Fast Search Algorithm for Sequences of Hierarchically Structured Data

Masaki Murata[1], Masao Utiyama[1], Toshiyuki Kanamaru[1,2] and Hitoshi Isahara[1]

[1] National Institute of Information and Communications Technology,
3-5 Hikaridai, Seika-cho, Soraku-gun, Kyoto, 619-0289, Japan,
{murata,mutiyama,isahara}@nict.go.jp,
WWW home page: http://www.nict.go.jp/jt/a132/members/murata/
[2] Kyoto University,
Yoshida-nihonmatsu-cho, Sakyo-ku, Kyoto, 606-8501, Japan,
kanamaru@hi.h.kyoto-u.ac.jp

Abstract. We developed an algorithm for quickly searching sequences of hierarchically structured data, such as tagged corpora where each word includes information on its part of speech (POS) and minor POS and the word itself. Using our method, we first make a data item where each data item in a lower level is surrounded by two data items in a higher level. We then connect these data items to make a long string and store the string in a database. We use suffix arrays to query the database. Our experiments showed that our method was 194 times faster than a conventional method at fastest and 24 times faster on average. Our method can be used for other kinds of hierarchically structured data, such as Web applications. Methods that can be used on such data are in high demand. For example, our method can be used to retrieve Web text that includes hierarchical information of low, middle, and high semantic levels. If we use our method for such Web text, we can query using the terms, "High semantic level: method", "Word: in", and "Low semantic level: group"; in other words, our retrieval method is more useful and convenient than conventional Web retrieval.

1 Introduction

We developed an algorithm for quickly searching sequences of hierarchically structured data.[3] For example, in the Kyoto Text Corpus [1], each word has information on its part of speech (POS) and minor POS and the word itself. (A minor POS is a more specific POS.) The POS, minor POS, and word can be considered data of the highest layer, the second-highest layer, and the lowest layer. Hierarchically structured data, as explained below, has data from the lowest to the highest layers. The Kyoto Text Corpus has such a structure, so it can be considered to be sequences of hierarchically structured data. The algorithm we propose is for quickly searching sequences of such data. Our algorithm

[3] We obtained a Japanese patent for the algorithm.

© A. Gelbukh (Editor)
Advances in Natural Language Proccesing
Research in Computing Science 18, 2006, pp. 163-174