III межвузовская конференция

Бизнес-аналитика. Использование аналитической платформы Deductor в учебном процессе вуза

BaseGroup Labs
ТЕХНОЛОГИИ АНАЛИЗА ДАННЫХ

Опыт решения учебных и практических задач на аналитической платформе Deductor

на кафедре «Информатика и программное обеспечение»
Брянского государственного технического университета

Подвесовский Александр Георгиевич
заведующий кафедрой, к.т.н., доцент

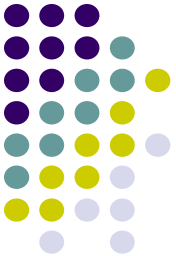
apodv@tu-bryansk.ru

Лагерев Дмитрий Григорьевич
к.т.н., доцент

lagerevdg@mail.ru

г. Москва, 28 июня 2016 г.

Содержание



- О кафедре «Информатика и программное обеспечение»
- О курсе «Интеллектуальный анализ данных»
- Проекты, выполненные с использованием платформы Deductor
- Дальнейшие планы

Содержание

- **О кафедре «Информатика и программное обеспечение»**
- О курсе «Интеллектуальный анализ данных»
- Проекты, выполненные с использованием платформы Deductor
- Дальнейшие планы

Краткая историческая справка

- Кафедра «Информатика и программное обеспечение» создана в Брянском государственном техническом университете в 1989 г.
  - (до 1997 г. называлась «Вычислительная техника и прикладная математика»)
- 1995 г. – первый набор на специальность «Программное обеспечение вычислительной техники и автоматизированных систем»
- 2004 г. – первый набор на специальность «Математическое обеспечение и администрирование информационных систем»
- 2009 г. – открытие магистратуры по направлению «Информатика и вычислительная техника»
  - (до 2014 г. в магистратуру принимались выпускники специалитета)
- 2011 г. – переход на двухступенчатую систему подготовки, первый набор на направление подготовки бакалавров «Программная инженерия»
- 2015 г. – открытие магистратуры по направлению «Программная инженерия»

Кафедра сегодня

- Крупнейшая выпускающая кафедра университета
  - Более 300 студентов дневной формы обучения
  - Более 250 студентов заочной формы обучения и слушателей программ дополнительного образования
  - Численность профессорско-преподавательского состава – 36 человек (± 1-2)
    - в том числе 1 профессор и 26 доцентов
    - средний возраст ППС – менее 40 лет
- Имеется аспирантура, и ведутся научные исследования по следующим направлениям
  - многомерные и пространственно-временные структуры данных, методы поиска информации
  - интеллектуальные системы поддержки принятия решений в управлении и проектировании
  - интеллектуальный анализ данных и машинное обучение
  - мобильные системы и робототехника

Реализуемые образовательные программы (на основе ФГОС ВО 3+)

Бакалавриат

02.03.03 – Математическое обеспечение и администрирование информационных систем

09.03.01 – Информатика и вычислительная техника

- Программное обеспечение вычислительной техники и автоматизированных систем

09.03.04 – Программная инженерия

- Разработка программно-информационных систем

Магистратура

09.04.01 – Информатика и вычислительная техника

- ***Компьютерный анализ и интерпретация данных***
- Информационное и программное обеспечение вычислительных систем

09.04.04 – Программная инженерия

- Проектирование программно-информационных систем

Аспирантура

05.13.18 – Математическое моделирование, численные методы и комплексы программ

05.13.10 – Управление в социальных и экономических системах

Магистерская программа «Компьютерный анализ и интерпретация данных»

- Реализуется в рамках направления «Информатика и вычислительная техника» с 2009 г. (первая магистерская программа, открытая на кафедре)
  - 2009-2010 гг. – на основе ГОС ВПО 2-го поколения
  - 2011-2013 гг. – на основе ФГОС ВПО 3-го поколения
  - с 2014 г. – на основе ФГОС ВО 3+ (программа академической магистратуры)
- К настоящему моменту выпущено 25 магистров
  - с 2017 г. планируется выпускать до 10 магистров ежегодно
- Основной целью является углубленная (на базе соответствующих направлений бакалавриата) подготовка профессиональных разработчиков программного обеспечения и системных аналитиков со специализацией в следующих областях
  - **обработка и анализ больших объемов данных, методы машинного обучения**
  - **модели и методы поддержки принятия решений**
  - **интеллектуальные системы на основе мягких вычислений**
  - **обработка и анализ изображений, машинное зрение**
  - **цифровая обработка сигналов**

Ключевые дисциплины учебного плана

1. Базовые дисциплины направления

- Методы оптимизации
- Теория принятия решений
- Компьютерное моделирование
- Теория систем и системный анализ
- Технология проектирования, разработки и верификации программного обеспечения

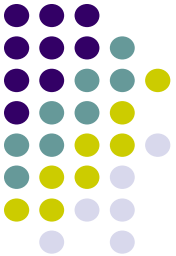
2. Математический аппарат, методология и инструменты анализа данных

- Статистический анализ данных
- **Интеллектуальный анализ данных**
- Теория нейронных сетей
- Интеллектуальные системы

3. Специальные разделы и приложения методологии анализа данных

- **Хранилища данных**
- Системы с параллельной обработкой данных
- Обработка и анализ изображений
- Системы машинного зрения
- Цифровая обработка сигналов

Содержание



- О кафедре «Информатика и программное обеспечение»
- **О курсе «Интеллектуальный анализ данных»**
- Проекты, выполненные с использованием платформы Deductor
- Дальнейшие планы

Интеллектуальный анализ данных: общие сведения о курсе

- ***Цели дисциплины***
  - изучение современных информационных технологий, предназначенных для интеллектуального анализа данных
  - формирование целостного представления об анализе и интерпретации данных, как о процессе поиска, так и о методологии применения скрытых в них закономерностей для достижения поставленных целей
- ***Структура курса***
  - изучается на 1-м курсе во 2-м семестре
  - лекции – 17 часов
  - лабораторные работы – 34 часа
  - курсовой проект
  - экзамен
- ***Теоретическая часть:***
  - структура и возможности аналитических систем
  - основные методы интеллектуального анализа и предобработки данных
  - процесс ETL
  - ансамбли моделей
- ***Практическая часть:*** использование **Deductor Studio Academic**

Интеллектуальный анализ данных: курсовой проект

- Цель – приобретения ***практических навыков*** в работе по освоению и применению методов и инструментов интеллектуального анализа данных для решения прикладных задач в различных предметных областях
- Выполняется только на ***реальных данных***
- Задачи магистранта при выполнении курсового проекта:
  - Найти реальный набор данных
    - Если кто не может, ему выдается база данных магазина или оптового склада, содержащая данные за год
  - Выдвинуть гипотезы
  - Обеспечить выгрузку требуемых данных в формат *.csv
  - Выполнить импорт данных в Deductor Studio Academic
  - Выполнить необходимую трансформацию данных (оценка качества, очистка, фильтрация и т.д.)
  - Выполнить интеллектуальный анализ данных
    - Методы могут быть предложены преподавателем либо выбраны самим магистрантом
  - Сделать выводы

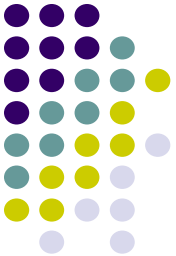
Курсовой проект: задача поиска ассоциативных правил

- Любимая задача слабых и/или не инициативных магистрантов – поиск ассоциативных правил в данных о продажах
- Задача обычно решается «в лоб», что приводит к нахождению правил вида
  - если *хлеб*, то *пакет* (поддержка 8,21 %, достоверность 50,64 %)
  - если *пакет*, то *хлеб* (поддержка 8,21 %, достоверность 32,98 %)
  - если *пакет*, то *молоко* (поддержка 7,27 %, достоверность 29,18 %)
  - если *молоко*, то *пакет* (поддержка 7,27 %, достоверность 54,46 %)
  - если *батон*, то *хлеб* (поддержка 1,58 %, достоверность 48,99 %)
  - если *сыр*, то *колбаса* (поддержка 2,04 %, достоверность 43,31 %)
  - и т.п.
- Более толковые магистранты пробуют выполнять поиск не универсальных ассоциативных правил, а более специализированных, анализируя данные о продажах в выходные и праздничные дни, летнее и зимнее время, обеденное и вечернее время и т.д.

Курсовой проект: типовые темы более высокого уровня

- Прогнозирование временных рядов на основе данных
  - о продажах товаров
  - о посещаемости интернет-ресурсов
  - о курсах валют и т.д.
- Кластеризация
  - студентов
  - покупателей
  - посетителей интернет-ресурсов
  - товаров и т.д.
- Классификация
  - студентов
  - покупателей
  - посетителей интернет-ресурсов,
  - товаров и т.д.

Пример интересного курсового проекта

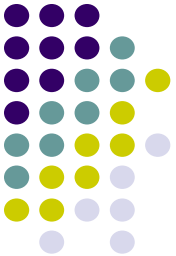


- Цель – анализ данных спортивной активности велосипедистов, предоставленных сервисом Strava, для последующего подбора каждому спортсмену подходящей группы для совместной велосипедной поездки / прогулки / тренировки
- Имеющиеся данные:
  - пол
  - возраст
  - расстояние на каждой поездке
  - скорость на каждой поездке
  - время, за которое спортсмен прошел поездку
  - пройденное расстояние при подъеме на каждой поездке
  - пройденное расстояние при спуске на каждой поездке
  - время начала каждой поездки.
  - время окончания каждой поездки.
- Были получены данные по 836 спортсменам

Пример интересного курсового проекта: выделенные кластеры

- **Молодые активные спортсмены (~ 40%)**
  - Бóльшая часть представителей катается на средней скорости 30 км/ч
  - Все представители преодолевают расстояния больше 70 км, а часть из них способна преодолеть и бóльшие расстояния
  - Бóльшая часть катается по пересеченной местности
  - Все представители совершают не менее 6 поездок за 4 недели
- **Спортсмены с малым количеством поездок (~ 30%)**
  - Количество поездок не превышает 8
  - Преодолевают расстояния менее 70 км
  - Средняя скорость не превышает 26 км/ч
  - Бóльшая часть представителей катается на ровной местности, из чего можно сделать вывод о преобладании поездок по городу
- **Взрослые спортсмены, у которых катание на велосипеде – неотъемлемая часть жизни (~ 30%)**
  - Количество поездок за 4 недели превышает 9
  - В отличие от первого кластера с похожими характеристиками, абсолютно все представители данного кластера катаются с такой периодичностью, что позволяет сделать вывод о профессиональном увлечении велоспортом

Преимущества использования Deductor Studio Academic



- Низкий порог вхождения
- Простой и понятный интерфейс
- Наличие подробной русскоязычной документации
- Доступность примеров готовых проектов
- Легкость выполнения сложных для программирования операций
  - например, поворот набора данных с группировкой и т.п.
- Возможность быстро выполнить анализ имеющихся данных и оценить применимость конкретного алгоритма
- Возможность легко сравнить имеющиеся алгоритмы анализа и выбрать наилучший в имеющихся условиях
- Простота построения ансамблей моделей
- Поддержка кластеризации категорийных данных (метод CLOPE)

Проблемы при использовании Deductor Studio Academic (1)

- ~~**Отсутствие возможности импорта данных напрямую из СУБД**~~
  - *Решена с приобретением Deductor Studio Professional*
- ~~**Отсутствие возможности экспорта результатов в MS Office**~~
  - *Решена с приобретением Deductor Studio Professional*
- **Закрытость**
  - Отсутствие возможности модифицировать алгоритм или подключить библиотеку с собственноручно написанным кодом демотивирует большинство сильных программистов
  - В результате становится невозможным использование Deductor Studio при выполнении большинства магистерских диссертаций
- **Особенности выполнения трансформации данных и их подготовки к анализу**
  - Хорошему программисту бывает проще и привычнее написать один SQL-запрос в СУБД и сразу получить нужные данные, чем разбираться с названиями операций и особенностями их реализации в Deductor Studio

Проблемы при использовании Deductor Studio Academic (2)

- **Недоступность серверной части**
  - Недоступность Deductor Analytic Server и Deductor Integration Server приводит к невозможности разрабатывать приложения (веб-, мобильные и т.д.)
  - Как следствие, это опять-таки не позволяет использовать Deductor Studio при выполнении большинства магистерских диссертаций
- **Ограничения по использованию аппаратных ресурсов компьютера**
  - Используется одно ядро процессора
- **Неустойчивое поведение**
  - «Зависает» при открытии некоторых больших проектов
- **Отсутствие некоторых методов анализа**
  - кластеризация последовательностей
  - поиск иерархических ассоциативных правил и т.д.

Неутешительный вывод

В настоящее время при выполнении магистерских диссертаций большинство магистрантов предпочитает использовать СУБД с возможностью анализа (например MS SQL Server) или библиотеки с открытым исходным кодом

Содержание

- О кафедре «Информатика и программное обеспечение»
- О курсе «Интеллектуальный анализ данных»
- **Проекты, выполненные с использованием платформы Deductor**
- Дальнейшие планы

Центр живой истории «Кветунь». Описание проекта

- Заказчик – Центр живой истории «Кветунь» в г. Брянске
  - Предмет деятельности Центра
    - воссоздание предметов средневековой материальной культуры территории Среднего Подесенья
    - освоение старинных видов ремесел и воинских искусств
  - Это был наш первый опыт работы с реальным заказчиком

- Предмет анализа – сведения об археологических материалах, найденных во время раскопок курганного некрополя X – XIII веков в составе Кветуньского археологического комплекса

Центр живой истории «Кветунь». Особенности данных

- Заказчиком были предоставлены сведения по 213 погребениям
- Для описания использовалось 84 признака, описывающие:
  - параметры кургана (высота и диаметр земляных насыпей)
  - особенности погребения (трупоположение в яме, на горизонте, в насыпи и т.п.)
  - возраст, пол, ориентировка погребенного
  - наличие в захоронении предметов быта, оружия, украшений и амулетов
- Данные были представлены в бинарном виде (факт наличия или отсутствия признака у каждого погребения)
  - Некоторые данные по смыслу были количественными (например, высота и диаметр насыпи), но они также были представлены в бинарном виде
- *Заказчик не предоставил никаких гипотез для проверки, поэтому пришлось экспериментировать с разными видами моделей*
  - Использовались ассоциативные правила, деревья решений и кластеризация
  - Исполнители работали изолированно друг от друга; каждому было поручено исследовать один из видов моделей
- Ансамбль моделей в этом проекте не применялся

Центр живой истории «Кветунь». Трансформация данных

- Для поиска ассоциативных правил исходные данные были преобразованы следующим образом:
  - каждая характеристика кургана была сделана отдельной записью (преобразование к виду транзакций)
  - бинарные характеристики были заменены параметрами, содержащими исходное название параметра и «+» или «–» в зависимости от значения параметра

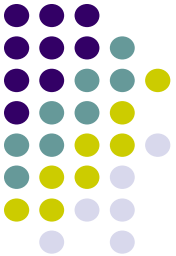
	A	B
64	1	Спиралевидные-
65	1	Трехбусинные (минский тип)-
66	1	Перстенеобразные+
67	1	Кресты-
68	1	Лунницы-
69	1	Амулеты (птицы)-
70	1	Амулеты (предметы быта)-
71	1	Монеты византийские-
72	1	Монетовидные привески-
73	1	Монеты западные-
74	2	Диаметр насыпи (свыше 10 м)-
75	2	Кольцо суглинка-
76	2	Трупоположение в яме-
77	2	Трупосожжение на стороне-
78	2	Трупосожжение на месте-
79	2	Трупоположение на горизонте+

Преобразованные данные (столбец «А» – номер погребения; столбец «В» – наличие или отсутствие признака)

Центр живой истории «Кветунь». Ассоциативные правила

- Большинство правил получились тривиальными. Примеры:
  - ***Если в насыпи были обнаружены угли и пепел***, ***то угли и пепел на уровне захороненного*** (поддержка – 3,08 %, достоверность – 70 %)
  - ***Если при захоронении были обнаружены кресала***, ***то также были обнаружены ножи*** (поддержка – 3,08 %, достоверность – 87,5 %)
  - ***Если при захоронении были обнаружены ножи***, ***то был захоронен мужчина*** (поддержка – 13,10 %, достоверность – 69,77 %)
- Также были найдены нетривиальные правила, которые заказчик признал интересными, например:
  - ***Если была западная ориентировка, имелись перстнеобразные кольца, и было трупоположение в яме***, ***то была захоронена женщина*** (поддержка – 7,05 %, достоверность – 100 %)
  - ***Если при захоронении были обнаружены грибовидные пуговицы***, ***то также был обнаружен бисер*** (поддержка – 5,29 %, достоверность – 92,31 %)
  - ***Если при захоронении были обнаружены ножи***, ***то было парное погребение*** (поддержка – 5,68 %, достоверность – 30,23 %)

Центр живой истории «Кветунь». Деревья решений



- В результате анализа было построено несколько деревьев решений
  - Исполнитель экспериментировал с целевым признаком: следствия подбирались вручную
  - В качестве целевых признаков были выбраны пол погребенного и наличие украшений
- Результаты
  - Если в качестве целевого признака выбирались украшения, то наиболее значимым атрибутом оказывался пол погребенного
  - Если у детей удавалось определить пол, то он преимущественно был женским
  - Наличие драгоценностей позволяло достоверно определить пол человека и наоборот
  - Факт парного захоронения в некоторой степени зависел от наличия ножей и кольца суглинка
  - Факт обнаружения перстнеобразных колец сильно коррелировал с полом погребенного
  - Пол погребенного достаточно хорошо коррелировал с его возрастом

Центр живой истории «Кветунь». Итоги исследования

- Кластеризация не дала полезных результатов, поскольку все данные были категориальными, а алгоритм CLOPE на тот момент не поддерживался
- По итогам обсуждения результатов, полученных разными исполнителями независимо друг от друга, был сделан вывод о том, что результаты применения одной модели целесообразно использовать при построении или уточнении другой
  - Например, деревья решений позволяли выдвинуть некоторые гипотезы, которые далее было целесообразно проверить с помощью ассоциативных правил
  - Либо можно было бы применять деревья решений для интерпретации результатов кластеризации
  - Таким образом, возникла мысль об ансамбле моделей …
- Результаты были переданы заказчику, который обещал дать их интерпретацию и предоставить гипотезы для последующей проверки – эту проверку планировалось выполнить с помощью ансамбля моделей
- Впоследствии заказчик потерял интерес к данному проекту, а без участия специалиста в предметной области дальнейшая работа была невозможна

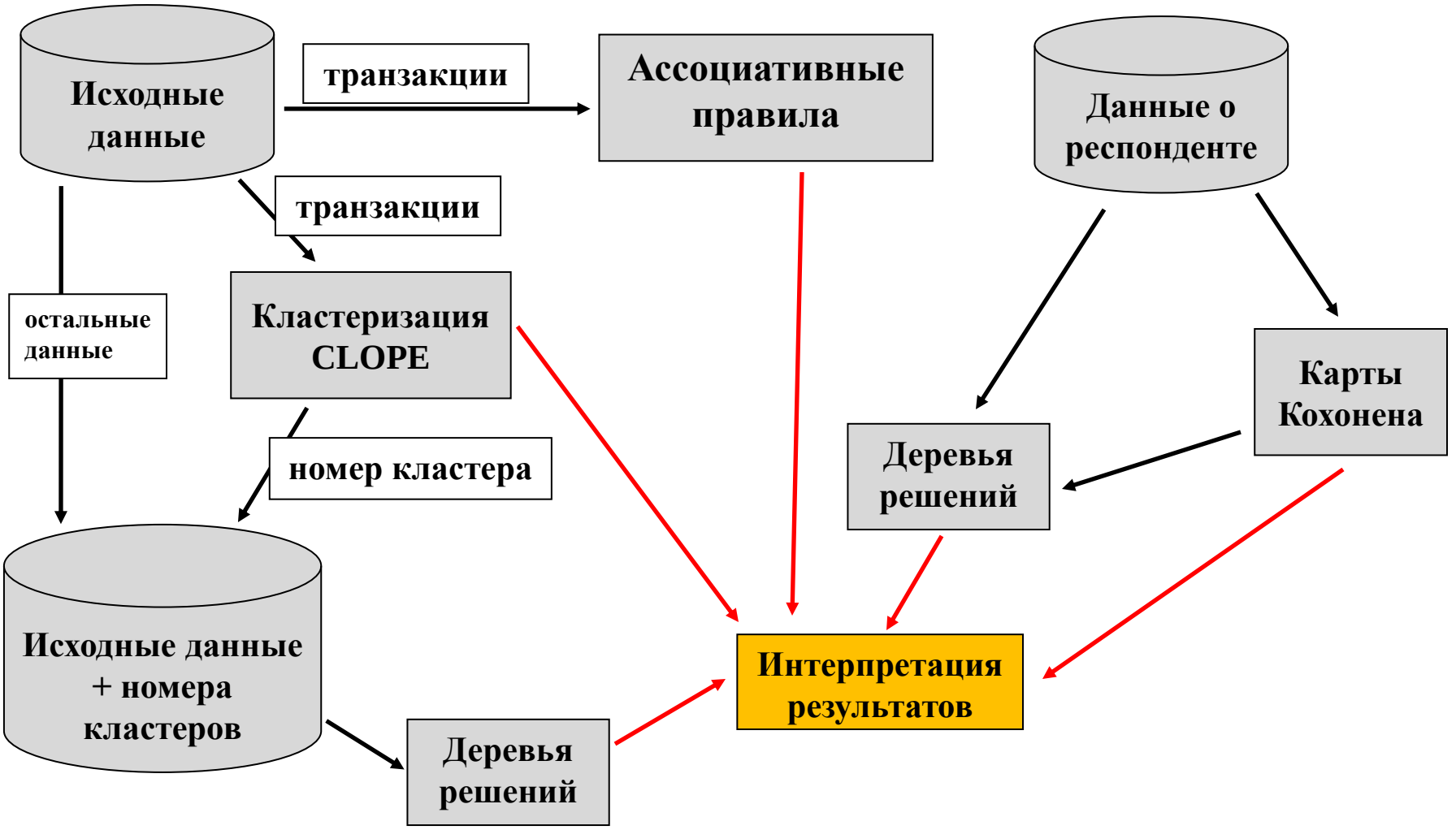
Проект «Мониторинг наркоситуации». Описание проекта

- Исходные положения
  - Анализ и оценка состояния наркоситуации в Брянской области
  - Проводится ежегодно по заказу УФСКН РФ по Брянской области
  - Исполнитель – кафедра социально-гуманитарных дисциплин Брянского филиала Российской академии народного хозяйства и государственной службы
    - Мы приняли участие благодаря имеющимся научным связям с данной кафедрой
  - Основной метод – социологический опрос населения, в том числе пациентов наркологических клиник
    - Единая форма анкеты (в рамках РФ)
    - Рекомендована статистическая обработка результатов анкетирования
  - Социологический опрос проводился в 2013-2016 гг.
    - 2013 г. – 1875 респондентов
    - 2014 г. – 2211 респондентов
    - 2015 г. – 1915 респондентов
    - 2016 г. – проект планируется продолжить, но данные пока недоступны
- Цель проекта – повышение эффективности обработки данных за счет применения методов интеллектуального анализа

Проект «Мониторинг наркоситуации». Структура данных

- Анкета содержит следующие типы вопросов
  - Данные о респонденте (пол, возраст, образование)
  - Жизненные ориентиры респондента (наиболее острые проблемы, жизненные ценности, проведение свободного времени)
  - Вопросы, касающиеся здоровья респондента (оценка здоровья, наличие вредных привычек)
  - Отношение респондента к наркотикам и наркомании
- Структура анкеты в 2013-2015 гг. незначительно менялась. Общее число вопросов – от 37 до 42, среди которых
  - с ответами в категориальной шкале – от 32 до 36 вопросов
  - с ответами в порядковой шкале – от 1 до 2 вопросов
  - с ответами в числовой шкале – от 3 до 4 вопросов
- *В этом проекте был впервые применен алгоритм кластеризации CLOPE и построен ансамбль моделей*

Проект «Мониторинг наркоситуации». Ансамбль моделей



Проект «Мониторинг наркоситуации». Ассоциативные правила (2013 год)

- Всего найдено 3310 правил.
  - максимальная поддержка – 4,86 %
  - максимальная достоверность – 90 %
- Часть правил получились тривиальными, например:
  - ***Если респондент пробовал наркотики, но перестал употреблять***, ***то он попробовал их только один раз***
    (поддержка – 2,13 %, достоверность – 67,8 %)
  - ***Если респондент попробовал наркотики из интереса, любопытства***, ***то он не проходил лечение от наркомании***
    (поддержка – 3,31 %, достоверность – 86,11 %)
- Также были найдены нетривиальные правила. Примеры:
  - ***Если респондента «угощают» наркотиками***, ***то он употребляет их путем курения*** (поддержка – 2,99 %, достоверность – 87,5 %)
  - ***Если респондент не проходил лечение от наркомании***, ***то он употребляет наркотики путем курения*** (поддержка – 3,95 %, достоверность – 80,43 %)

Проект «Мониторинг наркоситуации». Результаты кластеризации CLOPE

Номер и состав кластера	Характеристика представителей
0 (1580 анкет)	• никогда не употребляли наркотики • никогда не использовали ресурсы Интернет, чтобы узнать о наркотиках • отказались бы от употребления наркотиков, если бы им предложили • среди основных жизненных ценностей указали здоровье
1 (284 анкеты)	• пробовали наркотики из любопытства • в кругу общения много тех, кто употребляет наркотики • основной способ употребления наркотиков – курение
2 (10 анкет)	• проходили курс лечения от наркомании и однократно проходили курс реабилитации после лечения • наркотики добывают следующим способом – берут в долг, а деньги на наркотики получают незаконными путями • впервые предложил попробовать наркотики кто-то из членов семьи • способ употребления наркотиков – внутривенное введение

Проект «Мониторинг наркоситуации». Карты Кохонена, деревья решений

Состав кластера (число анкет)	Пол	Возраст	Социально-профессиональное положение	Материальное положение	Номер кластера CLOPE
179	Мужской	18 и более	Разное	Разное	1
349	Мужской	до 18	Разное	Среднее	0 (2)
416	Мужской	23 и более	Служащий, специалист	Разное	0 (2)
108	Женский	от 18 до 35	Разное	Преимущ. среднее	1 (2)
37	Женский	23 и более	Не указано	Среднее либо выше среднего	0
430	Женский	от 23 до 35	Специалист	Среднее	0
335	Женский	от 15 до 23	Разное	Среднее либо ниже среднего	0 (2)
20	Женский	23 и более	Разное	Обеспеченные	0

- В результате построения деревьев решений установлено, что наиболее значимыми атрибутами при разбиении на кластеры оказались:
  - пол респондента
  - возраст респондента
  - номер кластера CLOPE

Проект «Мониторинг наркоситуации». Итоги исследования

- Применение методов ИАД позволило обнаружить закономерности, которые не были обнаружены методами классической статистической обработки данных
- Поскольку число респондентов, признавших себя наркозависимыми, составило всего 10 человек, то не удалось выявить характерные черты представителей данной группы
  - Изначально никто из респондентов не отнес себя к наркозависимым, поэтому эти данные были добавлены искусственно путем опроса пациентов наркологического диспансера
- Сравнение результатов обработки данных, относящихся к разным годам
  - Ассоциативные правила в целом похожие, незначительно изменились их характеристики
  - Результаты кластеризации с помощью карт Кохонена различаются значительно, как по составу кластеров, так и по значимости атрибутов кластеризации

Проект «Вежливая регистратура». Описание проекта

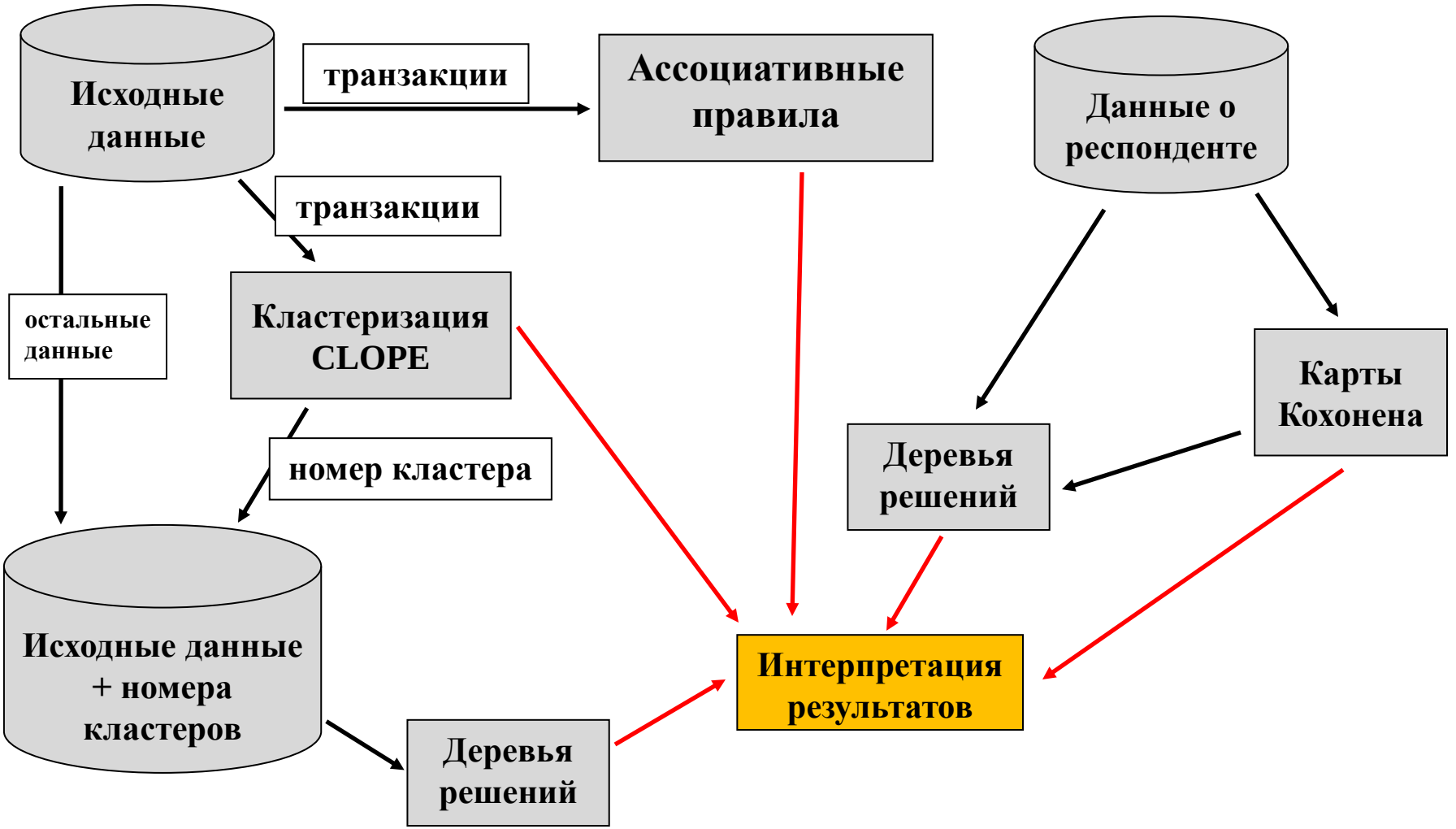
- Цель проекта – оценка качества взаимодействия сотрудников регистратуры поликлиник г. Брянска с пациентами
  - Заказчик – Департамент здравоохранения Брянской области
  - Исполнитель – кафедра социально-гуманитарных дисциплин Брянского филиала Российской академии народного хозяйства и государственной службы
    - Мы приняли участие благодаря имеющимся научным связям с данной кафедрой
- Методы сбора данных
  - социологический опрос пациентов двух поликлиник
  - мониторинг поведения пациентов при общении с регистраторами
    - в частности, подсчет процентного соотношения тех, кто здоровается с работниками регистратуры
- В 2016 году в опросе участвовало 496 респондентов
- **Наша цель** – исследование применимости ансамбля моделей, построенного ранее в рамках проекта «Мониторинг наркоситуации», для анализа данных в другой предметной области, но при этом схожих по структуре и по типу

Проект «Вежливая регистратура». Структура анкеты

- Социологическая анкета содержала вопросы:
  - о респонденте (пол, возраст, образование, социальный статус)
  - об особенностях общения с работниками регистратуры, в частности:
    - здоровается ли респондент с работниками регистратуры при обращении
    - считает ли респондент, что работа в регистратуре тяжелая
    - считает ли респондент, что сотрудники регистратуры вежливы и доброжелательны
  - об особенностях обслуживания в поликлинике, в частности:
    - как долго пришлось ждать в очереди
    - каким образом респондент записался на прием к врачу
  - смотрит ли респондент телепередачи о здоровье

- Общее число вопросов в анкете – 14, из них:
  - с ответами в категориальной шкале – 13
  - с ответами в числовой шкале – 1 (возраст респондента)

Проект «Вежливая регистратура». Ансамбль моделей



Проект «Вежливая регистратура». Ассоциативные правила

- Всего найдено 417 правил
  - максимальная поддержка – 6,65 %
  - максимальная достоверность – 88,9 %
- Примеры найденных правил:
  - ***Если возраст респондента 55-59 лет***, ***то респондент регулярно смотрит телепередачи о здоровье*** (поддержка – 3,83 %, достоверность – 50 %)
  - ***Если респондент – домохозяйка в возрасте 25-34 лет***, ***то день обращения в поликлинику – понедельник*** (поддержка – 1,81%, достоверность – 60 %)
  - ***Если респондент не считает, что сотрудники регистратуры вежливы и доброжелательны***, ***то респондент регулярно смотрит телепередачи о здоровье*** (поддержка – 3,63%, достоверность – 60%)

Проект «Вежливая регистратура». Результаты кластеризации CLOPE

Состав кластера (число анкет)	Характеристика представителей
309 анкет	• возраст до 60 лет • иногда, редко или вообще не смотрят телепередачи о здоровье • всегда или иногда здороваются с сотрудниками регистратуры при обращении • обращались в эту поликлинику и ранее
187 анкет	• возраст свыше 60 лет (пенсионеры) • регулярно смотрят телепередачи о здоровье • никогда не здороваются с сотрудниками регистратуры при обращении • день посещения – вторник • не рекомендовали бы данную поликлинику друзьям и родственникам

Проект «Вежливая регистратура». Карты Кохонена, деревья решений

Состав кластера (число анкет)	Пол	Возраст	Соц. статус	Время ожидания в очереди	Рекомендовали ли бы поликлинику	Здоровается ли	Время поиска карты	Работа в регистратуре тяжелая	Телепередачи о здоровье	Ранее обращались в поликлинику
149	Жен	Свыше 55 лет	Пенсионеры	Не помнят, либо практически не пришлось ждать	Примерно поровну да/нет	Всегда	Менее 3 минут	Согласны	Да или иногда	Да
103	Муж	Разный	Разный	Практически не пришлось ждать	Разные ответы	Всегда	Менее 3 минут	Согласны	Разные ответы	Да
192	Жен	25-44 лет	Разный	Практически не пришлось ждать	Преимущ. да	Всегда	Менее 3 минут либо 5-7 минут	Преимущ. согласны	Разные ответы	Да
52	Муж, Жен	До 55 лет	Разный	Практически не пришлось ждать	Разные ответы	Всегда либо иногда	Менее 3 минут	Преимущ. согласны	Нет или иногда	Нет

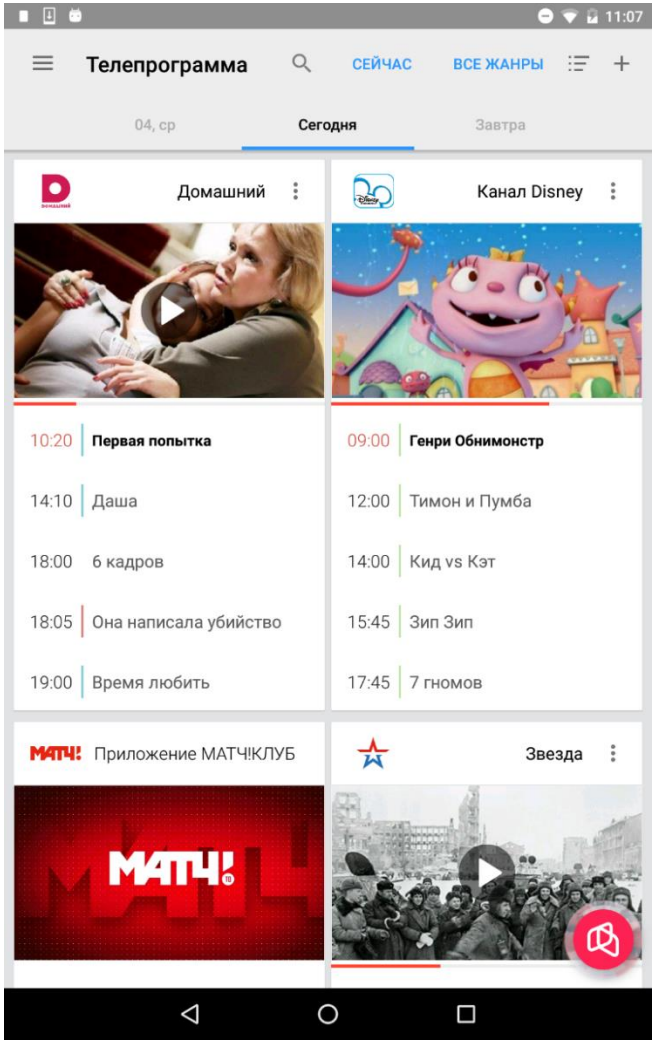
- В результате построения деревьев решений установлено, что наиболее значимыми атрибутами при разбиении на кластеры оказались:
  - пол респондента
  - возраст респондента
  - обращался ли ранее респондент в эту поликлинику

Проект «Вежливая регистратура». Итоги исследования

- Одна из целей заказчика состояла в том, чтобы выявить различия между пациентами, которые демонстрируют вежливое и невежливое поведение при общении с работниками регистратуры
- Кластеризация не привела к получению значимых результатов, поскольку многие респонденты давали «социально приемлемые» ответы на вопросы анкеты
  - По данным мониторинга, здоровались 43 % обращавшихся в регистратуру
  - По данным анкетирования, это показатель составил 78 %
- Наилучшие результаты были получены на основе ассоциативных правил
- Найденные ассоциативные правила хорошо коррелируют с результатами статистической обработки, которую самостоятельно проводил заказчик
  - Но при этом гипотезы формировались заказчиком самостоятельно
  - Также был получен ряд гипотез, неожиданных для заказчика, которые были признаны им интересными для дальнейшего анализа
    - Например, гипотеза о посещении поликлиники разными категориями населения в определенный день

Проект «Сегментация пользователей приложения Tviz». Описание приложения

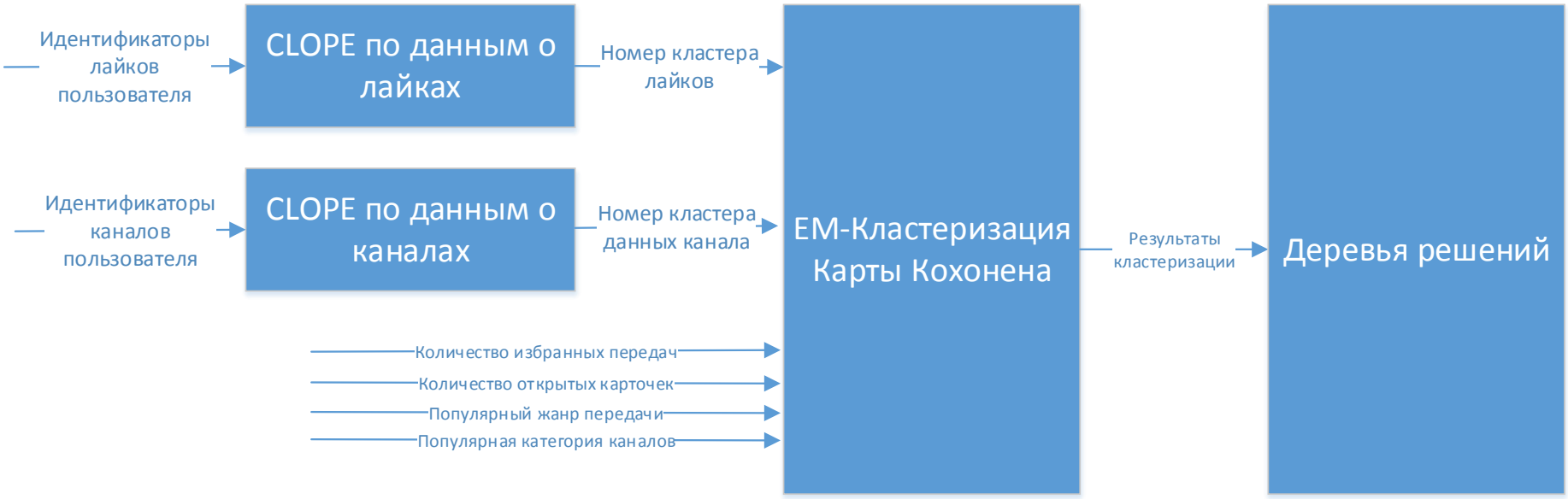
- Функции мобильного приложения Tviz
  - телепрограмма и описание телепередач
  - просмотр каналов онлайн
  - автоматическое распознавание просматриваемого канала
  - возможность сохранения выбранных каналов («Избранное»)
  - установка оповещений на выбранные передачи
  - возможность комментировать передачи и ставить «лайки»
  - открытие карточки передачи
- Используемые технологии
  - технология распознавания эфирного live-контента
  - технология «второго экрана»



Проект «Сегментация пользователей приложения Tviz». Исходная ситуация

- Потребность заказчика – вовлечение как можно большего числа людей, установивших приложение, в активное пользование им, за счет:
  - выдачи всплывающих сообщений о передачах, которые могут быть им интересны
  - стимулирования просмотра карточек передач, формирования расписаний и напоминаний о передачах
- Риск потери механизма воздействия на пользователя, если первые несколько уведомлений окажутся для него не интересными
- В настоящий момент для рассылки уведомлений используется система фильтров и четких правил вида «если … то …»
  - Правила задаются аналитиком вручную, индивидуально для каждого уведомления
  - Ввиду высокой трудоемкости возникает дилемма – уменьшать охват аудитории либо использовать менее персонифицированные правила
    - … или увеличивать количество аналитиков

Проект «Сегментация пользователей приложения Tviz». Ансамбль моделей

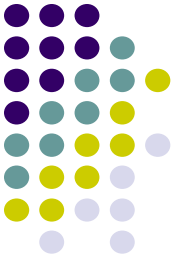


Сегментация пользователей приложения Tviz. Предполагаемая архитектура

Проект «Сегментация пользователей приложения Tviz». Итоги исследования

- Результаты построения и экспериментальной проверки ансамбля моделей позволили сделать вывод о применимости технологии ИАД для решения поставленной задачи
- Особенностью задачи является необходимость использования ансамбля моделей в динамике, поскольку данные об активности пользователей постоянно меняются, вследствие чего будут меняться и рассылаемые сообщения
- Для получения полноценного решения, пригодного к использованию заказчиком, требуется:
  - дальнейшее развитие и совершенствование ансамбля моделей
    - в частности, целесообразно учитывать время суток, в которое пользователь проявляет ту или иную активность, и на основе этого определять содержание уведомлений
  - использование подсистем Deductor Integration Server и Deductor Analytic Server

Содержание



- О кафедре «Информатика и программное обеспечение»
- О курсе «Интеллектуальный анализ данных»
- Проекты, выполненные с использованием платформы Deductor
- **Дальнейшие планы**

Дальнейшие планы и пожелания

- **Обучение и сертификация**
  - Пройти обучение двум преподавателям дисциплины «Интеллектуальный анализ данных»
  - Пройти сертификацию этим преподавателям
  - Рассмотреть возможность сертификации студентов в дистанционном режиме
- **Расширение количества лицензий Deductor Studio Professional**
  - В настоящее время имеется одна лицензия, и для комфортного выполнения курсовых проектов этого недостаточно
- **Учет упоминавшихся ранее особенностей обучения магистрантов *со специализацией в области программирования***
  - Курсовые проекты и магистерские диссертации должны быть направлены не просто на анализ данных, а на *разработку приложений анализа данных*, что возможно только при наличии серверных компонентов Deductor
  - При получении такого опыта обязуемся рассказать об этом на следующей конференции

III межвузовская конференция

Бизнес-аналитика. Использование аналитической платформы Deductor в учебном процессе вуза

Спасибо за внимание!

Подвесовский Александр Георгиевич
заведующий кафедрой, к.т.н., доцент

apodv@tu-bryansk.ru

Лагерев Дмитрий Григорьевич
к.т.н., доцент

lagerevdg@mail.ru

г. Москва, 28 июня 2016 г.