Analysis and Design of Bio-Info-Collaborator Service Model to Collaborate Multiple Heterogeneous Bioinformatics Web Services

Isha Soni
M.Tech,
BRCM CE&T,
Bahal, Bhiwani

Amit Nain
Asstt. Professor
IT Dept. BRCM CE&T,
Bahal, Bhiwani

ABSTRACT

In this paper, an approach to generate a mechanism for bioinformatics web services to register them with a Data Collaborator Web Service has been presented. The aim of this approach is to create new collaborator service model using existing web services. Up to now, the focus has been on answering specific user queries. On the contrary, the plan is to focus on an automated mechanism that can work as a data collaborator web service. This mechanism would be able to respond to a range of inputs to provide desired and efficient output that are required by the service user.

General Terms

Existence of several large bioinformatics databases on Global Internet resulted the need for an approach to collaborate information from multiple data sources and services. In recent years, technology has moved so fast to develop major advances to access data on the web. There was a time when data access was limited to text files and that was accessible to a limited number of people. Now technology gave us so much huge and powerful resources like online data systems capable to store several terabytes of data.

Keywords

Bioinformatics, Web Services, XML Web Services, Collaborator Web Services, Heterogeneous Databases.

1. INTRODUCTION

Bioinformatics is unified discipline formed from combination of biology, information technology and computer science. It basically deals with biological information like Data collection, Data Storage, Data Searching and Retrieval. In general, if we notice the real scenario, a large number of bioinformatic datasets are available online which provides on the fly data in various formats. Each of the data format has its own pros and cons. It is hard to change the existing services because of their heterogeneous structure & size. This problem led the need to find out a common solution or a common approach to collaborate information from these different datasets. Only a favorable pro is that, unlike other domains, the bioinformatics uses web standards such as XML and web services for its enhancement.

A Web service is a program that can be executed on a remote machine using standard protocols, such as WSDL and SOAP. There exists a large number of bioinformatics data sources that are either accessible as web services or provide data using XML [13]. Most of the available bioinformatics web services are information providing services. These services do not change the state of the world in any way. For example, if a user queries the web service for details of a protein, the user provides a Protein Name and gets back the information about the protein; the presence of the large number of information providing services has highlighted the need for a framework to integrate information from the available data sources and services.

2. BIOINFORMATIC DATA

2.1 Management of Bioinformatics Data

Today a bioinformatics information system typically deals with large data sets reaching a total volume about one terabyte. The problem of managing this information is not solved satisfactorily, although it has been recognized as a key component of today's genome/biology research. [5]

2.2 Present Web Service Scenario

In present scenario, If a user need some information for a particular virus, protein or any other data component, he need to go through various resources to collect the information and there after he again need to work (manually or using any other tool) to consolidate the fetched data set.

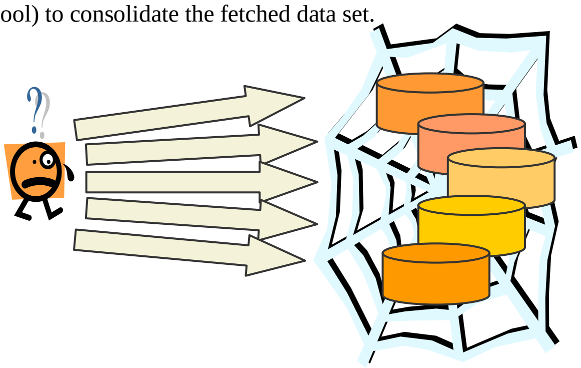


Fig 1: End User's Request to Multiple Services

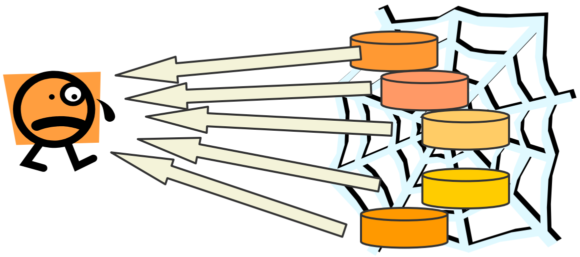


Fig 2: Web Service Response back to User.

2.3 Challenges/Issues in Data Management

- Bioinformatic datasets are complex to be modeled. Various relationships make it even harder to use any existing model for each bioinformatics data structure. [14]
- Because of regular researches, new type of data is being emerged on a regular basis. For instance complete sequences of genomes, micro arrays, interaction maps of proteins, proteomics. Biggest issue with these data is that each molecule emerges with different property. So the problem is not only to model this new type of data, but to modify the perception of old data model.[1]
- The volume of data grows exponentially, doubling in less than two years [14]
- Data are updated very frequently, access intensively and exchanged very often by researchers on the internet.
- Data analysis generates new data that also have to be modeled and integrated.[6]
- Always there is a need to access the existing information because scientists regularly need to return to raw data to confirm computer interpreted results.[12]
- Granularity of data is rather fine, and the terabyte of bioinformatics data consists of a large number of objects.
- Complex queries can be issued by different users..[6]
- Data is stored in a web of different databases that are duplicated in several repositories. There is a need to discard duplication information.[9]
- Most of bioinformatics databases have heterogeneous formats.

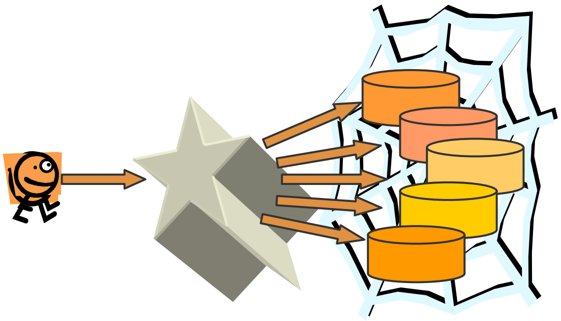
3. EXISTING ADVANCEMENTS

For the problems described above, various researchers put their efforts to draw or develop some sort of mechanism that could help a user to retrieve the information efficiently. But most of the work revolved around the data modeling techniques [4], query processing, speed & performance of data retrieval [3], Efficient Query plans, Data Integration Techniques [11][12] & Semantic web techniques [14]. Although most of researchers got successful what they want to achieve, but still no one is still able to provide an efficient solution for all of the problems mentioned above. Most importantly handling of heterogeneous structure datasets was itself a big challenge. Some of the researchers provided mediator approach and got succeed up to a certain level. But the proposed models could prove themselves when there was huge change in source data structures[1]. Handling unknown structure of source data models was itself a big problem, but frequent structure updates made this a huge challenge to be resolved efficiently. Almost all of the mediator solution got failed when there was a big change in source data structures.

4. PROPOSAL: BIO-INFO COLLABORATOR SERVICE

During the preliminary study of existing solutions, it was discovered that none of existing mechanism is capable to resolve all of the above mentioned problems. This led the research work to move the brain towards a totally different thought. Rather than updating existing data models, the complete effort was put and concentrated towards developing a mechanism that would convert existing heterogeneous data models to the famous relational data model. As a solution of the problems described above, this new mechanism is proposed (i.e. Bio-Info-collaborator model). As a solution, a mediator collaborator service model was developed. This model is capable to register and collaborate the results provided by other heterogeneous bioinformatics web services and finally provides the consolidated results in formatted and user friendly structure (e.g relational databases). To achieve this there are some pre-initialization steps that are required to take care before starting any data providing services. This proposed mechanism (Collaborator web service model) needs some kind of information (Metadata in the form of XML) from the other existing web services while registering them. This information (XML) set contains the meta data for the database structure as well as the data modeling information (For example: Database structures, database Schema, Column details, Size, DataType, Primary/Candidate keys etc.). When this information is submitted by the requesting web service, this Collaborator service is able to know whether which service contains which information. Once the web service is registered with this system, it starts storing all the information structure into its own database (not the actual data, but the schema information only). So that if the End User tries to issue a Query with this Collaborator web service and searches for some information, this collaborator service requests to other registered heterogeneous web services and obtains the desired data. Now it is responsibility of this service to convert all the heterogeneous data to homogeneous form. As it already know the structure / schema/ modeling technique of source data, the service can easily convert the data to relational form.

.



This way the end user will get all the consolidated & collaborated information from a single source rather than fetching the data from multiple sources. Given below is the comparison table that separates this model with existing models:

Operation Performed by Mediator Services	Bio-Info-Collaborator	Other Mediator Services
Able to handle Complex Data Structure	YES	YES
Handles Large Volumes of Data	YES	YES
Handles Frequently Changing Information	YES	YES
Provides Consistent modeled Result Set	YES	YES
Handles Redundancy of Data	YES	YES
Handling Complex Search Queries	YES	YES
Handles Source Data Structure Updates	YES	NO

Comparison between Bio-Info-Collaborator & Other Services

For testing authenticity and efficiency of this collaborator service model, an in-house application was developed to implement the architecture of collaborator service. The testing was done by developing and implementing three different heterogeneous web-services providing information for viruses with different data structure and then thereafter collaborating those service results through a sample application service based on this bio-info-collaborator model. The testing was done in four different phases. During each phase, data structure modifications were done in result sets given by source services and Collaborated result was reviewed thereafter. Finally after getting all positive results, authenticity of the application was proved.

5. FUTURE SCOPE AND APPLICATION

One interesting future application that can be visualized for this work is an advanced, intelligent Internet search mechanism for Bioinformatics, a simple program that would collect the results of a search from each of the major Bioinformatics Web Services and represent those results as logic programs. With this system, complex queries could be performed against the combined knowledge base to select precisely the items of interest, or weed out the non-interesting results. An area for further research is the possibility of constructing distributed simulations, using the strength of the logic-programming paradigm in abstract modeling. The vision is to construct a logical model of the system being modeled, where each component in the system is represented by a logical term. Possible solutions for each term are drawn from a distributed knowledge base that can be linked together in real time to solve the overall system. With an intermediary similar to that mentioned above for the search engine interface, results from numerical calculations or other computations could be combined and included in this logical model.

The focus is on automatically generating special Mechanism that can be hosted as web service that respond to a range of inputs to provide desired and efficient output that are required by the service user.

6. CONCLUSION

During the research work, various technologies like HTML, XML, ASP.NET, and ADO.NET were reviewed for their role in the field of biology. The finding was that there is a need to improve existing mediator service models because the data is very large and being updated frequently. So a need was felt to develop frequent web services according to data that is being updated regularly. Although there are many development technologies that can help to enhance existing web services but with the help of XML and .Net framework development tools web services can be enhanced more frequently and effectively.

At last, a registrar based approach was introduced to automatically generate integration mechanism that can be hosted as a web service. And finally it was proved that these existing data integration techniques could be extended to generate integration mechanism for new collaborator web services.

7. REFERENCES

[1] William M. Shui - Utilizing Multiple Bioinformatic Information Sources: An XML Database Approach , Sydney, NSW 2052, Australia

[2] Sandeepan Banerjee - A Database Platform for Bioinformatics - Redwood Shores, CA, USA

[3] Stephen W. Ryan and Arvind K. Bansal Applying Java for the Retrieval of Multimedia Knowledge Distributed on High Performance Clusters on the Internet Proceedings of the International Conference on Practical Applications of JAVA, London, UK, (1999), 1993 – 2003

[4] Zoe Lacroix, Terence Critchlow -Bioinformatics – Managing Scientific Data - 2003

[5] Eckman, B.A., lacroix Z. Raschid L: Optimized seamless integration of biomelocular data. In: Proceedings of 2nd IEEE International Symposium on Bioinformatics and Bioengineering (2001)

[6] Ullman, J. : Principles of Data and Knowledge-Base Systems. Computer Sciences Press, New York(1988).

[7] Bultan, T. Fu, X. Hull, R. Su, J. : Conversation specification: a new approach to design and analysis of e-service composition. In: Proceedings of 12th International World Wide Web Conferences(www) – 2003

[8] Kei-Hoi Cheung, Kevin Y. Yip, Andrew Smith, Remko deKnikker, Andy Masiar, Mark Gerstein, -"YeastHub: - A semantic web use case for integrating data in the life sciences domain"

[9] Duncan Hull, Robert Stevens, and Phillip Lord "Describing Web Services for user-oriented retrieval"

[10] Stuart E. Madnick –"The Misguided Silver Bullet: What XML will and will NOT do to help Information integration"

[11] Marie-Dominique Devignes and Malika Smaïl – "Integration of Biological Data From Web Resources : Management of Multiple Answers Through Metadata Retrieval" UMR LORIA 7503, CNRS-University Henri Poincaré, BP 239, 54506 Vandoeuvre-lès-Nancy, France.

[12] Shirley Crompton, Brian Matthews, Alex Gray, Andrew Jones, Richard White- "Data Integration in Bioinformatics Using OGSA-DAI"

[13] Remko de Knikker1, Youjun Guo1, Jin-long Li1, Albert KH Kwan2, Kevin Y Yip2, David W Cheung2 and Kei-Hoi Cheung*1,3 – "A web services choreography scenario for interoperating bioinformatics applications"

[14] David Buttler, Matthew Coleman, Terence Critchlow, Renato Fileto, Wei Han, Ling Liu, Calton Pu, Daniel Rocco, Li Xiong – "Querying Multiple Bioinformatics Information Sources: Can Semantic Web Research Help?"

[15] Tony Hey and Anne Trefethen- "The Data Deluge: An e-Science Perspective"