Анализ данных социальных сетей

Социальные сети могут стать источником дополнительных данных о клиентах, однако для его использования требуются специализированные инструменты. Открытые технологии из стека Hadoop позволяют строить платформы, способные в режиме массовой обработки извлекать ценную информацию для обогащения профилей клиентов.

Ключевые слова: социальные сети
Keywords: Hadoop, MapReduce, Social networks, Spark

Ольга Горчинская,
Андрей Ривкин

Понятие социальной сети использовалось социологами еще в 20-х годах прошлого века для изучения взаимосвязей между участниками различных сообществ. Психолог и психотерапевт Якоб Морено предложил социограммы, на которых отдельные индивиды представлялись в виде точек, а взаимосвязи между ними — в виде линий. Идею использования аппарата теории графов для изучения взаимоотношений и взаимосвязей между людьми подхватили специалисты в области социологии, психологии, антропологии, политологии, экономики — так сформировалось направление Social Network Analysis, изучающее структурные свойства социальных взаимосвязей, моделируемых в виде графов и сетей. Важным, но весьма трудоемким этапом такого исследования было построение модели на основе различных данных из печатных источников, дополнительных опросов и анкетирования.

Современные социальные сети существенно изменили постановку вопроса — сегодня у исследователей имеется «бесплатный» ресурс для изысканий [1], а стремительное распространение социальных онлайн-сервисов и развитие технологий Больших Данных инициировали интерес к использованию сведений из социальных сетей в различных отраслях. Совместное использование структурных и контентных данных потенциально позволяет применять социальные сети для решения широкого круга бизнес-задач: борьбы с мошенничеством, управления брендом, рекламы товаров и услуг, формирования новых каналов сбыта и др.

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 компаний. Для этого прежде всего необходимо идентифицировать клиента в каждом источнике, что позволяют сделать далеко не все ресурсы — на многих из них люди не регистрируются либо указывают недостаточно идентифицирующих данных. Даже там, где данных для идентификации клиента достаточно, может не оказаться полезных дополнительных сведений о нем. Социальные сети в этом отношении являются наиболее подходящим источником, содержащим и информацию для идентификации клиента, и дополнительные данные о предпочтениях, семейном положении, образовании, круге общения и др.

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

Самый простой способ сбора данных — воспользоваться услугами специализированных компаний, собирающих и постоянно обновляющих данные из множества источников. Главное преимущество здесь — быстрота получения информации, что существенно при больших объемах клиентской базы и использовании различных социальных сетей. Недостаток — платная подписка на обновления данных.

Следующий способ — использовать программные интерфейсы, предоставляемые почти всеми популярными социальными сетями. Для различных сетей API отличаются набором доступных данных, ограничениями на количество запросов и стоимостью доступа к интерфейсам. Например, если с помощью программного интерфейса сети «В контакте» можно получить полную информацию о пользователе, то Facebook предоставляет API, возвращающий практически «нулевые» сведения о пользователе. К недостаткам этого метода относится ограничение на количество одновременных запросов и на количество обращений, которые приложение может делать в единицу времени. Кроме того, необходимо постоянно отслеживать изменения в API и обновлять приложение по сбору данных, причем некоторые социальные сети предоставляют важные данные только на платной основе. Преимуществами метода являются возможность получения данных об одном клиенте в структурированной форме (JSON или XML), а также простота интеграции вызовов API в собственное приложение.

Еще один способ — ручной разбор веб-страниц социальных сетей, а также использование готовых краулеров для сбора данных с последующим разбором. В этом случае имеется доступ ко всем открытым данным и отсутствуют ограничения на скорость их сбора. К недостаткам следует отнести сложность реализации — веб-страница каждой социальной сети уникальна, поэтому каждый раз придется разрабатывать свои правила разбора. Недостатками являются также сложность поддержки и необходимость больших вычислительных ресурсов, правда, этот процесс хорошо распараллеливается.

Идентификация клиента — обнаружение всех профилей, представляющих конкретного клиента в социальных сетях. Исходными данными для поиска могут служить паспортные данные, однако будет полезна и дополнительная информация. Сузить круг и помочь при поиске человека могут такие сведения, как название компании, в которой

он работает, номер телефона, адрес почты, место учебы и список друзей.

Самым простым способом идентификации является поиск по точному совпадению всех известных характеристик клиента, однако необходимо учитывать, что соответствующие характеристики в социальных сетях достоверны лишь до определенной степени — они могут отсутствовать, быть заведомо ложными либо допускать различные варианты написания. Поэтому перед проведением идентификации необходимо произвести очистку и нормализацию данных, а также проверить правильность указанных в профиле параметров — например, город пользователя можно уточнить на основе анализа его подписок, постов и статусов.

Некоторые параметры можно восстановить, анализируя профиль пользователя или его друзей. Например, женщины очень часто не указывают год рождения, тогда как имеется год окончания университета или школы.

Каждая характеристика, используемая при идентификации, имеет некий вес — сумма всех весов при совпадении всех параметров должна быть равна единице. Так, фамилия, имя и пол — одни из самых важных параметров во время идентификации, и если эти данные указаны неверно, то с высокой степенью вероятности идентифицировать этого пользователя не удастся. На втором месте стоят день и месяц рождения. Эти данные поддаются восстановлению, но без них шанс на удачную идентификацию также весьма низкий. Город и год рождения имеют самый низкий вес. Однако именно эти параметры лучше всего поддаются восстановлению на основе других данных.

Кроме данных, которые пользователи сети явно указывают в своих профилях, многое можно узнать, анализируя посты, группы подписки и фотографии. При этом интерес представляют дополнительные факты, которые можно извлечь из этой неструктурированной информации. Например, если в большинстве записей на стене речь идет о впечатлениях о фильмах, то ясно, что пользователь интересуется кино.

Автоматический анализ текстов невозможен без лингвистических технологий. Кроме того, для решения многих задач полезны также статистические методы, технологии машинного обучения и углубленный анализ данных (data mining). Статистические исследования и работа с естественным языком обычно связаны с некоторой неточностью — в статистике речь всегда идет об определенных допущениях, эвристических предположениях, которые не всегда полностью выполняются, а в естественном языке всегда есть вероятность неоднозначного толкования утверждений и выводов. Правильное сочетание лингвистических и статистических подходов повышает качество результата и уровень его достоверности. Для иллюстрации возможного соотношения различных методов при текстовом обогащении данных рассмотрим несколько примеров.

Допустим, нам необходимо узнать, интересуется ли пользователь футболом. Определим, насколько часто в текстах на его стене встречаются соответствующие термины, и при достижении некоторого уровня их появления можно сделать определенные выводы. Для такого метода обогащения необходимо знать терминологию, получить которую можно из словарей или тезаурусов по конкретной предметной области. Кроме того, нужно еще и уметь правильно подсчитать количество употреблений — понимать различные формы одного и того же слова. Таким образом, для данного примера достаточно только лингвистических средств.

Второй пример относится к случаю, когда кроме лингвистической обработки необходимы методы машинного обучения. Предположим, что у пользователя не указана полная дата рождения и требуется определять возрастную группу на основе текстов, которые он пишет. Прежде всего формируется набор текстов пользователей, возраст которых известен. Затем для этого набора с помощью алгоритмов машинного обучения выявляются особенности текстов для каждой возрастной группы и формируется некоторая формальная модель, позволяющая для произвольного текста оценить возраст его автора. Алгоритмы машинного обучения обычно рассчитаны на структурированные данные, поэтому перед их применением тексты заменяются на наборы встречающихся в них слов или на набор тематик, характеризующих эти тексты. Для этого используются лингвистические алгоритмы выделения значимых слов, их нормализации, составления лексического профиля текста, определения тематик и др.

У одного клиента, заданного идентификационными данными, в социальных сетях может существовать много различных пользователей, имеющих достаточно высокий уровень достоверности идентификации. В этом случае возникает задача «объединения» данных нескольких пользователей в единый профиль клиента. Как именно соединять данные, зависит от конкретной задачи — например, для формирования общего списка интересов можно отбирать только интересы, присутствующие у каждого пользователя. Либо можно объединять интересы всех пользователей заданного клиента и использовать расширенный набор интересов.

Обогащение профилей клиентов

Платформа ForSMedia, разработанная компанией «Форс» на базе Hadoop [2], средств лингвистической обработки компании RCO и языка R, реализует все перечисленные методы и подходы. Важной особенностью решения является возможность массового обогащения данных для большого числа профилей клиентов в автоматическом режиме. Платформа может быть развернута на серверах пользователя или поставляться в виде облачного сервиса.

***

Социальные сети служат новым полезным источником дополнительных данных о клиентах любой компании. Использовать этот источник не так просто, и возникающие на этом пути проблемы требуют специализированных технологий и инструментов. Система ForSMedia, основанная на Hadoop и других технологиях Больших Данных позволяет автоматически в режиме массовой обработки обогащать профили клиентов не только данными, в явном виде указанными в социальных сетях, но и сведениями, неявно присутствующими в многочисленных текстах сообщений, постах, группах подписки. ■

ЛИТЕРАТУРА

1. Вэй Тан, Брайан Блейк, Иман Салех. Аналитика Больших Данных и социальные сети // Открытые системы.СУБД. — 2013. — № 8. — С. 37–41. URL: http://www.osp.ru/os/2013/08/13037856 (дата обращения: 18.09.2015).

2. Наталья Дубова. Как устроены Большие Данные // Computerworld Россия. — 2015. — № 16. — С. 13. URL: http://www.osp.ru/cw/2015/16/13046526 (дата обращения: 18.09.2015).

Ольга Горчинская (olga.gorchinskaya@fors.ru) — директор по исследовательским проектам, Андрей Ривкин (andrey.rivkin@fors.ru) — ведущий эксперт по технологиям Больших Данных, компания «ФОРС» (Москва).