SE305 Database System Technology – Lab Project

Final Report/Code/Data Due: Dec 7, 2018

Introduction

This is a group-based mini research project. You are given four types of data sources. Each group picks one data source of your choice. By Thursday**, Oct 25**, each group must email Shanshan(florahuangss@163.com) your choice of data source (give your 1st and 2nd choices). If a data source is chosen by too many groups, it will be allocated on a first-come, first-serve basis.

You task is as follows:

1. Study and understand the chosen data source, and design some useful queries.
2. Design a relational database (a set of schemas) for the data, so that you can load the data into database and submit queries you designed and the queries listed in the requirements.
3. Verify the design of your models. You should optimize your design (e.g., by implementing indexes), so that the database can return results of the queries within reasonable time.
4. Implement a simple graphical user interface to perform your queries, so that I can check the performance of your design.
5. Prepare a presentation of 5-10 mins to showcase your preliminary design.
6. Write a report to describe your design and implementation and hand out all source code (Crawler code, MySQL script, any other codes).

Specifications

1. Data

Each given data source is accompanied by a description and some example queries (known as requirements).

Description will be the basic information about the data, such as the format of file, the structure of the data, the meaning of the fields in data, etc.

Requirements are a number of queries written in English. These must be supported by your database design efficiently.

2. Model

A database model is a type of data model that determines the logical structure of a database and fundamentally determines in which manner data can be stored, organized, and manipulated. The most popular example of a database model is the relational model, which uses a table-based format. (See Wikipedia).

You need to design a model for the given data using ER diagrams, and convert the model to schemes of a real relational database (MySQL).

3. Optimization

The initial version usually cannot satisfy the all the demands. In order to make your model fast or efficient for your proposed queries, you need to revise the design of your model and adjust parameters of database to make your database run faster.

The potential optimization can include but is not limited to refining table design, building index, distributing nodes, etc. However, no matter what kind of optimization you use, you should give the persuasive reason that you really need to do it and it indeed has certain effects. In other words, you need to design solid experiments to demonstrate your design.

4. Experiment Design

Experiments are designed to demonstrate your ideas. An effective experiment should at least contain the following parts:

- Hardware Specifications
- Dataset
- Test Queries
- Initialization Scripts
- Experiment Procedure
- Result Analysis

For each of your optimization, you should design an experiment to demonstrate the advantages and disadvantages and whether you will take that optimization as part of your design.

Say, an index might improve speed of query. However, it also takes disk space to store them. As a result, if one column is not like to be filtered, there is no need to build index on it.

In case you still do not know how to design or set up experiment, you can refer to the first few chapters of *High Performance MySQL*.

5. Graphical User Interface

This is essential for us to check the performance of your design. There is no need to make it very fancy, but it should be able to let others perform your queries with different parameters.

Presentation

Depending on the schedule of this course, you may be required to present your ideas during the last class. The presentation should include your basic ideas and expected experiments. No concrete data is needed.

Deliverables

The final deliverables should include the following items:

- A well-written report to describe your ideas, design, experiments, conclusion, etc.
- All datasets
- Initializing scripts of database system (MySQL).
- All source codes (including SQL scripts).

Submit all of the above electronically to Shanshan(florahuangss@163.com)

Scoring Criteria

Your score will consist of three parts.

The first part comes from the designed queries. They should be meaningful and expressive. You will get up to 10% additional credit if you design additional and interesting queries.

The second part is from the design and optimization of your database model. It should contain your sufficient concern and well-designed experiments to demonstrate your design.

The last part (10%) is from the graphical user interface. It can be simple, but sufficient to test your design with different parameters and/or query inputs.

The percentage for first and second part are different in different task.

Data I - CN-DBpedia

Description:

CN-DBpedia is a large-scale general-purpose structured encyclopedia developed and maintained by the Knowledge Factory Laboratory of Fudan University. Its predecessor is the Fudan GDM Chinese Knowledge Map. CN-DBpedia has extended from the encyclopedia field to more than ten vertical fields such as law, business, finance, entertainment, science and technology, military, education, medical, etc., providing supportive knowledge services for intelligent applications in various industries. CN-DBpedia has the characteristics of huge volume, excellent quality, real-time update and rich API services. CN-DBpedia has become the industry's first choice for opening Chinese knowledge maps.

For our project, one is required to load the CN-DBpedia data dump into MySQL database system, and make suitable tables, indexes and queries for them. You will be learning some knowledge about dealing with hierarchical structure in knowledge graph.

The dataset provided in the website (http://kw.fudan.edu.cn/cndbpedia/search/). CN-DBpedia currently offers Dump data downloads. Contains 9 million+ encyclopedia entities and 67 million+ triad relationships. The description2entity information is 1.1 million+, the summary information is 4 million+, the label information is 19.80 million+, and the infobox information is 41 million+。

You are required to design a relational database schema for storing those encyclopedia entities , and triad relationships. A good starting point would be thinking about how to create tables for entities and the relation between those entities. Besides the given required queries you should design some interesting queries by yourself to demonstrate the design of your dataset model is good. The more and better queries you design, the higher score you will get. You can also extract the entity and relation tuple from wikipedia , Baidu Encyclopedia etc. The more information the better.

Required queries:

1. Given a question with entity and relations in CN-DBpedia and you can return the answer with another entity. For example ,“中国的官方语言是什么？”，“中国” is an entity, “官方语言” is a kind of relation. The answer you should return is “汉语（通用普通话)” which is an entity.
2. Given a question with entities and you can return the answer for the relation between the two or more entities. For example, “周杰伦和昆凌的关系？”, the right answer is “丈夫-

妻子” or “夫妻”.

(hint: you have to do some fuzzy match for entities)

Data II – CN-Probase

Description

CN-Probase is a Chinese Concept Graph that contains 17 millions of entities, 270 thousands of concepts and 33 millions of isA relations. The accuracy of isA relations is more than 95%. Compared to other concept graphs, CN-Probase has two striking advantages. First, it is the largest Chinese Concept Graph and covers most of the common entities. Second, it is strictly organized by entities, which is helpful to understand entities precisely.

It is similar with CN-DBpedia. The largest difference between these two knowledge graph is the relationships in CN-Probase are very simple, mostly is ISA relation. You can find the online demo from http://kw.fudan.edu.cn/cnprobase/search/ which is also maintained by Knowledge Factory Laboratory of Fudan University. However, the data is not directly given, and you should crawl or extract the ISA relation data firstly. You can also extract the entity and relation tuple s from wikipedia , Baidu Encyclopedia etc. The more information the better.

CN-Probase acts as central storage for the hierarchical **structured data** of hypernym and hyponym. You are required to create a relational database to store the entire data you get by yourself. You should carefully design those tables as well to fit the need of queries and large data quantity. A good starting point would be thinking about how to create tables for entities and ISA relationship. Besides the given required queries you should design some interesting queries by yourself to demonstrate the design of your dataset model is good. The more and better queries you design, the higher score you will get.

Required queries:

1. Given an entity, you can return all preceding categories (instance of and subclass of) it belongs to.
2. Given questions about whether two entities belongs to the same category, and you can judge with “True” or “False“. For example, “大豆和萝卜同属种子植物门吗”, the right answer is ”True“

(hint: you have to do some fuzzy match for entities)

Data III – Finance Analysis

Description

Nowadays, the finance analysis of stock market is very hot. There are large scale of data in finance analysis. And the factors for influencing the stock market are hard to analysis. However in fact, the stock changes with the world. That means if we know what happens about the world through news text, we can detect what changes about the world can influence the stock price. This is called finance analysis.

There are some finance knowledge bases which you can refer to, like http://doc.kuaiyutech.com/ . However the existing knowledge base are not update on time. The information are old, and the structure in these knowledge bases may not be suitable for our task. So the first requirement is to design your own database model and prepare the data by yourself. the richer the data, the better. And the data should contain the following content at least:

- More then 100 stock (the stock have come the market more than 5 years and belong to different sector) which contains the type, sector and other detailed information.
- The dairy market of each stock in 5 years (contains Day high, Day low, Daily limit, Down limit, yesterday's close, today's open)
- The finance news every day(mainly contains 2 parts, the latest news about a fixed stock and finance news not about the whole market, like the disaster happened to influence the fruit price)

The stock market is selective, you can choose China stock market or US stock market . And you can refer to many finance websites, like http://business.sohu.com for China and https://www.cnbc.com/ for US.

In this project, you are required to create a relational database to store the entire data. You should carefully design those tables as well to fit the need of queries and large data quantity. Besides the given queries, you can also design extra queries for bonus.

Required queries:

1. Given the name of the stock and the date, you can find all the information about this stock, including High, Low, limits, close open price and news about this stock.
2. Given a period of time, find the sector which performs best.
3. Given a stock, you can find when it comes to bull market and when it comes to bear market.(You can define the bull or bear market by yourself)
4. Given news text (2 kind of finance news), you can analysis which stock can be influenced. (hint: analysis the management domain of one stock and what domain can the news affect)

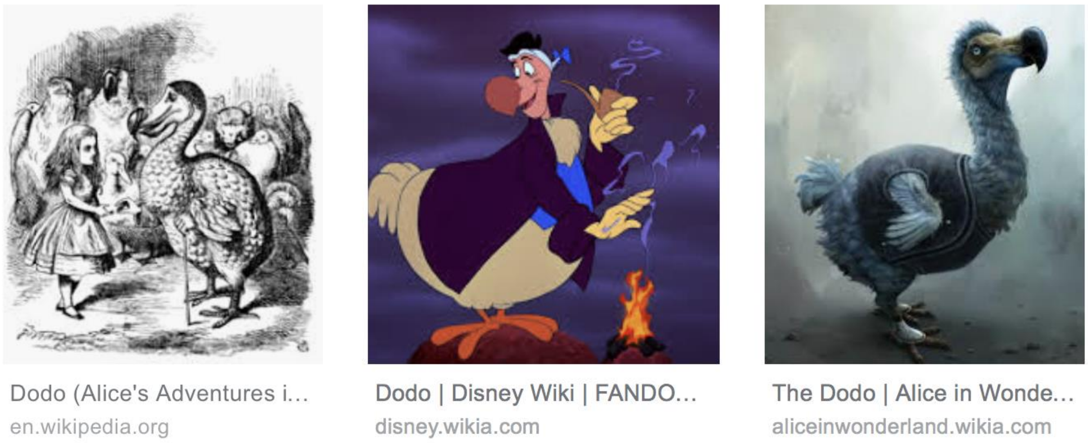
Data IV – Illustration of the books

Description

Books are a rich source of both fine-grained information. In order to improve the readability of the book, many books have illustrations which are strictly illustrative of the text. However the E-book rarely contains illustrations. So in this work you are required to choose illustrations for each chapter in a book. This project can be seen partly as an information retrieval task.

There is a dataset called BookCorpus, but this corpus has not been maintained. Firstly, you are required to crawl all the books in BookCorpus. You can find the booklist on https://www.smashwords.com/books/category/892 and you can download books from https://www.gutenberg.org or other website. What you also need to crawl is a picture dataset about the books. The requirement is that you need to crawl at least one picture about a named entity or a concept in *each* chapter of every book in the data set.

In this project, you are required to create a relational database to store the entire data, the book text and pictures. You can store the books with separate chapters. In each chapter, you are required to extract some important information which you believe. And then, you can label for pictures and match the pictures to the chapter according to the labels you get from some tools and the important information you extract from the chapter text. For example, I can search ”dodo“ in *Alice in Wonderland* , and I will get:



Dodo (Alice's Adventures i…
en.wikipedia.org

Dodo | Disney Wiki | FANDO…
disney.wikia.com

The Dodo | Alice in Wonde…
aliceinwonderland.wikia.com

Required queries:

1. Find the books according to the name, author.
2. Given a book and chapter number, you can return some relative pictures.
3. Given some words, you can find the relative pictures according to the words.