Search and Data Mining: Tools

Introduction

Anya Yarygina

Boris Novikov

Организация курса

- Теоретические занятия (в основном лекции)
  - Обсуждение методов, подходов, алгоритмов
  - Доклады студентов на теоретических занятиях (далее «Большие доклады, 45 минут)
- Практические занятия
  - Проекты, выполняемые самостоятельно
  - Обсуждение инструментов для выполнения проектов
  - Индивидуальные задания для проектов
  - Проверка результатов выполнения проектов
  - Короткие доклады (15 мин.)

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

OLAP

- Выполнение задания включает: выбор и подготовку модели; проектирование хранилища данных; генерацию и загрузку данных; проектирование запросов для получения нескольких отчетов; построение отчетов
- Возможные данные
  - Интернет-магазин
    - Товар
      - Категория (например: книги, музыкальные записи, электроника)
      - Группа (например, для книг и музыки – жанр, для электроники – компьютеры, телефоны, аудио-видео)
      - Модель
    - Покупатель
      - Почтовый индекс доставки
      - Населенный пункт
      - Регион
    - Факты описывают заказы и содержат количество единиц товара и стоимость.
  - Экзаменационная сессия
    - Студент
      - Год обучения
      - Направление
      - Группа
    - Дисциплина
      - Кафедра
      - Преподаватель
      - Тип (обязательная, по выбору, общая, специальная и т.п.)
    - Факты могут содержат экзаменационную оценку.
- 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)

Информационный поиск

- Выбрать коллекцию текстовых документов, загрузить в индекс и поэкспериментировать с разными типами текстовых запросов: AND, OR, LIKE, NEAR и запросами к разными полям.
- Возможные данные:
  - Новостные статьи, сгруппированные по авторам
    http://archive.ics.uci.edu/ml/datasets/Reuter_50_50
  - Мнения пользователей по отелям в разных городах
    http://archive.ics.uci.edu/ml/datasets/OpinRank+Review+Dataset
  - Мнения пользователей по маркам автомобилей в разные годы
    http://archive.ics.uci.edu/ml/datasets/OpinRank+Review+Dataset
- Краткий текстовый отчет
  - Неформальное описание задачи и данных, использованных при выполнении задания
  - Отображение формальной задачи на данные (объект, поля с типами), параметры выбранного индекса
  - Описание процесса загрузки данных
  - Описание процесса поиска: осмысленные запросы разного типа и результаты
  - Замеры времени выполнения запросов, если загружается вся коллекция или только половина.

Классификация

- Задание состоит в том, чтобы запустить в любом удобном инструменте минимум три существенно отличающихся алгоритма классификации, сравнить результаты. Для одного из алгоритмов классификации, выбранного на предыдущем шаге, проводят эксперименты с выбором атрибутов (любой алгоритм) и ансамблем (любой алгоритм bagging или boosting).
- Возможные данные:
  - Данные по определению вида растения по признакам изображения листа
    http://archive.ics.uci.edu/ml/datasets/One-hundred+plant+species+leaves+data+set
  - Данные по определению автора новости
    http://archive.ics.uci.edu/ml/datasets/Reuter_50_50
  - Данные по распознаванию букв
    http://archive.ics.uci.edu/ml/datasets/Letter+Recognition
  - Данные по классификации грибов
    http://archive.ics.uci.edu/ml/datasets/Mushroom
  - Данные по классификации вина по типу или качеству
    http://archive.ics.uci.edu/ml/datasets/Wine+Quality
    http://archive.ics.uci.edu/ml/datasets/Wine
- Краткий текстовый отчет
  - Неформальное описание задачи и данных, использованных при выполнении задания
  - Отображение формальной задачи на данные (объект, атрибуты с типами, класс), параметры выбранных алгоритмов
  - Описание результатов классификации (оценки качества и confusion matrix) и выводы по каждому эксперименту

Кластеризация

- Задание состоит в том, чтобы запустить в любом удобном инструменте минимум три существенно отличающихся алгоритма кластеризации и сравнить результаты на основе визуализации. Выполнение задания предполагает выполнение одного алгоритма кластеризации, параметризуемого числом кластеров (перебор в небольшом диапазоне значений 2-6), и одного алгоритма кластеризации с автоматическим определением числа кластеров с последующей визуализацией результата, что позволяет выбрать настоящее число кластеров и попробовать еще 1-2 алгоритма кластеризации с заданным числом кластеров.
- Возможные данные:
  - Двумерные данные на основе генерации кластеров по картинке с объектами http://www.math.spbu.ru/user/boris_novikov/courses/index.shtml
- Краткий текстовый отчет
  - Неформальное описание задачи и данных, использованных при выполнении задания
  - Отображение формальной задачи на данные (объект, атрибуты), параметры выбранных алгоритмов
  - Описание результатов каждого запуска алгоритма кластеризации с картинкой и выводами

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
  - Набор данных
  - Предметная область
  - Объекты, атрибуты, меры
  - Аналитическая задача