Title	ゲノムデータベースにおける柔軟なデータ加工およびマイニングシステムの構築に関する研究
Author(s)	布施田，敏樹
Citation	
Issue Date	2000-03
Type	Thesis or Dissertation
Text version	author
URL	http://hdl.handle.net/10119/631
Rights	
Description	Supervisor:佐藤 賢二，知識科学研究科，修士

A Study on a Data Mining System for Genome Databases with Flexible Data Tailoring and Cleaning Function

Yoshiki Fuseda

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

March 2000

Keywords: Genome Database, Data Mining, Association Rule, Genome Informatics ,Knowledge Discovery .

Abstract

Today, major sites of genome database services, including GenomeNet in Japan, are gathering and updating over ten millions of heterogeneous database entries. Furthermore, most of them equip retrieval functions and analysis tools for processing such data. This situation is similar to some sort of data warehouses since heterogeneous and large amount of data are gathered together for trans-database search and analysis.

In this study, a technology called "data mining", which is one of the most actively studied in the area of computer science, was applied to the analysis and knowledge discovery from large amount of various genome databases. In genome informatics, data mining against DNA, protein, and disease data is highly noticed as a promising approach to scientific discovery. Besides this study, there are several application studies which uses some kind of data mining algorithms for finding useful knowledge from genome databases. However, since most of data mining algorithms are combinatorial (it means high computational complexity), in typical cases, they use only small amount of static subset of genome database. There is one more reason of it, that is, data mining technology still remains the area of computer science (it means biologists could not utilize data mining tools without aid of computer scientist or skillful programmer). So, it is needed to develop a system which can treat large amount of genome databases and provides analysis and discovery services via WWW, but there is no such services or systems.

On the contrary, this study addresses to the application of a data mining algorithm against large amount of various genome databases maintained on GenomeNet. Using association discovery algorithm and dynamic data tailoring functions, the system developed in this study provides the ability of rapid and flexible knowledge discovery from genome databases based on the cross-reference information among them. The information, called LinkDB, is one of fundamental data in GenomeNet, and maintained by Kyoto University, University of Tokyo, and JAIST. By using it, I introduced a new framework called "data mining from cross-reference information" into the application area of knowledge discovery from scientific databases.

A user of this system can:

1) easily use it via WWW from his/her browser,

2) perform analysis and knowledge discovery from the latest genome databases in GenomeNet, and

3) rapidly obtain only the association rules related to the set of database entries which he/she gave to the system to express his/her interests.

By the last feature above, the system dynamically cuts down irrelevant cross-reference information and execute association discovery algorithm against very small but important information. It enabled fast data mining from huge amount of cross-reference information (10^{17} in the worst case) about heterogeneous databases.

Furthermore, a new problem of data mining arose, which was not imagined in the area of data mining from usual business data. It can be regarded as a problem of redundancy, that is, in the situation that many items have the same bit pattern, association rule finding ruins by the explosion of computation for meaningless rules, where the word "item" means some proposition on the genome data, and "bit" means the proposition holds. Such a situation rarely considered for managing business data, however, by extracting small subset of data related to a user's interests, the system should overcome the problem of redundancy above. To solve the problem, a new algorithm called "item-chunking" was introduced to the system. It was verified that the algorithm drastically cuts down meaningless association rules consist of redundant items.

As the future work of this study, some improvements are planned about the system developed.

Related research

Yoshiki Fuseda and Kenji Satou: Toward a Data Mining Service from Large and Heterogeneous Genome Databases in GenomeNet, Genome Informatics 1999, UNIVERSAL ACADEMY PRESS,INC. TOKYO, JAPAN.