Содержание:

Введение

Сегодня в прикладных социологических исследованиях происходит настоящая революция, связанная с появлением принципиально новых источников данных, прежде всего основанных на так называемой объективной регистрации реального поведения людей. На основе новых информационных технологий различные субъекты (госорганы и бизнесструктуры) собирают огромные массивы данных (Big Data), которые используются в социальной диагностике и прикладных исследованиях. Радикально настроенные аналитики даже предрекают смерть традиционным методам социологических исследований, в большей мере основанным на субъективной информации, получаемой в ходе разного рода опросов. Существует хорошее высказывание, что «за последние годы, когда, стремясь к повышению эффективности и прибыльности бизнеса, при создании БД все стали пользоваться средствами обработки цифровой информации, появился и побочный продукт этой активности – горы собранных данных. И все больше распространяется идея о том, что эти горы полны золота». В прошлом процесс добычи золота в горной промышленности состоял из выбора участка земли и многократного дальнейшего ее просеивания.

Термин Data Mining часто переводится как добыча данных, извлечение информации, раскопка данных, интеллектуальный анализ данных, средства поиска закономерностей, извлечение знаний, анализ шаблонов, раскопка знаний в базах данных.

Методология Data Mining – это мультидисциплинарная область, возникшая и развивающаяся на базе таких наук, как прикладная статистика, распознавание образов, искусственный интеллект, теория баз данных. Возникновение и развитие Data Mining обусловлено различными факторами, и основные среди них

− совершенствование аппаратного и программного обеспечения;

− совершенствование технологий хранения и записи данных;

− накопление большого количества ретроспективных данных;

− совершенствование алгоритмов обработки информации.

Основная часть

Data Mining – это процесс поддержки принятия решений, основанный на поиске в данных скрытых закономерностей (шаблонов информации) т.е. это процесс обнаружения в сырых данных ранее неизвестных, нетривиальных, практически полезных и доступных интерпретаций знаний, необходимых для принятия решений в различных сферах человеческой деятельности.

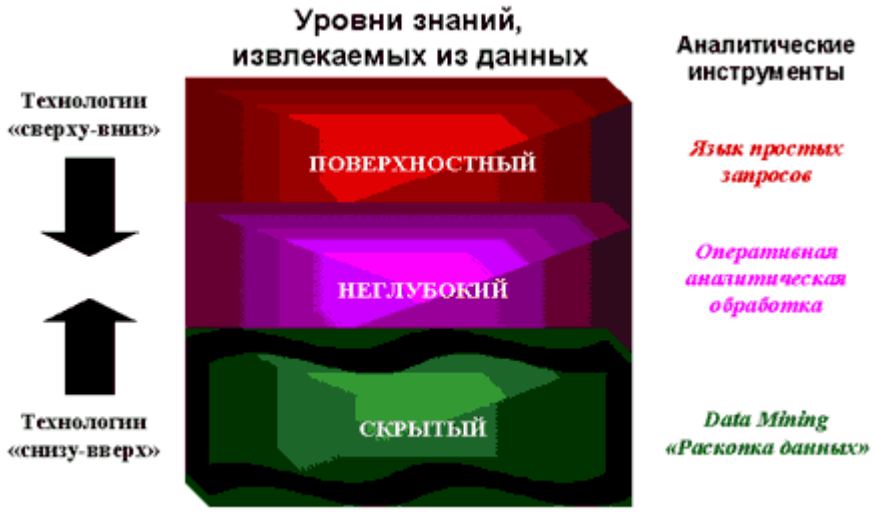

Рисунок 1

Data Mining как мультидисциплинарная область Суть и цель технологии Data Mining заключаются в извлечении из больших объемов данных неочевидных, объективных и полезных на практике закономерностей. Это значит, что найденные закономерности не обнаруживаются стандартными методами обработки информации или экспертным путем, они будут полностью соответствовать действительности и им можно найти практическое применение в социологии. В основу технологии Data Mining положена концепция шаблонов (patterns), которые представляют собой закономерности, свойственные выборкам данных, которые могут быть выражены в форме, понятной человеку. Цель поиска закономерностей – представление данных в виде, отражающем искомые процессы. Построение моделей прогнозирования также является целью поиска таких закономерностей. Чтобы максимально использовать мощность масштабируемых инструментов Data

Mining, в социологических исследованиях необходимо выбрать, очистить и преобразовать данные, иногда интегрировать информацию, добытую из внешних источников, и установить специальную среду для работы Data Mining алгоритмов. Результаты Data Mining в большой мере зависят от уровня подготовки данных, а не от чудесных возможностей некоего алгоритма или набора алгоритмов. Около 75% работы над Data Mining состоит в сборе данных, который совершается еще до того, как запускаются сами инструменты. Прежде чем использовать технологию Data Mining, необходимо тщательно проанализировать ее проблемы, ограничения и критические вопросы, с ней связанные, а также понять, что эта технология не может. Например, Data Mining не может заменить аналитика, она всего лишь дает ему мощный инструмент для облегчения и улучшения его работы. Необходимы тщательный выбор модели и интерпретация обнаруженных зависимостей или шаблонов. Поэтому работа с такими средствами требует тесного сотрудничества между экспертом в предметной области и специалистом по инструментам Data Mining. Построенные модели должны быть грамотно интегрированы в бизнес-процессы для возможности оценки и обновления моделей. В последнее время системы Data Mining поставляются как часть технологии хранилищ данных. В отличие от статистических средства Data Mining теоретически не требуют наличия строго определенного количества ретроспективных данных. Эта особенность может стать причиной обнаружения недостоверных, ложных моделей и, как результат, принятия на их основе неверных решений. Также необходимо контролировать статистическую значимость обнаруженных знаний. Традиционные методы анализа данных (статистические методы) и OLAP в основном ориентированы на проверку заранее сформулированных гипотез (verification-driven data mining) и на грубый разведочный анализ, составляющий основу оперативной аналитической обработки данных (OnLine Analytical Processing, OLAP), в то время как одно из основных положений Data Mining – поиск неочевидных закономерностей. Инструменты Data Mining могут находить такие закономерности самостоятельно и также самостоятельно строить гипотезы о взаимосвязях. Поскольку именно формулировка гипотезы относительно зависимостей является самой сложной задачей, преимущество Data Mining по сравнению с другими методами анализа очевидно. Исследования отмечают, что существуют как успешные решения, использующие Data Mining, так и неудачный опыт применения этой технологии. Области, где применения технологии Data Mining, скорее всего будут успешными, имеют такие особенности: − требуют решений, основанных на знаниях;

− имеют изменяющуюся окружающую среду;

− имеют доступные, достаточные и значимые данные;

− обеспечивают высокие дивиденды от правильных решений.

И все эти характеристики присущи социологии. Таким образом, технология Data Mining постоянно развивается, привлекает к себе все больший интерес как со стороны научного мира, так и со стороны применения достижений технологии в бизнесе, социологических исследованиях. С сентября 2014 г. в Институте общественных наук создана и успешно функционирует кафедра прикладных информационных технологий, состоящая из математиков-информатиков, обладающих большим опытом использования, разработки и внедрения информационных технологий в различные прикладные области. В том числе есть специалисты, способные обучать и передавать знания методологии Data Mining, особенности использования алгоритмов и инструментов программных приложений для обработки и анализа структурированных данных.

Data Mining – это процесс обнаружения в сырых данных

- ранее неизвестных
- нетривиальных
- практически полезных
- и доступных интерпретации знаний,
- необходимых для принятия решений в различных сферах
- человеческой деятельности.

Сфера применения Data Mining ничем не ограничена – она везде, где имеются какие-либо данные. Но в первую очередь методы Data Mining сегодня, мягко говоря, заинтриговали коммерческие предприятия, развертывающие проекты на основе информационных хранилищ данных (Data Warehousing). Опыт многих таких предприятий показывает, что отдача от использования Data Mining может достигать 1000%. Например, известны сообщения об экономическом эффекте, в 10-70 раз превысившем первоначальные затраты от 350 до 750 тыс. дол. Известны сведения о проекте в 20 млн. дол., который окупился всего за 4 месяца. Другой пример – годовая экономия 700 тыс. дол. за счет внедрения Data Mining в сети универсамов в Великобритании.

Data Mining представляют большую ценность для руководителей и аналитиков в их повседневной деятельности. Деловые люди осознали, что с помощью методов Data Mining они могут получить ощутимые преимущества в конкурентной борьбе.

Классы систем Data Mining

Data Mining является мультидисциплинарной областью, возникшей и развивающейся на базе достижений прикладной статистики, распознавания образов, методов искусственного интеллекта, теории баз данных и др. Отсюда обилие методов и алгоритмов, реализованных в различных действующих системах Data Mining. Многие из таких систем интегрируют в себе сразу несколько подходов. Тем не менее, как правило, в каждой системе имеется какая-то ключевая компонента, на которую делается главная ставка. Ниже приводится классификация указанных ключевых компонент на основе работы. Выделенным классам дается краткая характеристика.

Рисунок 2

Статистические пакеты

Последние версии почти всех известных статистических пакетов включают наряду с традиционными статистическими методами также элементы Data Mining. Но основное внимание в них уделяется все же классическим методикам – корреляционному, регрессионному, факторному анализу и другим. Недостатком систем этого класса считают требование к специальной подготовке пользователя. Также отмечают, что мощные современные статистические пакеты являются слишком “тяжеловесными” для массового применения в финансах и бизнесе. К тому же часто эти системы весьма дороги. Есть еще более серьезный

принципиальный недостаток статистических пакетов, ограничивающий их применение в Data Mining. Большинство методов, входящих в состав пакетов опираются на статистическую парадигму, в которой главными фигурантами служат усредненные характеристики выборки. А эти характеристики, как указывалось выше, при исследовании реальных сложных жизненных феноменов часто являются фиктивными величинами. В качестве примеров наиболее мощных и распространенных статистических пакетов можно назвать SAS (компания SAS Institute), SPSS (SPSS), STATGRAPICS (Manugistics), STATISTICA, STADIA и другие.

Наибольшую известность получили исследования поведения пользователей сетевых сообществ и социальных сетей, а также психологическое профилирование личности по цифровым следам на основе модели “Большой пятерки”, быстро взятое на вооружение в политических, маркетинговых и корпоративных проектах. Технологии анализа цифровых следов как маркеров психологических характеристик были разработаны при изучении сообщений блогеров в Twitter и опирались преимущественно на компьютерную лингвистику. Прорывом в этой области стало сопоставление личных страниц в Facebook с ответами их владельцев на стандартизированные психологические опросники, осуществленное в рамках проекта MyPersonality.org, работы Психометрического центра Кембриджского университета и проекта “World Well-Being Project” Центра позитивной психологии Университета Пенсильвании. Начав в 2007 г. с рассылки приглашения 150 друзьям в социальной сети, М. Косинский и Д. Стилвелл за четыре года методом снежного кома собрали более 6 миллионов участников исследования. Уже к 2013 г. им удалось обеспечить прогностическую валидность отметок “like” для шкал пятифакторной модели личности Big Five до 0.43, а при оценке политических убеждений и социально-демографических характеристик коэффициенты составляли от 0.7 до 0.9. Проведенный в 2017 г. мета-анализ исследований, посвященных связи цифровых следов с “Большой пятеркой”, показывает, что их предсказательная сила колеблется от 0.29 для доброжелательности до 0.40 для экстраверсии, то есть не уступает стандартизированным опросникам. Разработаны алгоритмы, позволяющие судить об уровне интеллекта, удовлетворенности жизнью, склонности к самораскрытию и самомониторингу, ценностных ориентациях личности и характеристиках временной перспективы. Для психологического профилирования личности используются не только лингвистические маркеры и сетевой анализ, но и фотографии, размещенные на страницах Instagram и Facebook, видеозаписи, аудиозаписи речи, видеоблоги и датчики движения в смартфонах. Для получения цифровых маркеров психологических характеристик исследователи, как правило, совмещают:

1) компьютерный анализ текста при помощи систем автоматизированного распознавания речи и невербального поведения (чаще всего LIWC);

2) самоотчеты испытуемых и экспертные оценки, в том числе кодирование данных экспертами;

3) машинное обучение.

Растущее количество видеокамер в городах и совершенствование алгоритмов автоматизированного распознавания психологических состояний по поведенческим индикаторам превращают анализ потокового видео в наиболее перспективное направление психометрических исследований . Сетевой и автоматизированный анализ естественного языка позволяет изучать психологические феномены без проведения опросов. В качестве иллюстрации возможностей и ограничений такого подхода можно привести исследования распространения лжи в социальных сетях. Например, команда специалистов из Массачусетского университета проанализировала 120 000 новостей, которыми 4.5 млн. раз поделились 3 млн. пользователей Twitter в 2006–2017 гг. Сообщения были разделены на правдивые и ложные с опорой на 6 независимых фактчекинговых организаций. После выявления и устранения сообщений, сгенерированных ботами, оказалось, что ложные новости отличаются большей новизной, распространяются в 6 быстрее и в 10 раз шире, чем правда. Исследование показало, что в основе распространения дезинформации в сети лежат психологические закономерности массового сознания, а не только злой умысел “фабрик троллей”. Ложные новости были связаны с лингвистическими маркерами удивления, страха и отвращения в перепостах, а правдивые новости – с предвосхищением, досадой, радостью и доверием. При этом авторы исследования признают необходимость проведения интервью и экспериментов для интерпретации полученных данных. Еще одним примером исследования, проведенного без использования опросов, на этот раз в области клинической психологии, стало определение депрессии и склонности к суициду по цифровым следам. Исследователи отобрали три группы интернет-форумов, две из которых включали сообщества людей, проходящих лечение депрессивных расстройств, а также сообщества, для которых характерна идеализация смерти. Третья, контрольная группа была представлена нейтральными форумами, посвященными женской и мужской тематике, карьере и профессиональному развитию. Автоматизированный анализ текстов показал, что на форумах первых двух групп в 3 раза чаще используется абсолютистская лексика, то есть такие слова, как “всегда”, “никогда”, “совсем” и т.п.

Мотивационный профиль студентов – это тест, апробированный и разработанный специально для того, чтобы выявлять факторы мотивации, которые высоко оцениваются студентом, а также те факторы, которым он придает мало значения, как потенциальным источникам удовлетворения выполняемой работой. Он позволит выявить потребности и стремления студента, и, тем самым получить некоторое представление о его мотивационных факторах. В основу теста положено сопоставление значимости ряда мотивационных факторов, представляющих важность с точки зрения руководства вуза. Мотивационные факторы (внутренние факторы) - вызывают чувство удовлетворения работой, это внутренние факторы удовлетворенности, направленные на:

1. успех, достижение (факт достижения цели),
2. продвижение по службе,
3. признание и одобрение результатов работы,
4. высокая степень ответственности за выполняемое дело,
5. возможность творческого и делового роста,
6. сама работа (насколько она интересна, содержательна)

Шкалы мотивационного профиля:

П - поддержание жизнеобеспечения;

К - комфорт;

С - социальный статус;

О - общение;

Д - общая активность;

ДР - творческая активность; ОД - общественная полезность.

Главной проблемой факторного анализа является выделение и интерпретация главных факторов. При отборе компонент исследователь обычно сталкивается с существенными трудностями, так как не существует однозначного критерия выделения факторов, и потому здесь неизбежен субъективизм интерпретаций результатов

Факторный анализ данных относится к линейным методам снижения размерности. Этот метод направлен на нахождение нового координатного пространства , в котором каждая координатная ось является линейной комбинацией исходных

признаков. Популярность данного подхода объясняется тем, что линейные комбинации признаков хорошо интерпретируются – коэффициенты в уравнениях координатных осей трактуются, например, как веса или вклады признаков.

Заключение

Благодаря новым методам в области Data Mining можно будет обнаружить новые связи, закономерности в исходных данных. Найденные новые закономерности, связи помогут проанализировать мотивационный профиль личности. Таким образом, технология Data Mining постоянно развивается, привлекает к себе все больший интерес как со стороны научного мира, так и со стороны применения достижений технологии в бизнесе, социологических исследованиях. С сентября 2014 г. в Институте общественных наук создана и успешно функционирует кафедра прикладных информационных технологий, состоящая из математиков-информатиков, обладающих большим опытом использования, разработки и внедрения информационных технологий в различные прикладные области. В том числе есть специалисты, способные обучать и передавать знания методологии Data Mining, особенности использования алгоритмов и инструментов программных приложений для обработки и анализа структурированных данных.

Список литературы

1. Паклин Н.Б., Орешков В.И. Бизнес-аналитика: от данных к знаниям. - СПб.: Питер.
2. «Что такое Data Mining» https://habr.com/ru/post/95209/
3. «Возможности использования методов в Data Mining» https://cyberleninka.ru/article/n/vozmozhnosti-ispolzovaniya-metodov-data-mining-pri-mediko-laboratornyh-issledovaniyah-dlya-vyyavleniya-zakonomernostey-v-massivah/viewer
4. Рисунок 1 https://blog.iteam.ru/data-mining-intellektualnyj-analiz-dannyh/
5. Рисунок 2 https://mypresentation.ru/presentation/Data-Mining--intellektualnyj-analiz-dannyx
6. «DATA MINING ПРИ РЕШЕНИИ ЗАДАЧ ОБРАБОТКИ СОЦИАЛЬНЫХ ДАННЫХ» http://ecsocman.hse.ru/data/2015/10/23/1250980860/2015_127_12_scientific_life.pdf