Title	データマイニング技術を用いたゲノムデータベースの要約手法に関する研究
Author(s)	土橋，潤也
Citation	
Issue Date	2002-03
Type	Thesis or Dissertation
Text version	author
URL	http://hdl.handle.net/10119/353
Rights	
Description	Supervisor:佐藤 賢二，知識科学研究科，修士

Research on the Summarization of Entries in Genome Database using Data Mining Technology

Junya Dobashi

School of Knowledge Science,
Japan Advanced Institute of Science and Technology
March 2002

Keywords: summarization, data mining, association rule discovery, grouping.

Research background and purpose

The University of Kyoto Chemistry Research Institute and The University of Tokyo Medical Science Institute Human Genome Center had been developing an information infrastructure named GenomeNet in order to answer for explosive increasing of data about life. The database services of GenomeNet are aiming at unifying the various knowledge, information and data of biology and medical science on the desktop of each researcher. [1]

There are many kinds of database services in GenomeNet. For example, keyword search, homology search, motif search, and pathway search are representative ones. Though they perform different processing, they are united in returning a set of documents (in genome databases, they are called *entries*) as a result. So, in case that a user obtained a long list of entries, it is not so easy to understand the meaning of it. Abstractly speaking, it is needed to know ``how a set of entries can be summarized''.

The Approach

Techniques for summarization have been studied in the research field of natural language processing. However, they could not be directly apply to entries in genome databases since they include various kinds of information besides texts written in natural language (e.g. numerical data, sequence data, structure data, and so on). In this study, we focused on a huge collection of technical terms extracted from genome databases in GenomeNet[2]. Since this collection includes millions of technical terms with their occurrences in databases, it can be

used for characterizing each entry. Then, it is needed to establish the way to know which terms are enough significant to remain as a part of summary of entries. For this purpose, association discovery, which is one of the most popular techniques in data mining, could be available. Using the technique, we can obtain only the technical terms (items) common and specific to the entries in which a user is interested.

Construction of a prototype system

A prototype system for summarizing a given set of entries was developed based on the approach above. In addition to a set of entries, the system receives some optional parameters for specifying database name, field name, maximum item combination in associarion rule, and order of priority of significance measures for sorting items. After the summarization, the system returns the following two kinds of information as summaries of specified entries.

- A list of significant items sorted by the order of specified measures
- A table of binary relation between items and entries (truth table)

By checking these summaries, a user can grasp the meaning of a set of entries.

Evaluation

Evaluation of the effectiveness of summarization were performed by the experiments using three kinds of sets of entries, that is, result of full-text search, result of homology search, and artificially collected subset of ENZYME database. Through these experiments, the following things were found.

- In most cases, the system tends to pick up significant items common and specific to a given set of entries.
- If a given set of entries contains many noises, e.g. unwished entries in the result of full-text search, the truth table might not so informative. In other words, it might not include meaningful groups in the set of entries. On the other hand, in case of homology search result and subset of ENZYME, clear groups or subgroups were appeared in truth table. It suggests that the system is useful for summarization and knowledge discovery from a set of entries in which a user is interested.

References

[1] Database directions of a genome network [the 2nd edition]:Toshihisa Takagi, Minoru Kanehisa, Kyoritsu publisher.

[2] Takuya Yagyuu and Kenji Satou: Toward Automatic Construction of Extensional Ontology from Genome Databases, Genome Informatics 2000, pp.442-443 (2000).