HTML is similar, the HTML has less expandability and flexibility than XML, and the XML has good scalability and class sex [1]. XML has not fixed format, and it is not a single language or defined description language, it is a Meta-Language [2]. In Figure 1, the XML file contains the data and structure, and it has become the standard for electronic data storage and exchange [3].

```
▼<toplevel>
▼<CompleteSuggestion>
  <suggestion data="中華大學資管系"/>
  <num_queries int="709000"/>
</CompleteSuggestion>
</toplevel>
```

Fig. 1. Figure 1 XML standard format

2.2 Chinese Word Segmentation

Lin et al. (2010) indicated the Chinese word segmentation is very important and basic when dealing with the analysis of Chinese documents. Chen(1992) also indicated that there are great differences in the structure and semantics of Chinese and English, the analysis of Chinese sentences single word may not be the smallest unit, for each word in terms of the English sentence is individually separated by a space[4].

In Chinese documents, to analyze the meaning of the string section, it must to execute the Chinese word segmentation processing, and then analyze the content of the vocabulary and comparison, to truly understand the words in the string to be expression meaning.