CSE601 Project 1: Biomedical Data Warehouse/OLAP System

In this project, you are asked to implement a clinical and genomic data warehouse based on your schema design using the Oracle system. A good data warehouse should satisfy the following requirements: 1) support regular and statistical OLAP operations; 2) be robust to potential changes in the future; and 3) support knowledge discovery.

The original data will be provided in the plain text files under the directory /projects/azhang/cse601. A detailed description of the file format is attached at the end. The information related to Oracle system can be found at: http://www.cse.buffalo.edu/HELP/UNIX/Oracle.html.

Part I:
You are required to implement your data warehouse schema in the Oracle system. Then populate your data warehouse with the provided data sets.

Part II:
Your data warehouse is supposed to support the regular OLAP operations (e.g., roll-up, drill down, slice, dice and pivot), as well as some statistical operations (e.g., t-test, ANOVA, and correlation). In the following are some typical queries by users. You may use either SQL, PL/SQL, or external programs (e.g. in Java) to answer the queries. Notice that you should retrieve the data from the Oracle system instead of the original plain text files. Report your approach and the results returned by your data warehouse.

- List the number of patients who had “tumor”, “leukemia” and “ALL”, respectively.
- List the types of drugs which have been applied to patients with “tumor”.
- For each sample of patients with “ALL”, list the mRNA values (expression) of probes in cluster id “00002” for each experiment with measure unit id = “001”.
- For probes belonging to GO with id = “0012502”, calculate the t statistics of the expression values between patients with “ALL” and patients without “ALL”.
- For probes belonging to GO with id=”0007154”, calculate the F statistics of the expression values among patients with “ALL”, “AML”, “colon tumor” and “breast tumor”.
- For probes belonging to GO with id=”0007154”, calculate the average correlation of the expression values between two patients with “ALL”, and calculate the average correlation of the expression values between one “ALL” patient and one “AML” patient.

Part III:
Use your data warehouse and the OLAP operations to support knowledge discovery.
1. Given a specific disease, find the informative genes.
For example, suppose we are interested in the cancer “ALL”.

1) Find all the patients with “ALL” (group A), while the other patients serve as the control (group B).
2) For each gene, calculate the t-statistics for the expression values between group A and group B.
3) If the p-value of the t-test is smaller than 0.01, this gene is regarded as an “informative” gene.

2. Use informative genes to classify a new patient.
For example, given a new patient P_N, we want to predict whether he/she has “ALL”.

1) Find the informative genes w.r.t. “ALL”.
2) Find all the patients with “ALL” (group A).
3) For each patient P_A in group A, calculate the correlation r_A of the expression values of the informative genes between P_N, and P_A,
4) Patients without “ALL” serve as the control (group B).
5) For each patient P_B in group B, calculate the correlation r_B of the expression values of the informative genes between P_N, and P_B.
6) Apply t-test on r_A and r_B, if the p-value is smaller than 0.01, the patient is classified as “ALL”.

The data file with respect to each entity will start with a row describing the fields of the entity. Then each following row in the file corresponds to one instance of the entity.

1. Clinical data space
Entities: patient, disease, drug, test and sample
Fact table: clinical_fact

File: patient.txt

p_id	ssn	name	gender	DOB

File: disease.txt

ds_id	name	type	description

File: drug.txt

dr_id	name	type	description

File: test.txt

tt_id	name	type	setting

File: clinical_fact.txt

p_id	ds_id	sympton	ds_from	ds_to	dr_id	dosage	dr_from	dr_to	tt_id	result	tt_date	s_id

2. Sample data space
Entities: sample, marker, assay, term

Fact table: sample_fact

File: sample.txt

s_id	source	amount	sp_date

File: marker

mk_id	name	type	locus	description

File: assay.txt

as_id	name	type	setting	description

File: term.txt

tm_id	name	type	setting

File: sample_fact.txt

s_id	mk_id	mk_result	mk_date	as_id	as_result	as_date	tm_id	tm_description

3. Microarray and proteomic data space
Entities: probe, measureUnit
Fact table: microarray_fact

File: probe.txt

pb_id	UID	name	description	isQC

File: measureUnit.txt

mu_id	name	type	description

File: microarray_fact.txt

s_id	e_id	pb_id	mu_id	expression

4. Gene data space
Entites: gene, go, cluster, domain, promoter
Fact table: gene_fact

File: gene.txt

UID	seqType	accession	version	seqDataset	speciesID	status

File: go.txt

go_id	accession	type	name	definition

File: cluster.txt

cl_id	num	pattern	tool	tSetting	description

File: domain.txt

dm_id	type	db	accession	title	length	description

File: promoter.txt

pm_id	type	sequence	length	description

File: gene_fact.txt

UID	go_id	cl_id	dm_id	pm_id	UID2

5. Experiment data space
Entities: experiment, project, platform, norm, person, protocal, publication
Fact table: experiment_fact

File: experiment.txt

e_id	name	type

File: project.txt

pj_id	name	investigator	description

File: platform.txt

pf_id	hardware	software	settings	description

File: norm.txt

nm_id	type	software	parameters	description

File: person.txt

pn_id	name	labName	contact

File: protocal.txt

pt_id	name	text	createdBy

File: publication.txt

pu_id	pub_med_id	title	authors	abstract	pubDate

File: experiment_fact.txt

e_id	nm_id	pj_id	pn_id	pf_id	pt_id	pu_id