Социальные сети:
новый источник данных о клиентах



Ольга ГОРЧИНСКАЯ,
директор по исследовательским проектам, компания «ФОРС»

Социальные сети служат новым полезным источником дополнительных данных о клиентах любой компании. Однако использовать его не так просто – требуются специальные методики и инструменты. Основанная на технологиях больших данных система ForSMedia, разработанная ФОРС, позволяет автоматически в режиме массовой обработки пополнять профили клиентов и анализировать полученные результаты.

От качества и полноты сведений о клиентах зависят прибыльность и успешное развитие любого бизнеса. До недавнего времени единственным источником такой информации были внутренние системы компании. Вместе с тем в социальных сетях, на форумах, новостных и развлекательных порталах и в блогах содержится много ценного материала, из которого можно «добыть» информацию о предпочтениях и особенностях людей и организаций. Для того чтобы получить дополнительные данные о клиенте, прежде всего необходимо найти или идентифицировать этого клиента в каждом источнике, но далеко не все ресурсы позволяют это сделать. На многих из них пользователи не регистрируются либо оставляют при регистрации недостаточно идентифицирующих данных. Даже там, где данных для идентификации клиента достаточно, может не оказаться полезных дополнительных сведений о нем. Социальные сети оказались наиболее подходящим источником, позволяющим и идентифицировать клиента, и получить дополнительные данные о его предпочтениях, семейном положении, образовании, круге общения и пр.

Алгоритм обогащения профиля клиента

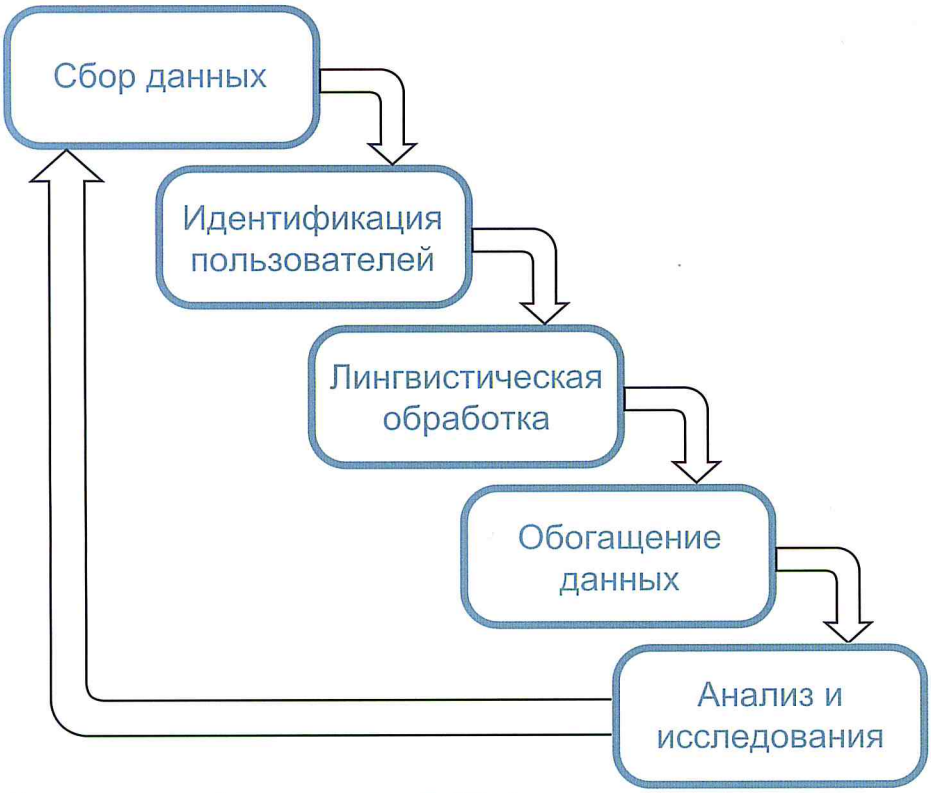
В общем случае задача обогащения профиля клиента состоит в следующем. Компания предоставляет базовые данные (имя, фамилия, дата рождения, город) о своих клиентах, на основе которых необходимо найти дополнительные сведения, например определить круг интересов, социальный статус, область профессиональной деятельности, музыкальные предпочтения и т. д. Для решения этой задачи следует собрать данные из социальных сетей, идентифицировать клиентов, обогатить данные и сформировать единый профиль для каждого клиента.

Платформа ForSMedia, разработанная компанией «ФОРС» (www.fors.ru) на базе технологий больших данных, обеспечивает выполнение всех этих функций. Особенностью решения является возможность массового обогащения данных для большого количества профилей клиентов в автоматическом режиме. При этом система позволяет получить не только данные, указанные в явном виде, но и косвенно представленные сведения, которые можно извлечь из текстов сообщений, групп подписки и действий пользователей в социальных сетях.

ForSMedia основана на технологиях больших данных – инструментарии Hadoop, средствах лингвистической обработки RCO и языке R. Для анализа результатов используется инструментальная среда класса data discovery, обеспечивающая высокий уровень интерактивности работы с данными и помогающая создавать и проверять различные гипотезы при гибкой смене критериев поиска.

Основой для хранения и обработки данных является Apache Hadoop. Выбор Hadoop в указанном случае обоснован следующими факторами: большое количество неструктурированных данных, подлежащих обработке, включая тексты, фотографии, видео; необходимость выполнения сложных вычислений, требующих больших ресурсов; потребность в лингвистической обработке. Дополнительное преимущество – доступная стоимость этого продукта.

Для хранения информации, собранной из социальных сетей, используется нереляционная распределенная база данных



Логическая схема функционирования ForSMedia

Apache HBase, которая работает над файловой системой HDFS (Hadoop Distributed File System) и позволяет хранить большой объем разнотипных данных с высоким уровнем отказоустойчивости. Эта база данных не поддерживает SQL и работает с данными как с массивами байт, что позволяет хранить не только структурированные характеристики пользователей, такие как имя, дата рождения, город, но и картинки, фотографии, видео. Процедуры загрузки и обработки данных основаны на фреймворках Apache Spark и Hadoop MapReduce, обеспечивающих автоматическое распараллеливание обработки большого объема информации.

Общая логическая схема функционирования ForSMedia представлена на рисунке.

Информация о пользователях социальных сетей загружается в HBase и в дальнейшем регулярно обновляется. В целях сокращения временных затрат на начальном этапе загружаются только основные характеристики профилей пользователей, чтобы идентифицировать клиентов. В отдельный набор данных HBase загружается базовая информация о клиентах, необходимая для их идентификации среди всех пользователей социальных сетей. Специализированные процедуры идентификации сопоставляют эту информацию с профилями пользователей и для каждого из них определяют степень схожести с данными клиентов. В качестве основных идентификационных параметров используются имя, фамилия, дата рождения, город. При вычислении степени схожести учитываются возможные неточности в написании имен и фамилий, неполно заданные даты, ошибки правописания в профилях пользователей. Дополнительно предусмотрено использование других контекстных данных для повышения точности идентификации. Связи между профилями пользователей и клиентами, характеризующиеся ненулевой степенью схожести (либо выше пороговой), сохраняются в HBase и используются в дальнейшем для обогащения выявленных профилей.

На следующем этапе обеспечивается лингвистическая обработка тех профилей пользователей, которые имеют степень схожести выше заданного порога. В ForSMedia загружаются стены постов, описания групп подписки и другая неструктурированная текстовая информация. С помощью лингвистических продуктов RCO выделяются тематики стен, анализируются группы подписки, формируются дополнительные характеристики пользователей, например интересы, увлечения. Кроме того, с помощью технологии машинного обучения и средств, встроенных в систему R, производится автоматическое построение классификационных моделей для определения социодемографических параметров пользователей. Эти модели применяются к профилям с неизвестными параметрами, что позволяет уточнить

> ForSMedia основана на технологиях больших данных – инструментарии Hadoop, средствах лингвистической обработки RCO и языке R.

или дополнить информацию, в явном виде указанную в социальных сетях.

В основном хранилище данных ForSMedia сохраняется набор обогащенных профилей пользователей для каждого клиента, на базе которого можно объединять их параметры разными способами. Логика объединения в значительной мере зависит от специфики бизнес-задач, поэтому непосредственно внутри ForSMedia такое объединение не осуществляется, а реализуется в виде дополнительных настроек. Поддерживаются несколько стандартных сценариев – объединение всех значений профилей, превышающих заданный уровень схожести, использование только тех значений, которые встречаются у всех пользователей, и др. Кроме того, предоставляется возможность формировать единый профиль клиента по любой заранее формализованной методике. Таким образом, единые профили клиентов в ForSMedia формируются на уровне выгрузки или экспорта результатов обработки в CRM, аналитическую или другие внешние системы.

Аналитика средствами Big Data Discovery

Наряду с основной функциональностью, обеспечивающей пополнение данных о клиенте, ForSMedia предоставляет дополнительный модуль для анализа обогащенных данных. Этот компонент основан на технологии data discovery, получившей распространение в связи с развитием направления больших данных. Главные ее отличия от классического бизнес-анализа – более высокий уровень интерактивности работы с данными и поддержка формирования и проверки различных гипотез при гибкой смене критериев поиска.

Аналитический модуль ForSMedia основан на новом продукте Oracle Big Data Discovery, предназначенном для работы с данными, хранящимися в Hadoop. На базе этого продукта в ForSMedia реализованы аналитические интерфейсы, ориентированные на маркетологов и предназначенные для детального исследования существующих и потенциальных клиентов при планировании маркетинговых мероприятий, продвижении продуктов и услуг, формировании целевой аудитории и т. д. Помимо бизнес-пользователей с обогащенной информацией о клиентах могут работать аналитики и специалисты по работе с данными, исследующие их на более общем уровне. Для таких специалистов в ForSMedia предусмотрены средства Big Data Discovery, а также дополнительные оригинальные инструменты для структурного анализа взаимосвязей с использованием методов и моделей теории графов. С помощью этих средств можно исследовать, преобразовывать и визуализировать обогащенные профили клиентов для решения различных бизнес-задач.

Платформа ForSMedia максимально использует продукты OpenSource и в минимальной конфигурации может быть установлена на любой аппаратной платформе, отвечающей требованиям конкретной бизнес-задачи.

При наличии высоких требований к производительности можно использовать специальную версию решения, оптимизированную для работы на новейших программно-аппаратных комплексах Oracle класса Engineered Systems. Хранение и обработка данных выполняются в этом случае на машине Oracle Big Data Appliance. Аналитический модуль ForSMedia показывает наивысшую производительность при работе на программно-аппаратном комплексе Oracle Exalytics In-Memory Machine, предназначенном для анализа данных с использованием технологий обработки в оперативной памяти.

Аналитический модуль ForSMedia основан на новом продукте Oracle Big Data Discovery, предназначенном для работы с данными, хранящимися в Hadoop.

Дополнительные возможности для эффективного применения ForSMedia и интеграции ее с системами CRM или хранилищами данных предоставляет инновационная технология Oracle Big Data SQL, которая позволяет работать с хранящимися в Hadoop обогащенными данными о клиентах непосредственно из базы данных Oracle. Реализованный в Oracle Database язык SQL расширен возможностями выполнения запросов в Hadoop без необходимости перегрузки данных в реляционное хранилище. Это означает, что реляционные данные внутренней системы CRM можно легко совмещать с обогащенной информацией о клиентах, используя любые SQL запросы. Платформа ForSMedia может быть развернута на серверах заказчика или поставляться в виде облачного сервиса. ■