

# An Efficient Pathway Search Using an Indexing Scheme for RDF

Akiyoshi Matono<sup>1</sup>

akiyo-ma@is.aist-nara.ac.jp

Toshiyuki Amagasa<sup>1</sup>

amagasa@is.aist-nara.ac.jp

Masatoshi Yoshikawa<sup>2</sup>

yosikawa@itc.nagoya-u.ac.jp

Shunsuke Uemura<sup>1</sup>

uemura@is.aist-nara.ac.jp

<sup>1</sup> Graduate School of Information Science, Nara Institute of Science and Technology, 8916-5 Takayama-cho, Ikoma-shi, Nara 630-0192, Japan

<sup>2</sup> Information Technology Center, Nagoya University, Furo-cho, Chikusa-ku, Nagoya-shi 464-8601, Japan

**Keywords:** pathway, RDF, suffix array, index, directed graph

## 1 Introduction

Most chemical reaction mechanisms in living organism are translated from a compound (substrate) to a compound (product) by an enzyme acting as the initiator. Such a series of reactions is generally called a pathway. It is important to compare and analyze pathways in order to understand the process of creating compounds and the evolutive relevance between organisms. The pathway data needs to be efficiently processed because the amount of known pathways is increasing due to the development of the technology for genome data. Some of pathway databases are constructed to manage the pathway data (e.g. KEGG [2]).

However, as far as we know, all of the pathway databases are constructed on relational databases or XML databases, and thus the storage model is not based on the structure of pathway data. The structure of pathway data cannot be easily represented by flat tables or trees; we need a directed graph in which compounds are vertexes and enzymes are arcs.

In this paper, we propose a scheme to search pathway data efficiently. The basic idea is that we represent the pathway data by RDF (Resource Description Framework) [3], and apply an indexing scheme for RDF data [1]. We believe that RDF is adequate for representing the pathway data because both of them are modeled as directed graphs. The indexing scheme [1] makes it possible to search RDF data efficiently, which can be done by suffix arrays for RDF data, which is a modified version of a well-known indexing structure for full-text searching. It is possible to search the pathway data efficiently by applying the scheme to the pathway data.

In fact, RDF is a specification of the framework to describe and manipulate metadata. Additionally, RDF is important to realizing Semantic Web as a next-generation web. For this reason, researches and systems for RDF will be expected to increase. By representing pathway data using RDF, we can thus apply the researches and systems to pathway data.

## 2 Our Approach

The goal of this paper is to achieve an efficient search of pathway data. We first transform pathway data in XML format, such as the KEGG, to RDF using XSLT. Figure 1 shows a pathway data example represented by an RDF graph. Then, we apply a suffix array for RDF to the pathway data transformed to RDF.

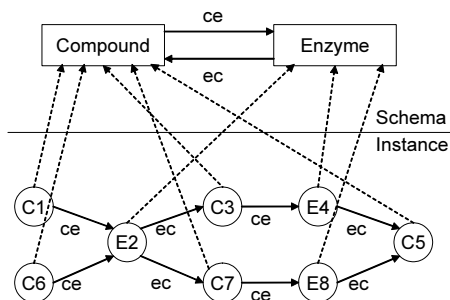


Figure 1: An RDF graph of pathway.

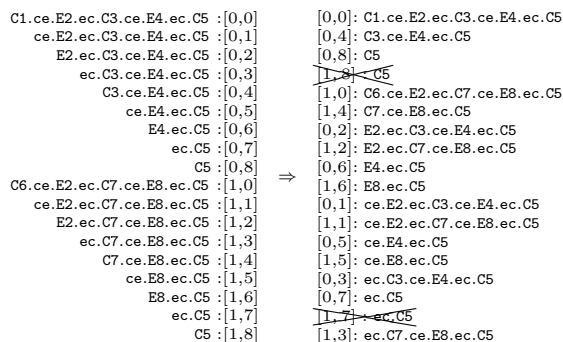


Figure 2: Creation of suffix array.

The creation of the suffix array for RDF can be summarized as follows: 1) We extract four types of subgraph in which arcs represent the distinct relationship from the RDF graph. 2) We extract all path expressions from the subgraph and assign an identifier to each path expression. In the case of Figure 1, the extracted path expressions are C1.ce.E2.ec.C3.ce.E4.ec.C5 and C6.ce.E2.ec.C7.ce.E8.ec.C5 3) We then create all suffixes from each path expression and assign an identifier to each suffix (left side of Figure 2). We define that a pair of path identifier and suffix identifier as an indexing point. 4) We finally sort all suffixes in lexicographical order and remove the redundancy (right side of Figure 2). The list of suffix indexing points are the suffix arrays for RDF. The suffix array in Figure 2 is [0,0] [0,4] [0,8] [1,0] [1,4] [0,2] [1,2] [0,6] [1,6] [0,1] [1,1] [0,5] [1,5] [0,3] [0,7] [1,3].

We think that the case in which a particular compound or enzyme is used as the query key is general in the search for pathway data. Our scheme can be used more efficiently in the case of a compound's reaction process as query path expressions than in the case of using compounds or enzymes as query keys.

### 3 Conclusions

In this paper, we focused on the structure of RDF as a directed graph that is transformed from XML pathway data to RDF, and applied an indexing scheme for RDF data to a search of pathway data. As a result, we achieved an efficient search of pathway data.

In the future, we will implement the proposed scheme and evaluate its performance. Additionally, we plan to consider updates of the indexing data.

### Acknowledgments

Akiyoshi Matono's work was in part supported by a Grant-in-Aid for the 21st COE Research Program from the Ministry of Education, Culture, Sports, Science and Technology of Japan.

### References

- [1] Matono, A., Amagasa, T., Yoshikawa, M., and Uemura, S., An indexing scheme for RDF and RDF schema based on suffix arrays, *First International Workshop on Semantic Web and Databases (SWDB), co-located with 29th International Conference on Very Large Data Bases (VLDB2003), Berlin, Germany, First International Workshop on Semantic Web and Databases (SWDB), Berlin, Germany*, September 7-8, 2003.
- [2] Kyoto Encyclopedia of Genes and Genomes: <http://www.genome.ad.jp/kegg/>.
- [3] World Wide Web Consortium. Resource Description Framework (RDF) Model and Syntax Specification. <http://www.w3.org/TR/1999/REC-rdf-syntax-19990222/>, W3C Recommendation 22 February 1999.