

РАЗРАБОТКА ПОДСИСТЕМЫ «АНАЛИТИК» МНОГОФУНКЦИОНАЛЬНОГО ВЕБ ПОРТАЛА ДЛЯ СОЦИАЛЬНО-ПСИХОЛОГИЧЕСКИХ ИССЛЕДОВАНИЙ

С.В Романчуков

Научный руководитель: д.т.н., профессор Берестнева О. Г.,
Томский политехнический университет
inoy@vtomske.ru

Введение

Важную роль в обеспечении благополучия человека и общества играют психологические исследования и службы, в том числе, проводимые на базе высших учебных заведений. Одним из средств проведения подобных тестирований являются сетевые ресурсы, которые также должны предоставлять возможность накопления данных об участниках, результатах и процессе прохождения тестов и обработки имеющейся информации.

Научно-учебной лабораторией информационных технологий в социальных и медицинских исследованиях разработан и поддерживается мультифункциональный тестовый портал «MultiTest», реализующий ряд тестовых методик.

В процессе функционирования данной системы был накоплен обширный материал, нуждающийся в дальнейшей обработке. В соответствующих базах данных хранятся как сырые данные о процессе и результатах тестирования, так и результаты тестов. Имеющийся функционал позволяет оперативно вносить корректировки в тестирующие приложения, эмулировать проведение тестирования, извлекать из базы данных информацию о любом конкретном тестирующемся, но, к сожалению, пока не предоставляет простого в обращении набора инструментов для обработки больших объёмов накопленных результатов. Заявляет о себе необходимость создания подсистемы, дающей аналитикам и психологам возможность работы с информацией без использования навыков программирования (которые психолог или математик иметь не обязан) и, помимо этого, расширяющий возможности обратной связи и взаимодействия психолога собственно с коллективом программистов, разработавших и сопровождающих данный ресурс.

Требования к подсистеме

Аналитическая подсистема разрабатывается с целью расширения функционала веб-портала «MultiTest», что накладывает на неё определённые ограничения.

В базе данных портала уже накоплен значительный объём информации о многолетних результатах исследований, для хранения которой системе применяется СУБД MySQL. Исходя из этого, разрабатываемый модуль должен взаимодействовать с СУБД MySQL, осуществляя выгрузку данных из неё для обработки и сохранения в базу данных её результаты.

Непосредственно для описания тестов и методик используется универсальный формат представления данных XML, что позволяет, например, для создания новых тестов включаемых в портал, применять сторонние программы, в том числе обычные редакторы текста. В этом же формате осуществляется и хранение в базе данных информации о результатах тестирования[1].

Аналитический модуль размещается на одном сервере с уже существующими компонентами портала, представляющем собой HTTP-сервер apache на платформе FreeBSD и реализовываться с использованием php, html и AJAX, аналогично уже размещённым компонентам системы, поддерживать единые с ними стандарты.

Кроме того сервис должен позволять разделение пользователей на различные категории, несущие различные функции и осуществлять разделение прав доступа различных категорий, гарантируя безопасность хранения информации и защиту от несанкционированного доступа.

Роли пользователей

Кратко всех посетителей аналитического модуля можно отнести к одной из трёх групп:

- 1) Администраторы;
- 2) Аналитики;
- 3) Незарегистрированные гости;

В обязанности администраторов входит поддержание базы данных, резервное копирование и восстановление информации. Кроме того на представителей данной группы возлагается ответственность по обеспечению безопасности, назначению и удалению привилегий пользователей и прав доступа.

Группа аналитиков и психологов является, по сути, целевой группой пользователей, для которой и разрабатывается данный модуль. Очевидно, для пользователей данной группы необходимо предоставить доступ к выбранным инструментам анализа данных и сегментам базы данных, относящихся к сфере их компетенции. Аналитик также должен обладать возможностью эмулировать проведение любого теста с помощью программы-имитатора. Также в рамках подсистемы необходимо развернуть элементы, обеспечивающие быструю передачу данных между аналитиками, психологами и администраторами. При этом пользователь-аналитик не должен иметь возможности вносить изменения в таблицы исходной базы данных во избежание коллизий и утраты накопленных данных. Вместо этого предполагается создание вре-

менных записей в специально отведённых для этого таблицах базы данных, приданной к аналитическому модулю.

Методы обработки данных

При решении данной задачи применение многих традиционных статистических методов нерационально и не позволяет добиться желаемого результата. Более эффективным представляется использование методов DataMining. DataMining — это процесс обнаружения в сырых данных ранее неизвестных, нетривиальных, практически полезных и доступных интерпретации знаний.[2] Выделяют пять стандартных типов закономерностей, которые позволяют выявлять методы DataMining: ассоциация, последовательность, классификация, кластеризация и прогнозирование.

В рассматриваемом приложении на первый план выходят задачи классификации и кластеризации данных.

С помощью классификации выявляются признаки, характеризующие группу, к которой принадлежит тот или иной объект. Кластеризация отличается от классификации тем, что сами группы заранее не заданы.

Алгоритмы кластерного анализа отличаются большим разнообразием. На практике в результате испытаний различных алгоритмов исследователями был сформирован ряд рекомендаций по их применению. С учетом чувствительности к зашумлению и способности восстановлению структуры данных наилучшим является алгоритм Уорда, наихудшим — метод ближайшего соседа[3].

Программное обеспечение

Компьютерная обработка данных предполагает математическое преобразование данных с помощью определенных программных средств. Для этого необходимо иметь представление как о математических методах обработки данных, так и о соответствующих программных средствах[4].

Наиболее популярными являются пакеты статистического анализа данных, такие как SPSS, Statistica, STATGRAPHICS. Однако специфика задачи накладывает свои ограничения, делающие использование данного ПО нежелательным:

- 1) Предпочтение программных продуктов, созданных на веб-ориентированных языках (Java, Python и т.д.);
- 2) Ограничения, накладываемые структурой портала;
- 3) Предпочтение программных продуктов с открытым исходным кодом
- 4) Предпочтение свободного ПО.

Всё вышесказанное обращает нас к opensource продуктам в области статистического анализа и

DataMining. К числу наиболее популярных и детально разработанных приложений в данной категории можно отнести Weka, RapidMiner и KNIME (написаны на Java), Rattle (основан на R) и Orange (интегрирован с Python). По сумме их характеристик представляется целесообразным использование в рамках разрабатываемого кабинета аналитика программ WEKA и Orange, разработанных на языке Java и предоставляющих достаточные возможности для проведения анализа статистических данных и визуализации его результатов.[5]

Заключение

В результате выполнения данной работы были разработаны модели поведения потенциальных пользователей, составлен комплекс требований к аналитическому модулю многофункционального веб-портала. На основании этого были подобраны подходящие методы компьютерной реализации аналитических компонентов данного модуля. Рассмотрены открытые программные продукты для data mining. Процесс итоговой программной реализации модуля на данный момент продолжается и далёк от завершения, однако уже предоставил некоторые достаточно значимые результаты. Так, например, был обнаружен один из сбоя в базе данных с результатами экспериментов.

Литература

1. Моисеенко А. В., Берестнева О. Г., Щербakov Д.О. Развитие информационного ресурса для оценки компетентности ИТ-специалистов// Информационные и математические технологии в науке, технике, медицине: труды Всероссийской конференции с международным участием. — Томск, 2012 г. — Т.2 - С. 12-14
2. Григорий Пятецкий-Шапиро, DataMining и перегрузка информацией // Вступительная статья к книге: Анализ данных и процессов / А. А. Барсегян, М. С. Куприянов, И. И. Холод, М. Д. Тесс, С. И. Елизаров. 3-е изд. перераб. и доп. СПб.: БХВ-Петербург, 2009. 512 с. С.13.
3. Мандель И. Д., Черный Л. М. Экспериментальное сравнение алгоритмов кластерного анализа // Автоматика и телемеханика, 1988, № 11. — 252 с.
4. Берестнева О.Г., Е.А. Муратова Компьютерный анализ данных. — Томск: ТПУ, 2010. — 144 с.
5. US Centers for Disease Control and Prevention, Open Source Data Mining Software Evaluation [Электронный ресурс] URL: <http://www.phiresearchlab.org/downloads/OpenSourceDataMining.pdf> доступ свободный