Search and Data Mining: Tools

Text Search

Anya Yarygina

Boris Novikov

Оценки

Активность	**Макс. Оценка**
Проект (каждый из 5)	20
Большой доклад	15
Короткий доклад	5
Устный экзамен (в сессию)	20

Итоговая оценка	**Баллы**
Отлично	> 81
Хорошо	61 - 80
Удовлетв.	40 – 60

Как добиться снижения оценки	
Некачественное выполнение задания (небрежный отчет, отсутствие выводов)	-5
Неполностью выполненное задание (некоторые пункты задания не выполнены)	-10
Задание не отправлено через 2 недели после формулировки	-5
Задание не отправлено через 4 недели после формулировки	-10

Проекты

Project #	Topic	Start Date	Due date
1	Data pre-processing	Feb 21	Mar 06
2	OLAP	Mar 07	Mar 20
3	Text search	Mar 21	Apr 03
4	Classification	Apr 04	Apr 17
5	Clustering	Apr 18	May 03

Отчет отправляется по электронной почте одним письмом на всех преподавателей с темой «Search and Mining Tools and Techniques» до 23:59 (UTC+04:00) дня указанного в таблице.

Структура практического занятия

- Постановка задачи
  - Формулировка задачи
  - Описание требований
  - Советы по выполнению задания
  - Вопросы по выполнению задания
- Обсуждение предыдущей задачи
  - «Идеальные» отчеты
  - Отсутствующие отчеты
  - Вопросы по отчетам

Задание 1: Подготовка данных

- Выбрать наборы данных для практических заданий из предложенных или согласовать любые другие и ознакомиться с ними
- Текстовый отчет с содержательным описанием 4 выбранных наборов данных для каждого практического задания

Задание 2: OLAP

- Выполнение задания включает
  - выбор и подготовку модели;
  - проектирование хранилища данных;
  - генерацию и загрузку данных;
  - проектирование запросов для получения нескольких отчетов;
  - построение отчетов
- Краткий текстовый отчет
  - Неформальное описание задачи и данных
  - Описание формализации задачи (какие размерности определены, какие меры, атрибуты и т.д.)
  - Описание запросов, извлекающих данные из полученного куба (какие содержательные результаты предполагалось получить и какими запросами это достигается)

Задание 3: Text Search

- Выполнение задания включает
  - выбор коллекции текстовых документов и задачи поиска
  - выбор структуры документов
  - загрузка документов в индекс
  - эксперименты с разными типами текстовых запросов:
    - Текстовые запросы
    - AND, OR, LIKE
    - Запросы к разными полям документа
- Краткий текстовый отчет
  - Неформальное описание задачи и данных
  - Описание отображения формальной задачи на данные (документ, поля с типами)
  - Описание инструмента индексирования и поиска (название, параметры, подходы)
  - Описание процесса загрузки данных
  - Описание процесса поиска: осмысленные запросы разного типа и результаты
  - Замеры времени выполнения запросов, если загружается вся коллекция или только половина

Задание 3: Инструменты

- Solr(Lucene)
- Database

Задание 3: Пример

- Lucene API
  - File system directory
- Reviews of cars
  - Year
  - Model
  - Author
  - Date
  - Text
  - Favorite

- Parse data
- Index data
- Query data
  - audi
  - audi AND bmw
  - text:(audi) AND favorite:(voice)
  - text:(audi OR bmw) AND favorite:"voice command"
  - model:acura*
  - date:"12/13/2008"
  - year:[2006 TO 2007]

Задание 3: Text Search

- Выполнение задания включает
  - выбор коллекции текстовых документов и задачи поиска
  - выбор структуры документов
  - загрузка документов в индекс
  - эксперименты с разными типами текстовых запросов:
    - Текстовые запросы
    - AND, OR, LIKE
    - Запросы к разными полям документа
- Краткий текстовый отчет
  - Неформальное описание задачи и данных
  - Описание отображения формальной задачи на данные (документ, поля с типами)
  - Описание инструмента индексирования и поиска (название, параметры, подходы)
  - Описание процесса загрузки данных
  - Описание процесса поиска: осмысленные запросы разного типа и результаты
  - Замеры времени выполнения запросов, если загружается вся коллекция или только половина

Задание 2: OLAP

- Выполнение задания включает
  - выбор и подготовку модели;
  - проектирование хранилища данных;
  - генерацию и загрузку данных;
  - проектирование запросов для получения нескольких отчетов;
  - построение отчетов
- Краткий текстовый отчет
  - Неформальное описание задачи и данных
  - Описание формализации задачи (какие размерности определены, какие меры, атрибуты и т.д.)
  - Описание запросов, извлекающих данные из полученного куба (какие содержательные результаты предполагалось получить и какими запросами это достигается)