УДК 004.622

В. В. Крайнов
Магистрант 2 года обучения,
направление подготовки «Прикладная информатика»
ФГБОУ ВО «Санкт-Петербургский государственный экономический университет»

М. И. Барабанова
Доцент кафедры Информатики
ФГБОУ ВО «Санкт-Петербургский государственный экономический университет»

ОБЗОР СОВРЕМЕННЫХ ПЛАТФОРМ ОБОГАЩЕНИЯ ДАННЫХ

Аннотация. Для современных организаций процесс Data Mining, в том числе в целях маркетинга, является одним из способов достижения конкурентного преимущества. Знания, полученные в результате применения таких методов, позволяют увеличивать лояльность клиентов, проводить успешные промоакции, создавать удобные и комфортные для клиентов интерфейсы. Однако точность методов Data Mining зависит от качества исходных данных, используемых для поиска знаний. Для повышения качества исходных данных, в том числе для целей Data Mining существуют метод обогащения данных. В статье проанализированы основные способы и инструменты для реализации такого метода.

Ключевые слова: data mining, data enrichment, CRM, знания, данные.

V.V. Krainov, M.I. Barabanova

OVERVIEW OF MODERN DATA ENRICHMENT PLATFORMS

Abstract. Data Mining process is one of the ways to achieve a competitive advantage for the modern organizations. The knowledge gained as the result of applying such methods allows to increase customer loyalty, conduct successful promotions, and create user-friendly and comfortable interfaces for the customers. However, the accuracy of Data Mining methods depends on the quality of the source data used to search for knowledge. To improve the quality of the source data there is a data enrichment method. The article analyzes the main approaches and tools for the implementation of that method.

Keywords: data mining, data enrichment, CRM, knowledge, data.

Введение

В современных условиях организациям приходится работать с огромными массивами данных для достижения конкурентного преимущества. По мнению экспертов, в ближайшие десятилетия данные станут одним из основных активов организаций [6, 15]. Так, например, Джек Ма создатель Alibaba Group, заявил, что, по его мнению, данные — это новая нефть. Это подтверждается и статистикой прироста объёмов создаваемых и обрабатываемых данных в мире. Для сохранения конкурентного преимущества и устойчивого развития организациям приходится применять инструменты и методы для работы с данными, например методы Data Mining [11]. Обычно их применяют для целей маркетинга, но возможны и другие сценарии их применения, например, для поиска квалифицированных специалистов и анализа информации о них [1, 3, 10]. В настоящее время, информационные системы, позволяющие применять указанные методы не являются чем-то новым, они широко распространены и являются скорее стандартом де-факто для средних и крупных организаций [4, 5]. Однако для применения методов Data Mining требуется, чтобы исходные данные отвечали заданным параметрам полноты, достоверности и качества. В связи с этим, перед применением методов Data Mining часто требуется очистка и дополнение (обогащение) исходных данных [9]. В рамках данной

статьи производится сравнительный анализ платформ, которые могут выступать в качестве источника новых (или уточнённых) данных для процесса обогащения данных.

Гипотеза

Для современной российской организации процесс обогащения данных для целей Data Mining следует проводить как с применением западных, так и отечественных платформ, содержащих наборы и базы данных, а также сети данных.

Методы

Предметом данного исследования является процесс обогащения данных, предшествующий применению методов Data Mining для задач маркетинга.

Какого-либо конкретного определения у термина «Data Mining» (DM) — нет, поскольку Data Mining (или интеллектуальный анализ данных, добыча данных глубинный анализ данных) — собирательное название, которое используется для наименования целого набора методов поиска и обнаружения в уже имеющихся данных ранее неизвестных, нетривиальных, при этом, практически полезных и доступных интерпретации знаний, которые необходимы для анализа и принятия решений в самых различных предметных областях [7, 8]. Термин предложен Григорием Пятецким-Шапиро в 1989 году.

В общем виде схема процесса Data Mining выглядит следующим образом (рис. 1).

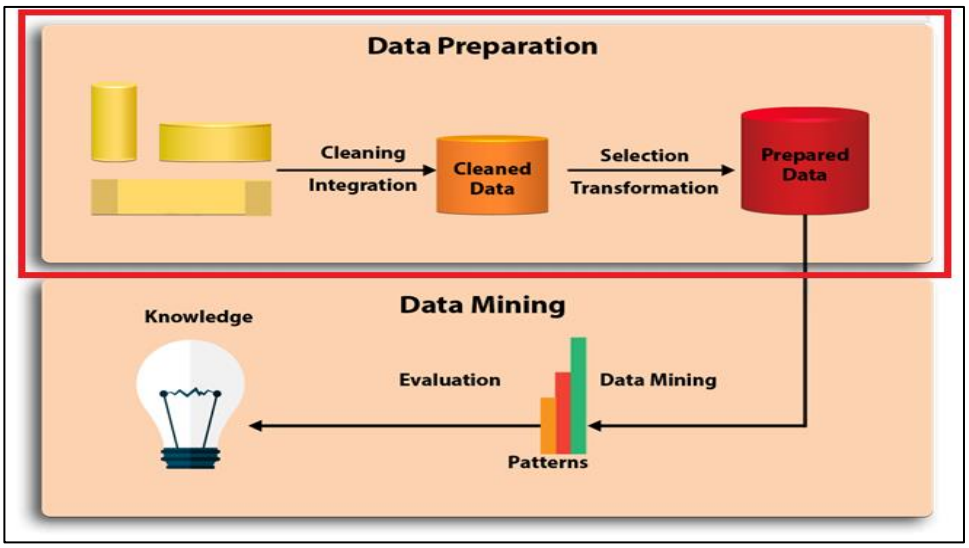

Рисунок 1. Схема процесса Data Mining в общем виде

Объектом исследования является среднестатистическая организация Российской Федерации, использующая CRM (Customer Relationship Management) систему.

Система управления взаимоотношениями с клиентами (CRM, CRM-система) — информационная система корпоративного уровня, основной функционал которой предназначен для автоматизации создания стратегий взаимодействия с заказчиками (клиентами), в том числе для увеличения объёмов продаж и улучшения клиентского сервиса путём сохранения и накопления информации о клиентах, а также истории взаимоотношений с ними с целью дальнейшего анализа [12, 16, 17].

Именно для такого анализа и могут применяться методы Data Mining.

В рамках данной статьи фокус сделан на первом этапе Data Mining (рис 1, Подготовка данных, Data Preparation).

Анализ предметной области показывает, что в большинстве организаций, использующих CRM-системы процесс заполнения данных в ней, выглядит следующим образом (рис. 2).

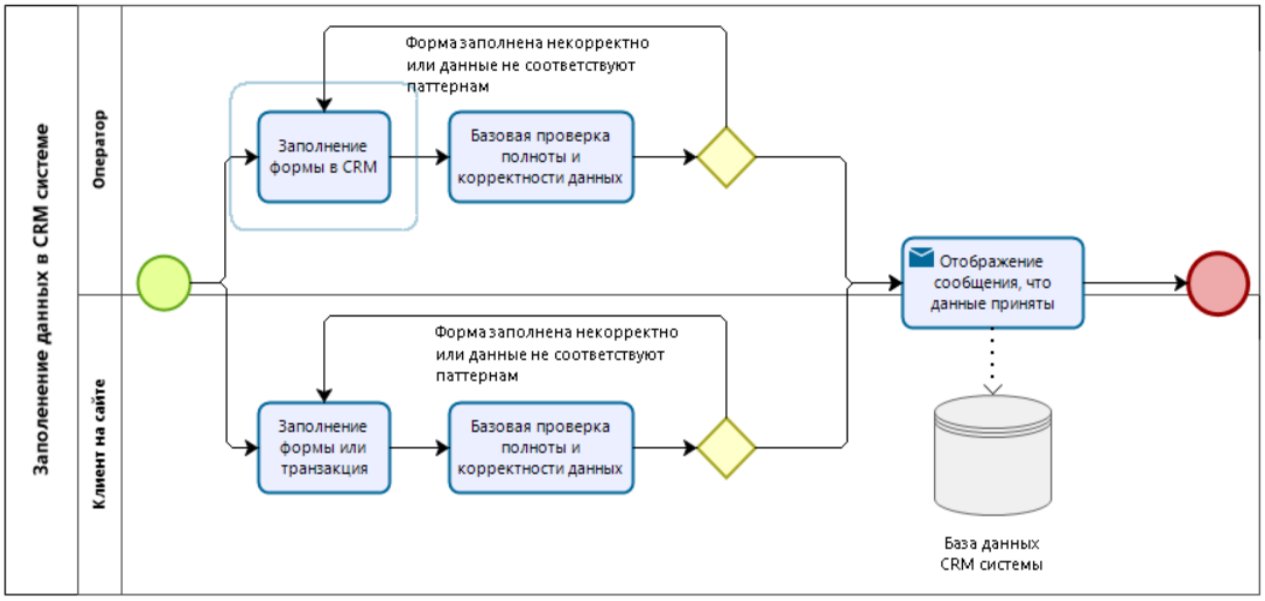

Рисунок 2. Типовой процесс внесения данных в CRM систему

Как в случае заполнения формы клиентом на сайте, так и в случае в заполнения формы оператором CRM системы, очевидно, что могут применяться стандартные методы, обеспечивающие уникальность данных (например, проверка на совпадения по ключевым полям), их полноту (например, проверка, что имя больше 1 символа, номер телефона состоит из 10 цифр и т.п.), соответствия заданным паттернам (например, что адрес E-mail содержит символ «@», а номер телефона начинается с «+7»). Хорошей практикой считается подключение к таким формам различных справочников для упрощения ввода стандартных реквизитов, таких как: страна, город и т.д.. Такой подход позволяет значительно снизить вероятность ошибок ввода по соответствующим полям.

Предметом исследования, как и было указано ранее является процесс обогащения данных для целей Data Mining. В рассматриваемой предметной области результатами такого процесса должен быть массив данных с дополнительными реквизитами и сведениями о людях и организациях, данные о которых хранятся в CRM системе, кроме того, при вводе данных в CRM-систему оператором в результате применения методов обогащения данных должны отображаться подсказки и возможные дополнительные реквизиты для добавления в новую или существующую запись.

Технически процесс обогащения данных может быть реализован по-разному [13, 14].

В общем случае, для сбора данных с различных сайтов (порталов) в автоматизированном режиме могут применяться веб-пауки и инструменты для веб-скрейпинга.

Веб-скрейпинг — это технология получения веб-данных путем извлечения их со страниц веб-ресурсов. Веб-скрейпинг может проводиться вручную, однако обычно термин относится к автоматизированным процессам, реализованным с помощью кода, который выполняет GET-запросы на целевой сайт.

Веб-скрейпинг используется для синтаксического преобразования веб-страниц в более удобные для анализа данных формы. Веб-страницы создаются с использованием текстовых языков разметки (HTML и XHTML) и содержат множество полезных данных в коде. Однако большинство веб-ресурсов предназначены для конечных пользователей, а не для удобства автоматического использования, поэтому существуют технологии, которые «очищают» веб-контент.

Второй вариант реализации процесса обогащения данных — работа с платформами, содержащими наборы и базы данных, а также сети данных через API. API (application programming interface, программный интерфейс приложения) — описание способов (набор классов, процедур, функций, структур или констант), которыми одна компьютерная система может взаимодействовать с другой [2]. Ввиду простоты и надёжности такого подхода, он является приоритетным для применения, однако, это требует, чтобы разработчики платформ предусмотрели соответствующий интерфейс, а также документацию для него. В случае, если такого интерфейса нет, он не открытый или отсутствует документация по нему следует использовать первый метод.

Мы провели сравнительный анализ функций платформ, содержащих наборы, базы данных и сети знаний. Безусловно, прямое сравнение приведённых ниже сервисов не имеет никакого смысла, поскольку они являются многофункциональными и направлены на решение различных задач. Мы собрали в таблице 1 те сервисы, которые позволяют обогащать указанные выше данные: реквизиты людей и организаций. Среди глобальных сервисов мы считаем целесообразным выделить представленные в таблице 1.

FullContact

Компания FullContact предоставляет сервис Enrich, позволяющий получать дополнительную информацию о людях или организациях, используя имеющиеся контактные данные. Недостатками данного сервиса являются:

- Размер базы данных. База данных FullContact содержит всего 1 миллиард профилей людей и 22 миллиона профилей организаций, в то время как база данных Diffbot KG включает в себя больше 10 миллиардов уникальных объектов
- Малое количество полей, по которым можно производить поиск. Согласно документации, информацию о человеке можно искать только по комбинации следующих полей:
    - E-mail
    - Аккаунт в Twitter
    - Номер телефона

Организации можно искать только по доменному имени. Таким образом, невозможно производить поиск организаций по названию или людей по полному имени.

- Отсутствие возможности извлечения наиболее актуальной по времени информации по запросу пользователя.

LeadIQ

Компания LeadIQ предоставляет сервис LeadIQ Prospector, API которого позволяет производить поиск информации о людях и организациях. Недостатками данного сервиса являются:

- Отсутствие возможности извлечения наиболее актуальной по времени информации по запросу пользователя.
- Размер базы данных. База данных LeadIQ содержит 10 миллионов контактов.

Таблица 1. Анализ функционала платформ (сервисов) для обогащения данных

	FullContact	Diffbot KG	LeadIQ	Clearbit	Synthio	Openprice	Kaggle
Основные функции	Сервис Enrich: Поиск информации о людях по комбинациям следующих полей: E-mail, аккаунт в Twitter, номер телефона. Поиск информации об организациях по доменному имени	Сервис The Knowledge Graph of the Public Web: Поиск данных об организациях и по всему миру по различным реквизитам, таким как: индустрия, адрес (страна, город, регион), названию, адресу в Twitter. О людях: по названию университета, по месту работы, по имени, по навыкам	Сервис LeadIQ Prospector, используя LeadIQ Search API: Поиск по людей по: ФИО и компании, ФИО и домену, профилю в LinkedIn, E-mail, или типу телефона (рабочий, личный). Поиск организаций по: E-mail, E-mail и телефону, по названию, домену	Сервис Enrichment: информация о персоне или компании по e-mail или домену. Возможность поиска данных по любым атрибутам и гарантия высокой актуальности данных	Сервис Vertify Enrich: Возможность обогащения данных B2B с по более чем 50 атрибутам	Сервис Data Enrichment for Marketing & Sales: Доступ к рынку данных Openprise сторонних поставщиков данных Доступ к каталогу открытых данных Openprise. В экосистеме сервисов решения для: очищения, дедупликации, унификации данных	Хостинг наборов данных по различным предметным областям
Объёмы данных	1 миллиард профилей людей и 22 миллиона профилей организаций.	более 10 миллиардов записей о людях и организациях по всему миру.	10 миллионов контактов.	250 общедоступных и закрытых источников данных, включающих в себя миллионы записей.	Платформа управления контактными данными Synthio обеспечивает единое целое для всех ваших потребностей в контактных данных B2B: более 160 млн. контактов, более 45 млн адресов электронной почты, более 25 стран, более 2 млн. обновлений контактов ежемесячно.	Большое количество источников данных.	Зависят от конкретного выбранного набора данных
Наличие API	Есть	Есть	Есть	Есть	Есть	Есть	Есть
Цена	По запросу	От 900 USD в месяц или по запросу	От 10080 USD в год	По запросу. Цена зависит от размера базы данных, трафика	От 599 USD в месяц	48000 USD в год при работе с количеством записей до 500000	Бесплатно
Официальный сайт	https://www.fullcontact.com/	https://www.diffbot.com/products/knowledge-graph/	https://leadiq.com/	https://clearbit.com	https://synthio.com/; https://vertify.com/	https://www.openprisetech.com	https://www.kaggle.com/

Synthio

Компания Synthio, основанная в 2011 году создала первую в мире платформу, осуществляющую очистку, стандартизацию и синтез контактных данных для того, чтобы сделать их доступными для немедленного использования. Данная компания предоставляет сервис Vertify Enrich, позволяющий пользователям дополнять существующую информацию о людях и организациях. Недостатками данного сервиса являются:

- Курируемость. Благодаря необходимости ручной валидации/наполнению базы данных повышаются расходы на содержание сервиса, а также уменьшается актуальность хранимых данных.
- Отсутствие возможности извлечения наиболее актуальной по времени информации по запросу пользователя.
- Размер базы данных. База данных Synthio содержит 160 миллионов контактов, в то время как база данных Diffbot KG включает в себя больше 10 миллиардов уникальных объектов.

Рассмотренные платформы позволяют получать готовые данные из заранее проиндексированных наборов данных.

Среди рассмотренных платформ, наиболее функциональной и богатой данными выглядит платформа Diffbot KG, при этом, она же является достаточно дорогой, а набор реквизитов, применимых для поиска не является самым большим.

Однако, очевидно, что на практике ни одна из проанализированных платформ (большинство из которых являются западными) не позволяет получать достаточные данные о российских организациях или персонах.

Выборочный запрос с данными об организациях малого и среднего бизнеса показал, что в самая богатая данными платформа (Diffbot KG) позволяет найти лишь около 10—15% организаций. Информацию о людях рассмотренные платформы чаще всего ищут либо в социальной сети LinkedIn (заблокирована в Российской Федерации), Twitter или Facebook, что значительно сужает объём доступной информации.

В связи с этим был произведён анализ отечественных платформ, которые могут выступать источником данных для процесса обогащения данных.

По организациям Российской Федерации основные реквизиты можно получить из системы СПАРК от Интерфакс или из системы Глобас от Credinform. В обзор, также, включена система КАД Арбитр, поскольку она может быть полезна как источник данных о судебных делах организаций или людей (например, если требуется выяснить являлась ли та или иная организация или человек ответчиками по судебным делам). Описание указанных систем представлено в таблице 2.

Таблица 2. Отечественные платформы для обогащения данных

	СПАРК	Глобас	КАД Арбитр
Основные функции	Поиск информации об организации по различным реквизитам, скоринг организаций.	Поиск информации об организации по различным реквизитам, скоринг организаций.	Информация о судебных делах по организациям и людям
Наличие API	Есть	Есть	Есть
Объёмы данных	146 млн. человек, 3,75 млн. юридических лиц	28 миллионов организаций и ИП России, более 400 млн. о компаниях по всему миру	Более 28000000 судебных дел в картотеке по состоянию на сентябрь 2020
Цена	По запросу	Тарифы могут быть персонализированы, по запросу	Бесплатно

С поиском данных о людях по отечественным источникам задача усложняется и её решение возможно с применением методов веб-скрейпинга по основным социальным

сетям: vk.com, ok.ru в рамках правил этих сервисов, а также действующего законодательства Российской Федерации.

Результаты и обсуждение

Результаты сравнительного анализа платформ для обогащения данных показывают, что не существует единой платформы, с которой возможно получить все необходимые данные. Для каждого конкретного проекта требуется выбор той платформы, которая содержит наибольшее количество данных требуемого типа. Для обогащения данных о западных организациях целесообразно применять одну из платформ, представленных в таблице 1, по нашему мнению, наилучшим выбором среди них является Diffbot KG. Для обогащения данных об отечественных организациях, следует применять одну из платформ, представленных в таблице 2. КАД Арбитр следует применять, когда требуются данные об истории судебных дел. Для обогащения данных о людях в Российской Федерации целесообразнее использовать методы веб-скрейпинга по отечественным социальными сетям и порталам. Безусловно, необходимо помнить, что все эти процессы должны осуществляться в соответствии с законодательством Российской Федерации.

Список литературы

1. Барабанова, М. И. Кадровое обеспечение цифровых предприятий в аспекте развития цифровой экономики / М. И. Барабанова, Е. А. Рыбакова // Технологическая перспектива в рамках Евразийского пространства: новые рынки и точки экономического роста: Труды 5-ой Международной научной конференции, Санкт-Петербург, 07–08 ноября 2019 года. – Санкт-Петербург: Центр научно-производственных технологий "Астерион", 2019. – С. 214-220.
2. Барабанова, М. И. Открытые системы и сети. Комплексная безопасность в системах и сетях современного предприятия : Учебник / М. И. Барабанова, А. В. Саитов, В. И. Кияев. – Санкт-Петербург : Санкт-Петербургский государственный экономический университет, 2019. – 496 с. – ISBN 9785731044509.
3. Барабанова, М. И. Проблемы формирования исследовательских компетенций в подготовке рабочих кадров и специалистов среднего профессионального образования / М. И. Барабанова // Профессиональное образование в современном мире: традиции и инновации : МЕЖДУНАРОДНАЯ ПРАКТИЧЕСКАЯ КОНФЕРЕНЦИЯ, Выборг, 19–20 февраля 2019 года. – Выборг: Государственный институт экономики, финансов, права и технологий, 2019. – С. 19-24.
4. Газуль, С. М. Особенности построения сервиса журналирования системных событий на основе blockchain-подобной платформы и технологий виртуализации / С. М. Газуль, В. И. Кияев // Технологическая перспектива в рамках Евразийского пространства: новые рынки и точки экономического роста : Труды 5-ой Международной научной конференции, Санкт-Петербург, 07–08 ноября 2019 года. – Санкт-Петербург: Центр научно-производственных технологий "Астерион", 2019. – С. 232-237.
5. Газуль, С. М. Формирование mashup-порталов с использованием контейнерной виртуализации / С. М. Газуль, В. И. Кияев // Конвергенция цифровых и материальных миров: экономика, технологии, образование : Сборник научных статей международной научной конференции, Санкт-Петербург, 21–22 июня 2018 года / Под редакцией В.В. Трофимова, В.Ф. Минакова. – Санкт-Петербург: Санкт-Петербургский государственный экономический университет, 2018. – С. 99-105.
6. Ильина О.П., Барабанова М.И. Методология гибкой цифровой трансформации предприятия // в сборнике "Технологическая перспектива в рамках евразийского пространства: новые рынки и точки экономического роста". Санкт-Петербург, 07-08 ноября 2019 г. - с. 223-232.
7. Ильина, О. П. Моделирование ценности сервисов информационных технологий для бизнеса / О. П. Ильина, М. И. Барабанова // Журнал правовых и экономических исследований. – 2019. – № 4. – С. 172-176. – DOI 10.26163/GIEF.2019.16.58.027.
8. Полякова А.Г., Шеханова А.С. Потенциал и особенности использования технологии Big Data //Вестник современных исследований. -2018. -№ 2.1 (17). -С. 99-101.
9. Филяк П.Ю. Сети, большие данные (big data), интеллектуальный анализ данных (data mining) и обеспечение безопасности//Информация и безопасность. 2017. Т. 20. № 4. С. 522-527.
10. Система формирования исследовательских компетенций и технологических заделов в научной и образовательной деятельности / В. В. Трофимов, Л. А. Трофимова, В. Ф. Минаков [и др.]. – Санкт-

Петербург : Санкт-Петербургский государственный экономический университет, 2018. – 199 с. – ISBN 9785731044240.

11. Трофимов В.В., Барабанова М.И., Ильина О.П., Макарчук Т.А. [и др.]. Информационно-образовательная среда экономического вуза. СПб.: Изд-во СПбГЭУ, 2018.
12. Трофимов В.В., Минаков В.Ф. Цифровая конвергенция в экономике / [В.В. Трофимов и др.]; под ред. В.В. Трофимова, В.Ф. Минакова. - СПб.: Изд-во СПбГЭУ, 2019. - 150 с.
13. Трофимов В.В., Трофимова Л.А., Минаков В.Ф., Барабанова М.И., Макарчук Т.А., Лобанов О.С. Единое информационное пространство взаимодействия субъектов научной и инновационной деятельности. СПб.: Изд-во СПбГЭУ, 2017.
14. Forming Ontologies and Dynamically Configurable Infrastructures at the Stage of Transition to Digital Economy Based on Logistics / S. Barykin, S. Gazul, V. I. Kiyaev [et al.] // Advances in Intelligent Systems and Computing (см. в книгах). – 2020. – Vol. 1116. – P. 844-852.
15. Silkina, G., Barabanova, M., Gazul, S., Kiyaev, V., 2019. Using Blockchain-based Approach for Building the System Events Logging Service. In: Journal of Physics: Conference Series. International Scientific Conference "Conference on Applied Physics, Information Technologies and Engineering - APITECH-2019", pp. 33-75
16. Creating Competitive Advantage from Big Data in Retail./Consumer Marketing Analytics Center. McKinsey & Company. -2012. -. URL: http://www.mckinsey.com/client_service/retail/expertise/~/media/mckinsey/dotcom/client_service/retail/articles/cmac_creating_competitive_advantage_from_big_data (дата обращения: 25.02.2018 г.).
17. Quality Management Of E-Learning In Information Technology Management Training / T. A. Makarchuk, V. F. Minakov, V. V. Trofimov [et al.] // The European Proceedings of Social & Behavioural Sciences EpSBS, Irkutsk, 26–28 апреля 2018 года. – Irkutsk: Future Academy, 2019. – P. 742-748. – DOI 10.15405/epsbs.2018.12.91.