ПОЛИТЕХ

Об обучении технологиям интеллектуального анализа данных в рамках курса «Корпоративные системы баз данных»

Нестеров Сергей Александрович
к.т.н., доцент кафедры
«Системный анализ и управление»
Санкт-Петербургский политехнический университет

Предыстория

Изначально курс по интеллектуальному анализу данных (data mining) средствами SQL Server задумывался как факультатив.

Цели:

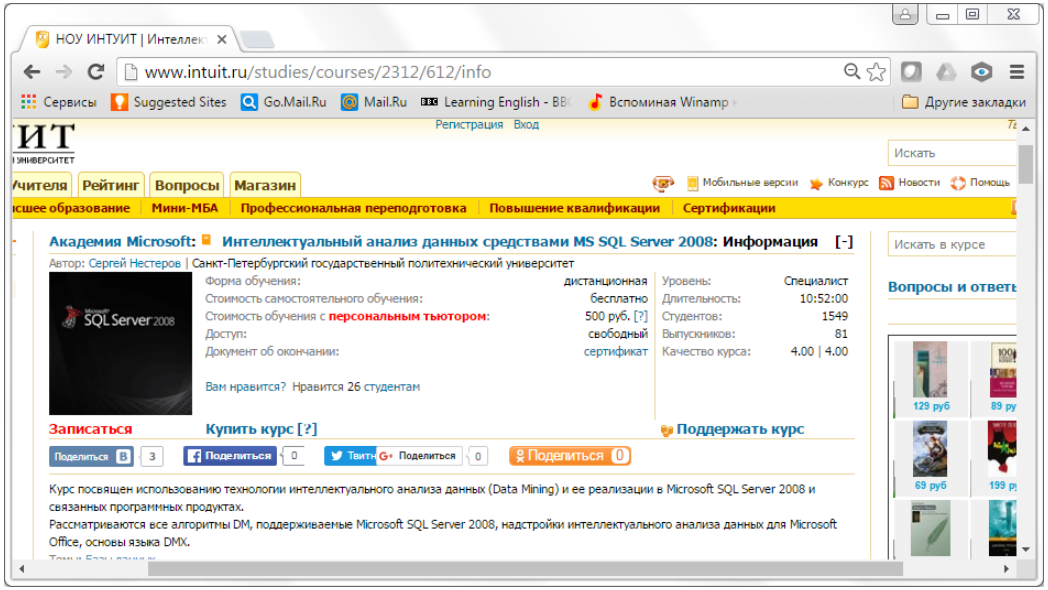
- Дать студентам представление о возможных приложениях data mining при анализе данных, организованных в виде баз данных или электронных таблиц.

- Развитие навыков работы с базами данных.

Предыстория

В 2011/2012 учебном году в рамках гранта Microsoft и ИНТУИТ.ру на разработку учебного курса, был подготовлен курс «Интеллектуальный анализ данных средствами MS SQL Server 2008».

http://www.intuit.ru/studies/courses/2312/612/info



Благодаря проекту

- Материалы подготовлены и оформлены в соответствии с требованиями ИНТУИТ.ру (что облегчило в дальнейшем и создание курса на портале СПбПУ на платформе Moodle)

- Разработан банк тестовых заданий

Корпоративные системы баз данных

В настоящее время на кафедре «Системный анализ и управление» СПбПУ данный материал преподается в рамках дисциплины «Корпоративные системы баз данных».

Направления подготовки:

27.04.03 - "Системный анализ и управление"

09.04.02 - "Информационные системы и технологии“

Продолжительность этой части курса:

6 недель (2 ч. лекций и 2 ч. лабораторных в неделю).

Основные разделы

- Введение: задачи, решаемые средствами интеллектуального анализа, примеры приложений, особенности работы с базами данных;

- Архитектура SQL Server, Analysis Services;

- Алгоритмы интеллектуального анализа данных;

- Основы языка DMX.

Достоинства используемого инструментария

- Знакомые программные средства: Excel надстройками интеллектуального анализа, SQL Server (знакомый по курсу «Базы данных»).

- В рамках академических программ SQL Server developer edition бесплатно предоставляется как учебному заведению, так и студентам (через DreamSpark).

Достоинства используемого инструментария

Известная из других курсов учебная база – AdventureWorks, на которой можно показать основные задачи data mining:

- кластеризация (сегментация клиентов магазина);
- классификация (покупателей);
- прогнозирование временных рядов (анализ продаж по месяцам);
- ассоциативные правила (анализ покупательской корзины).

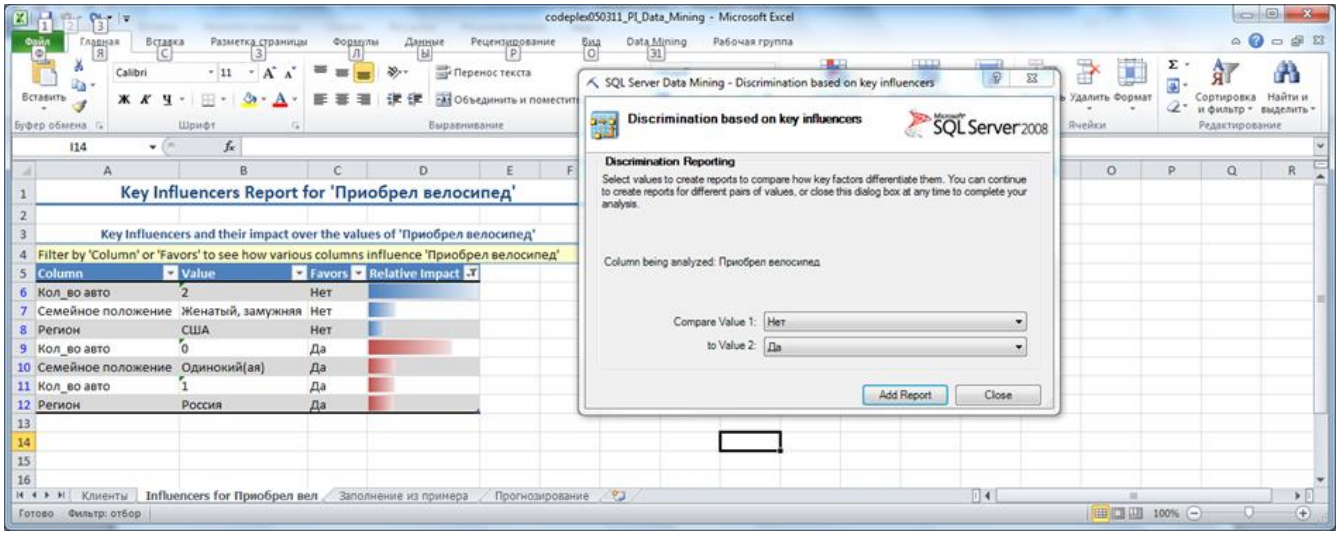
Дополнительные материалы

- Возможность использовать переведенные на русский материалы msdn в качестве дополнительных источников.

- Появление MOOC от Microsoft на платформе edx.org

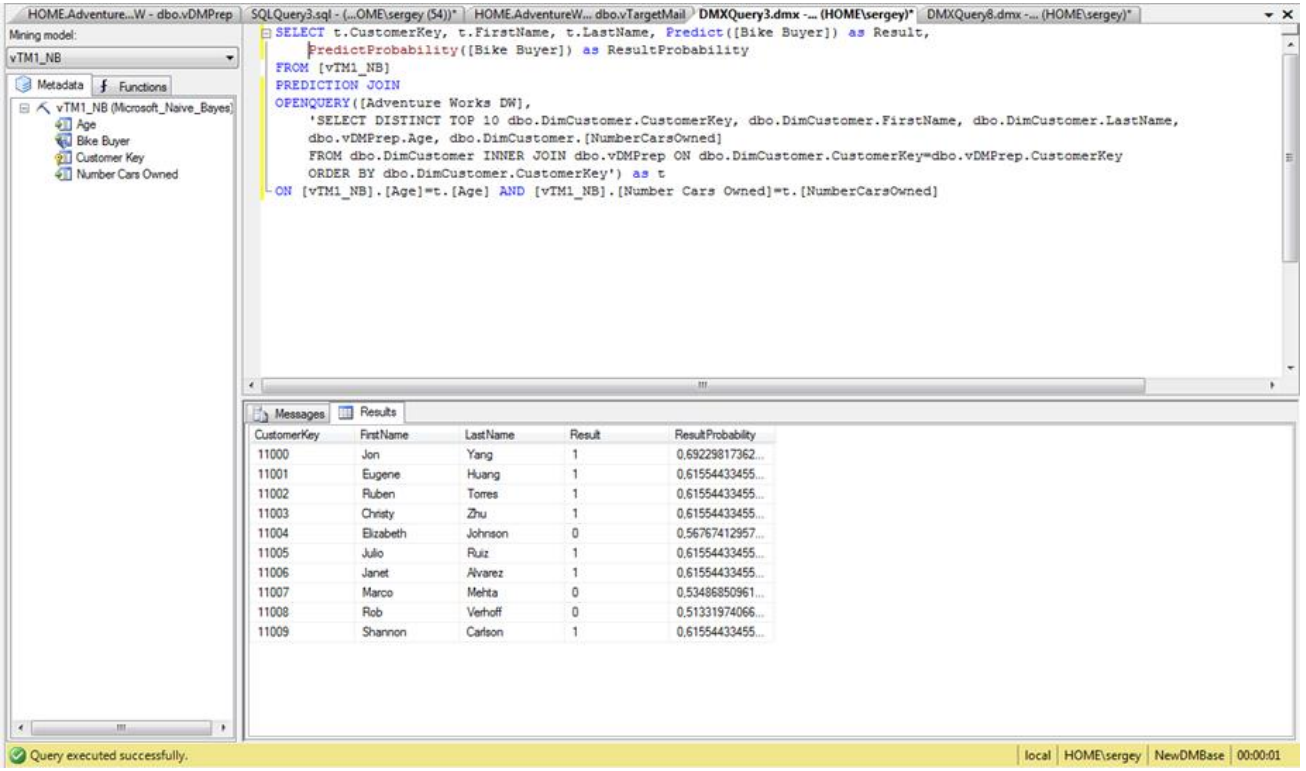
Лабораторный практикум

Часть 1: обработка данных с использованием Excel с надстройками интеллектуального анализа.
Цель: познакомить с задачами, используя простой инструментарий, не вдаваясь в подробности обработки данных на стороне сервера.



Лабораторный практикум

Часть 2: работа с аналитическими службами из BI Dev Studio (для SQL Server 2008 R2) и Management Studio.



Итоговое задание

Практическую часть курса завершает небольшой самостоятельный проект, включающий:

- Подготовку или выбор набора данных

(источники http://archive.ics.uci.edu/ml/datasets.html http://poligon.machinelearning.ru/DataSet/List.aspx и т.д.)

- Постановку и решение задачи интеллектуального анализа.

Итоговое задание

Некоторые темы:

- Классификация почтового траффика (обнаружение спама).
- Классификация ирисов.
- Классификация грибов (съедобные/ несъедобные).

Примеры тем выпускных работ

«Использование методов интеллектуального анализа данных в сфере Интернет-торговли»

Использование модели на основе алгоритма Microsoft Association Rules для анализа покупательской корзины в интернет-магазине: система предлагает покупателю дополнительные товары, основываясь на уже выбранных.

Примеры тем выпускных работ

«Разработка методики оценки тестовых заданий в среде дистанционного обучения»

Из системы дистанционного обучения Moodle после прохождения тестов брался набор показателей для тестовых заданий (коэффициент легкости, стандартное отклонение результата …). На основе типа тестового задания и полученных результатов проводилась кластеризация.

Цель – корректно назначить баллы за задание в соответствии с их трудностью, выявить некорректные или слишком простые задания.

Примеры тем выпускных работ

«Прогнозирование параметров функционирования теплотехнической установки с использованием методов интеллектуального анализа данных»

Анализировалась база данных с параметрами работы и данными о расходе топлива в автоматизированных котельных. Цель – определить режим работы, который ранее в аналогичных условиях был наиболее эффективен с точки зрения расхода топлива.

Некоторые итоги

- Несмотря на ограниченное время выделенное на модуль, связанный с data mining, использование знакомого инструментария позволяет достаточно быстро научить основам и добиться получения практических результатов.
- Очень важно выбрать такой учебный набор данных, чтобы предметная область была хорошо знакома студентам.
- В силу ограниченности во времени, в рамках курса пока не рассматриваются «нереляционные» источники данных.

Спасибо за внимание !