

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>

The screenshot shows a web browser window displaying the course page on the Intuit.ru website. The browser's address bar shows the URL www.intuit.ru/studies/courses/2312/612/info. The website header includes the Intuit logo and navigation links such as 'Сервисы', 'Suggested Sites', 'Go.Mail.Ru', 'Mail.Ru', 'Learning English - BB', and 'Вспомогательная Winamp'. The main content area features a yellow navigation bar with tabs for 'Читателя', 'Рейтинг', 'Вопросы', and 'Магазин'. Below this, a yellow banner highlights various educational services: 'Повышение квалификации', 'Сертификации', 'Мини-МБА', and 'Профессиональная переподготовка'. The main course listing is for 'Академия Microsoft: Интеллектуальный анализ данных средствами MS SQL Server 2008: Информация [-]'. The author is listed as 'Сергей Нестеров' from 'Санкт-Петербургский государственный политехнический университет'. The course details are as follows:

Форма обучения:	дистанционная	Уровень:	Специалист
Стоимость самостоятельного обучения:	бесплатно	Длительность:	10:52:00
Стоимость обучения с персональным тьютором:	500 руб. [?]	Студентов:	1549
Доступ:	свободный	Выпускников:	81
Документ об окончании:	сертификат	Качество курса:	4.00 4.00

Below the details, there is a 'Вам нравится?' section showing 'Нравится 26 студентам'. There are buttons for 'Записаться', 'Купить курс [?]', and 'Поддержать курс'. Social sharing options for Facebook, Twitter, and Google+ are also present. The footer contains a brief description of the course: 'Курс посвящен использованию технологии интеллектуального анализа данных (Data Mining) и ее реализации в Microsoft SQL Server 2008 и связанных программных продуктах. Рассматриваются все алгоритмы DM, поддерживаемые Microsoft SQL Server 2008, надстройки интеллектуального анализа данных для Microsoft Office, основы языка DMX.'

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.

The screenshot displays the Microsoft Excel interface with a 'Data Mining' add-in. The main window shows a 'Key Influencers Report for 'Приобрел велосипед'' with a table of factors and their relative impact. A dialog box titled 'SQL Server Data Mining - Discrimination based on key influencers' is open, allowing the user to select values for comparison.

Column	Value	Favors	Relative Impact
Кол_во авто	2	Нет	
Семейное положение	Женатый, замужняя	Нет	
Регион	США	Нет	
Кол_во авто	0	Да	
Семейное положение	Одинокий(ая)	Да	
Кол_во авто	1	Да	
Регион	Россия	Да	

Discrimination Reporting
Select values to create reports to compare how key factors differentiate them. You can continue to create reports for different pairs of values, or close this dialog box at any time to complete your analysis.

Column being analyzed: Приобрел велосипед

Compare Value 1: Нет
to Value 2: Да

Buttons: Add Report, Close

Лабораторный практикум

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

The screenshot displays the SQL Server Enterprise Manager interface. On the left, the 'vTMI_NB' mining model is selected, showing its metadata: Age, Bike Buyer, Customer Key, and Number Cars Owned. The central pane shows a DMX query that uses the Predict function to forecast 'Bike Buyer' status based on demographic data. The bottom pane shows the query results as a table with columns for CustomerKey, FirstName, LastName, Result, and ResultProbability.

```
SELECT t.CustomerKey, t.FirstName, t.LastName, Predict([Bike Buyer]) as Result,
       PredictProbability([Bike Buyer]) as ResultProbability
FROM [vTMI_NB]
PREDICTION JOIN
OPENQUERY([Adventure Works DW],
'SELECT DISTINCT TOP 10 dbo.DimCustomer.CustomerKey, dbo.DimCustomer.FirstName, dbo.DimCustomer.LastName,
dbo.vMPrep.Age, dbo.DimCustomer.[NumberCarsOwned]
FROM dbo.DimCustomer INNER JOIN dbo.vMPrep ON dbo.DimCustomer.CustomerKey=dbo.vMPrep.CustomerKey
ORDER BY dbo.DimCustomer.CustomerKey') as t
ON [vTMI_NB].[Age]=t.[Age] AND [vTMI_NB].[Number Cars Owned]=t.[NumberCarsOwned]
```

CustomerKey	FirstName	LastName	Result	ResultProbability
11000	Jon	Yang	1	0.69229817362...
11001	Eugene	Huang	1	0.61554433455...
11002	Ruben	Tones	1	0.61554433455...
11003	Christy	Zhu	1	0.61554433455...
11004	Elizabeth	Johnson	0	0.56767412957...
11005	Juko	Ruiz	1	0.61554433455...
11006	Janet	Alvarez	1	0.61554433455...
11007	Marco	Mehta	0	0.53486850961...
11008	Rob	Verhoff	0	0.51331974066...
11009	Shannon	Carlson	1	0.61554433455...

Query executed successfully. | local | HOME\sergey | NewDMBase | 00:00:01

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 научиться 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 !