Chapter 2
A Multiagent-Based Framework for Integrating Biological Data

Faheema Maghrabi
Ain Shams University, Egypt

Hossam M. Faheem
Ain Shams University, Egypt

Taysir Soliman
Asuit University, Egypt

Zaki Taha Fayed
Ain Shams University, Egypt

ABSTRACT

Biological data has been rapidly increasing in volume in different Web data sources. To query multiple data sources manually on the internet is time consuming for biologists. Therefore, systems and tools that facilitate searching multiple biological data sources are needed. Traditional approaches to build distributed or federated systems do not scale well to the large, diverse, and the growing number of biological data sources. Internet search engines allow users to search through large numbers of data sources, but provide very limited capabilities for locating, combining, processing, and organizing information. A promising approach to this problem is to provide access to the large number of biological data sources through a multiagent-based framework where a set of agents can cooperate with each other to retrieve relevant information from different biological Web databases. The proposed system uses a mediator-based integration approach with domain ontology, which uses as a global schema. In this paper we propose a multiagent-based framework that responds to biological queries according to its biological domain ontology.
INTRODUCTION

Recent advances in laboratory technology have resulted in massive amounts of biological data that are often deposited in Web databases. Clearly, access to this data is very important to biological researchers. However, heterogeneity among biological databases due to the incompatibilities in data formats, data representations, and data source schemas has impeded the accessibility to such databases. Three fundamental approaches have been used to address the challenges associated with the incompatibilities among biological databases: (a) data warehouse integration, (b) information linkage integration, and (c) mediator-based integration.

Data warehousing consists of materializing the data from multiple data sources into a local warehouse and executing all queries on the data contained in the warehouse rather than in the actual sources. Data warehousing suffers from a lack of scalability, when considering the exponential growth of biological databases. The information linkage integration, motivated by the fact that many of data sources on the Web are browsed instead of queried, the integration happens through links and applies to any collection of data sources which can be seen as a set of pages with their interconnections and specific entry point. Information linkage takes advantage of distributed resources. However, maintaining and updating the static links between the various databases is a challenge. Furthermore, the only queries that can be answered by information linkage based systems are those that are within the scope of the pre-existing static links (Hernandez & Kambhampati, 2004; Miled, Webster, & Liu, 2003).

The proposed framework is based on the third approach, namely, mediator-based integration. It establishes a transparent access to heterogeneous data sources without physically copying them into a single data repository. Database integration systems which use mediator-based integration consisting of three elements: wrappers, an integration layer, and a query interface. The wrappers provide uniform access to the heterogeneous data sources. The integration layer decomposes user queries, sends them to the relevant wrappers, and finally integrates the query results before the final result is returned to the user via the query interface (Köhler, Philippi, & Lange, 2003).

This class of integration systems can be divided into two subclasses: local as view (LAV) and global as view (GAV). In LAV, there is no global schema, and the user needs to specify the component databases in the query by using a multi-database query language. One of the disadvantages of LAV is that the component databases are not transparent to the user. This approach is used by BioKleisli (Davidson, Overton, Tannen, & Wong, 1997). In GAV, a global data schema is constructed, and the queries are expressed on this unified schema. The global schema integrates all the component schemas. Therefore, the component databases are transparent to the user. TAMBIS (Baker, Brass, Bechhoferb, et al., 1998) and BACIIS (Miled, Li, & Bukhres, 2005) are examples of GAV, where they use ontology as a conceptual model to integrate multiple biological Web databases.

Domain ontology and mapping schema are the two main components of the proposed system. The domain ontology can be used to define a common controlled vocabulary and to semantically define databases. It is designed with hierarchal structure describing the biological concepts and the relationship between them. When integrating heterogeneous databases, some issues have to be addressed, namely, semantic and syntactic variability. Resolving semantic variability consists of adequately interpreting and cross relating information stored in different databases. Syntactic variability arises from the heterogeneity of the database schema, data models, and query processing approaches. Domain ontology has been used successfully as a domain model in several general-purpose information integration systems. The domain ontology in this architecture serves as