Toward a Data Mining Service from Large and Heterogeneous Genome Databases in GenomeNet

Yoshiki Fuseda **Kenji Satou**
yosif@jaist.ac.jp ken@jaist.ac.jp
School of Knowledge Science, Japan Advanced Institute of Science and Technology,
1-1 Asahidai, Tatsunokuchi, Ishikawa 923-1292, Japan

1 Introduction

Today, major sites of genome database services 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.

The GenomeNet is one of such site, which contains about 20 sorts of genome databases (GenBank, EMBL, SWISS-PROT, PDB, KEGG, etc.) and provides many services including entry retrieval and sequence interpretation. However, there are few sites which provides data mining services covering all the database entries stored in them. The main difficulty here is data preparation for mining. As pointed out recent studies in computer science, it is hard work and requires deep knowledge about mining algorithm and structure of data itself. To solve this problem, we are developing a system for assisting flexible preparation of data for mining from the whole GenomeNet (Fig. 1).

2 System Design

Recently, data preparation for mining are roughly classified into some operations called feature extraction, feature selection, and so on [4]. From this point of view, we think that data preparation for mining from GenomeNet includes at least the following levels of operations. As to the algorithm of data mining, association rule discovery [1, 2, 3] is currently focused on.

1). "entry-entry relationship" It means hard links among entries like cross-reference information. In case of GenomeNet, it can be easily extracted from the data stored in LinkDB.

2). "content-content relationship" Automatically extractable contents of heterogeneous entries (e.g. classification, keyword, etc.) can be related to each other by using entry-entry relationship.

3). "relationship among synthetic data" The two relationships above are static and able to be prepared beforehand with user's request. We have to provide a way to incorporate dynamic data into the data set for mining, which might be calculated via the analysis tools in GenomeNet.

4). "relationship filtering" The relationships in 1), 2), and 3) should be filtered in accordance with user's interests since most of the users want to perform data mining on small subset of data in short time. Database selection and simple pattern matching might be used in this filtering.

Acknowledgments

This work was supported in part by Grant-in-Aid for Scientific Research on Priority Areas, "Genome Science", from the Ministry of Education, Science, Sports and Culture of Japan.

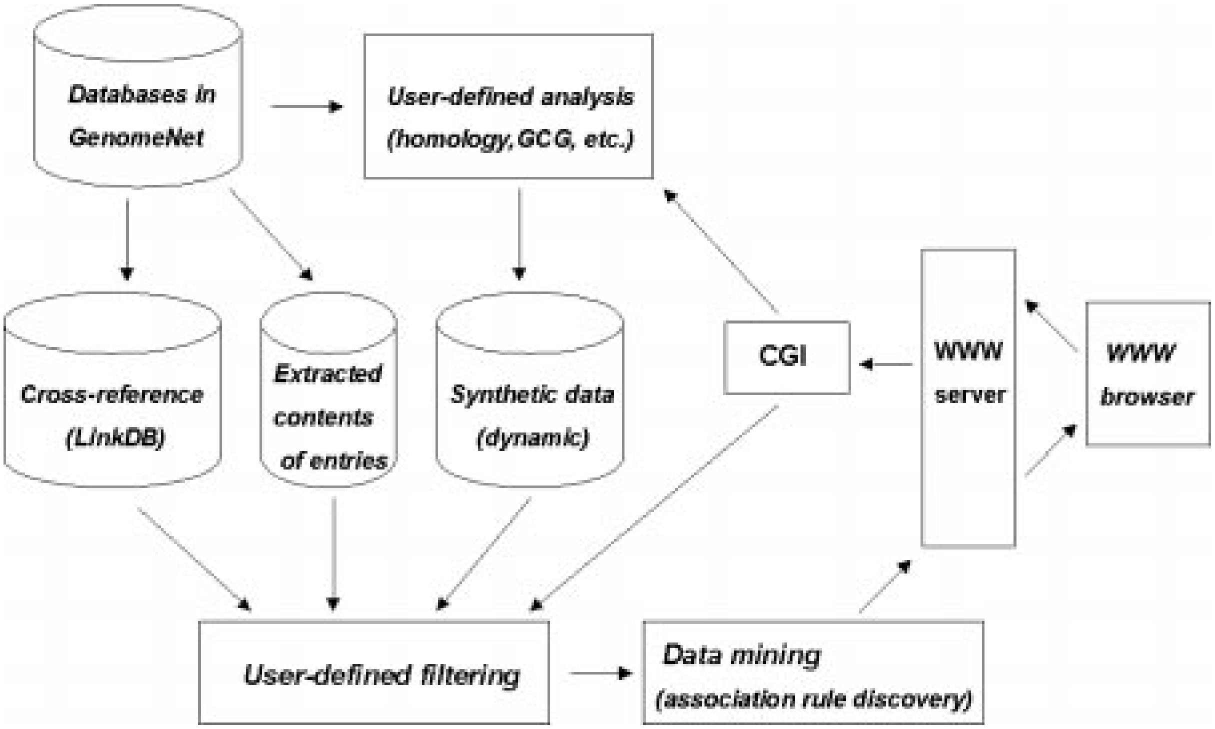

Figure 1: Overall Architecture.

References

[1] Agrawal, R., Imielinski, T., and Swami, A., Mining association rules between sets of items in large databases, *Proc. of ACM SIGMOD*, 207–216, 1993.

[2] Agrawal, R. and Srikant, R., Fast algorithms for mining association rules, *Proc. of VLDB*, 487–499, 1994.

[3] Kitsuregawa, M., Mining algorithms for association rules, *Journal of Japanese Society for Artificial Intelligence*, 513–520, 1997.

[4] Liu, H. and Motoda, H. (eds.), *Feature Extraction, Construction and Selection: A Data Mining Perspective*, Kluwer Academic Publishers, 1998.