Ph.D. in Information Technology:

Canakoglu and Venco Final Dissertations

DEIB Seminar Room

January 15th, 2016

10.00 am

First Ph.D. presentation and discussion:

Arif CANAKOGLU – XXVII Cycle

"Modular Schema Based Data Warehousing of Evolving, Redundant and Incomplete Data: Application to Biomolecular Knowledge Data Integration and Inference"

Supervisor: Prof. **Marco Masseroli**

Abstract:

Heterogeneous data integration is an important persistent problem that has been often faced in different domains. It is highly challenging when the heterogeneous data are very numerous, fast evolving, from different and distributed sources, and need to be efficiently and comprehensively evaluated in order to answer complex queries, possibly in short time. Data warehousing well supports applications where off-line processing of numerous data from various and dispersed sources is required, e.g. in order to comprehensively and efficiently mine the integrated data towards knowledge discovery. Yet, they leave open issues to solve, firstly, when the many heterogeneous data sources to be integrated are evolving in number, in type and also in their data schema (limitedly and less rapidly). Another difficulty is that the information from the sources may be complementary, but also overlapping. All these aspects require performing the integration by means of a well-defined but simple methodology, which is easily configurable and rapidly adjustable, in order to be able to cope with the challenges of these sources. All these difficulties and requirements are typical in bioinformatics. Such complex scenario, which is not present in other less challenging domains, led us to choose the bioinformatics domain to test and demonstrate the efficacy and effectiveness of our research, which aims at developing a domain independent abstracted and generalized data warehousing approach that can be straightforwardly customized and applied in different domains.

In this Thesis, the above issues and challenges are addressed by focusing on the integration of controlled annotation data, expressed through several terminologies and ontologies, from different sources, which represent the various aspects of the available knowledge. In order to meet these requirements, in our work the following steps were performed: 1) abstraction and generalization of the main features represented by the data to be integrated (e.g. biomolecular entities and biomedical-molecular features) and their

associations; 2) design of a modular global data schema according to such abstraction and generalization; 3) design of a multi-level data architecture (including import, aggregation and integration levels) and the metadata that describe it and the imported data integrated; 4) design and implementation of provenance recording and consistency checking of imported data, in order to ensure high quality of integrated data; and 5) development of a software framework for the automatic creation of a data warehouse implementing the designed global data schema, architecture, provenance tracking and quality checking. Furthermore, by leveraging association data integrated with the proposed approach, using different techniques new data associations can be inferred, which may help scientists to orientate their research and experiments.

After these generation of the data warehouse, to ease accessing, querying and extracting the many valuable data integrated in the created data warehouse, several different interfaces were developed. The data warehouse is publicly accessible through a basic and a more advanced web interface at http://www.bioinformatics.deib.polimi.it/GPKB/; furthermore, we created a Web Service interface to the data warehouse (http://www.bioinformatics.deib.polimi.it/GPKB-REST-client/). Towards these goals, the web service access to the data warehouse within different projects were leveraged, including the Bio-Search Computing (http://www.bioinformatics.deib.polimi.it/bio-seco/seco/) and Drug Repurposing ones. In the latter one, performed a mediator based integration of the web service has been designed to access our data warehouse and the web service to access the Semantic MEDLINE Database (SemMedDB, http://skr3.nlm.nih.gov/SemMedDB/) of the US National Library of Medicine. SemMedDB is a repository of semantic predications that are extracted from the very numerous publication abstracts included in PubMed (http://www.pubmed.org/) by using natural language processing methods. The integration of the two resources allowed supporting queries aimed at drug repurposing, potential discoveries or hypothesis formulation by using different concepts, in other words a drug may be determined to treat a disease other than the one for which it was intended.

Second Ph.D. presentation and discussion:

Francesco VENCO - XXVIII Cycle

“Modeling and Querying Genomic Data”

Supervisor: Prof. **Stefano Ceri**

Abstract:

In the last decades we witnessed the birth and growth of the new field of genomics, thanks to the recent opportunities given by high throughput DNA sequencing. In recent years, the quantity and quality of the

data produced have augmented, while the cost of sequencing is dropping; experts foresee that it will soon be possible to obtain the full genome of a person for less than 1000$.

There are many algorithms and dedicated tools to efficiently solve specific problems in the fields, but there is a notable lack in standards and systems to query heterogeneous genomic data. This thesis presents the first results of the efforts of the recently born Genomic Computing Group at Politecnico di Milano. We present a novel approach to design and manage genomic data, starting from a modern Laboratory Information Management System. We did not want to interfere with biologists' pipelines; instead we modeled the information obtained at the end of the most common workflows with a unique model, the Genomic Data Model, or GDM. We present the GenoMetric Query Language, or GMQL, a novel high level system for querying our GDM. GMQL can be used to manage and a retrieve information over vast repositories of genomic data, making integration of different sources possible and easy. We will the details of the first implementation of GMQL and its core algorithms, and we also give some insight of the most recent development of the second version of the system.