3.2. ИСПОЛЬЗОВАНИЕ ОНТОЛОГИЙ В КОРПОРАТИВНЫХ АВТОМАТИЗИРОВАННЫХ СИСТЕМАХ

Горшков С.В., директор ООО «ТриниДата»

Любая крупная организация накапливает огромное количество информации в информационных системах, автоматизирующих ее деятельность. Как правило, данные в разных системах структурированы без единой концепции, слабо связаны, плохо поддаются поиску и использованию вне заранее предусмотренных рутинных бизнес-процессов. В результате накопленная информация становится неэффективно используемым активом, на поддержку которого тратятся значительные средства, но который приносит довольно ограниченную выгоду. Потенциал использования этой информации огромен, но для его раскрытия необходимо не часто встречающееся на практике сочетание управленческой воли и мотивации, видения ИТ-архитектуры и адекватных технологических средств с возможностью воплотить в жизнь амбициозный проект.

Среди средств повышения эффективности использования накопленной информации выделяются семантические технологии, или технологии «семантической паутины» (semantic web). Они представляют собой набор стандартов и программных средств, позволяющих создать машинно-читаемое представление концептуальной модели определенной предметной области (онтологию), использовать его для представления конкретных данных в целях анализа и интеграции, а также автоматизировать получение логических выводов.

Средства обработки концептуализированных знаний и автоматизации логического вывода известны с 1970-х годов, когда они применялись в экспертных системах. Современное поколение семантических стандартов разрабатывается с начала 2000-х годов, то есть тоже не является технологической новинкой. Несмотря на это, потенциал использования этих технологий начинает раскрываться только в последние годы. Это связано, на наш взгляд, с несколькими факторами. Только сейчас складывается сочетание факторов, которые делают применение семантических технологий методически необходимым и технически возможным: объем и сложность структуры накопленной информации, зрелость задач по ее обработке, вычислительные мощности.

Создание и применение онтологий требует серьезной аналитической работы, владения непростыми методиками моделирования. Этот фактор не позволяет онтологиям стать «модными» и идет вразрез с тенденциями современных ИТ, ориентированными на минимизацию аналитической деятельности в пользу эвристических подходов. Получение полезного эффекта от использования онтологий требует постановки стратегических задач и планирования их реализации хотя бы в среднесрочной перспективе – в отличие от превалирующих сейчас управленческих и технологических подходов, требующих «быстрых побед» и адаптации процесса создания ИТ-систем под постоянно меняющиеся краткосрочные цели. Тем не менее, использование онтологий обеспечивает фундаментальные преимущества, благодаря которым в ряде сфер применения им просто не существует альтернативы. Рассмотрим в качестве примера системы поддержки принятия решений.

Сфера применения продуктов этого класса очень широка и простирается от корпоративного управления до ситуационных центров, от медицины до промышленной безопасности. Сегодня популярной идеей является использование в таких системах технологий машинного обучения и нейросетей, которые способны имитировать принятие решений человеком в определенных классах ситуаций на основании «наблюдения» за решениями, которые принимали в аналогичных ситуациях реальные люди. Но решения, предложенные такими системами, не могут быть верифицированы. Подобные системы не анализируют суть ситуаций и не воспроизводят ту логику, которой явно или неявно руководствуется в принятии решений человек, а всего лишь выдают достоверный с какой-то вероятностью результат, полученный в результате вычислений по математической модели, нейтральной к предметной области и смыслу обрабатываемой информации. Это качество делает указанные технологии непригодными к применению в тех ситуациях, когда существуют нормативные требования к обоснованности и верифицируемости принимаемых решений.

Автоматизация получения логических выводов на основе семантических технологий свободна от такого недостатка. Каждое решение, выданное системой, может быть логически проверено, декомпозировано до набора фактов и аксиом, каждый из которых кем-то подтвержден или закреплен нормативно. Системы на основе онтологий воспроизводят формальные цепочки рассуждений, приводящие к тому или иному решению.

Однако, на практике достаточно сложно сформулировать тысячи логических правил, которыми руководствуются люди в принятии решений даже в ограниченных диапазонах ситуаций. Здесь на помощь приходит принцип машинного обучения, при помощи которого можно реализовать автоматизированное составление формальных правил принятия решений на основании массива данных о решениях, фактически принятых людьми в разных ситуациях, оказавшихся верными или неверными. Каждое правило, сформулированное подобным алгоритмом, предъявляется эксперту для логической проверки и после верификации начинает использоваться в работе системы. Такой способ является, на наш взгляд, одним

из существенных элементов, необходимых для широкомасштабного и успешного применения семантических технологий, поэтому он находится в фокусе разработок нашего коллектива.

Другая технологическая проблема, которая иногда рассматривается как препятствие к использованию семантических технологий, состоит в относительно небольшой емкости и производительности графовых баз данных, которые используются для хранения онтологий (в сравнении с реляционными базами или noSQL-решениями). Решением здесь является синтез онтологий и технологий «больших данных» (big data) и уже упомянутых noSQL, in-memory хранилищ. В сочетании с принципом логической витрины данных этот синтез позволяет обеспечить хранение практически не ограниченных объемов информации, с применением онтологий для их структурирования и обработки. Проект Optique (http://optique-project.eu/) является примером успешной реализации этого принципа в мировой практике.

Обратимся к другому сценарию использования онтологий в корпоративных информационных системах. Достаточно распространенной является задача обмена информацией между различными организациями: в рамках холдинга, промышленной кооперации, или даже межгосударственного сотрудничества. Каждая из обменивающихся сторон имеет большое число автоматизированных систем, содержащих информацию об одних и тех же объектах и событиях; однако структура этой информации может существенно отличаться как из-за особенностей хранения в разных системах, так и из-за различия точек зрения владельцев информации на описываемые объекты и события. Кроме того, структура информации со временем постоянно изменяется. Решение интеграционной задачи в таких условиях при помощи стандартных средств (ETL-процедур, веб-сервисов и др.) приводит к росту стоимости поддержки решения в геометрической прогрессии по мере добавления новых источников информации. Лучший способ избежать этого, на наш взгляд, состоит в том, чтобы сделать структуру данных одним из видов данных, то есть стереть границу между метаданными и собственно данными. Нужно создать репозиторий информационной модели, доступный всем участникам обмена при помощи программного интерфейса (или федерацию таких репозиториев). Необходимо предусмотреть возможность создания частных фрагментов модели, отражающих разные точки зрения, а также внести в модель правила сопоставления (мэппинга) между элементами таких моделей. Так, одна и та же торговая операция для одного участника является «покупкой», а для другого – «продажей»; отпускная цена для одного участника является закупочной для другого. Разумеется, создание комплексных интеграционных решений с использованием онтологий требует и наличия развитого инфраструктурного программного обеспечения – хранилища моделей с программным интерфейсом доступа, сервисной шины ESB для реализации транспортного слоя обмена, универсальных адаптеров для преобразования информации.

Нельзя обойти вниманием сценарий, связанный с обеспечением доступности знаний, накопленных организацией. Обычные системы управления контентом предлагают инструменты поиска, основанные прежде всего на полнотекстовой индексации всех доступных документов, в лучшем случае – их сегментации при помощи «тэгирования» или создания метаописаний на основе фиксированных наборов признаков. Все это предоставляет пользователю довольно скудный инструментарий осмысленного поиска знаний; результаты поиска представляют собой обширные наборы записей, в какой-то степени релевантных запросу пользователя, среди которых ему предстоит самостоятельно найти нужную информацию.

Семантический поиск основан на ином принципе. Он работает не столько с документами, представляющими собой фрагменты слабо структурированного содержания, сколько с фактами, извлеченными из этого содержания и представленными в соответствии с концептуальной моделью. При использовании семантического поиска пользователь задает системе вопрос (в терминах концептуальной модели) и получает точный, логически верифицируемый ответ. Можно легко сравнить эти два способа поиска, набрав в любом интернет-поисковике фразу «*самый дешевый смартфон на Android 5 в Москве*» и сформулировав аналогичный запрос в любой системе структурированного поиска товаров. В первом случае результатом будут сотни тысяч страниц, среди которых найти действительно самый дешевый смартфон практически нереально. Во втором случае набор результатов будет содержать только товары, точно соответствующие условиям поиска, и при помощи сортировки среди них легко можно будет найти самый дешевый (в качестве еще одной аналогии знатоки фантастической литературы могут вспомнить «Большой Всепланетный Информаторий», описанный братьями Стругацкими, и сравнить его с нынешним интернетом).

Создав единый логический массив всей корпоративной информации и предоставив корпоративным пользователям возможность семантического поиска по нему, компания получит инструмент, достойный носить название «база знаний» или «система управления знаниями». С помощью такого инструмента сотрудники компании смогут задавать вопросы автоматизированной системе так же, как если бы они общались с реальным человеком-экспертом, и получать гарантированно верные ответы, соответствующие смыслу вопроса. Таким образом, преобразование информации в соответствии с онтологией, описывающей предметную область, обеспечивает качественный переход от работы с данными к работе со знаниями.

Большинство сценариев использования семантических технологий, описанных в этой статье, применяются нами на практике и подтверждены конкретными проектами, реализованными в крупных рос-

сийских компаниях и организациях. Накопленный опыт позволяет утверждать, что использование онтологий для решения перечисленных классов задач позволяет создавать значительно более гибкие и надежные системы, чем «традиционные» средства, затрачивать на их создание и особенно поддержку меньшие средства, а в ряде случаев – выполнять функциональные задачи, принципиально не решаемые иными способами.

Горшков Сергей Вадимович (serge@trinidata.ru)

Ключевые слова

корпоративные автоматизированные системы, онтология, автоматизация знаний, семантические технологии

Gorshkov S.V. Use ontologies in the corporate automated systems

Keywords

corporate automated systems, ontology, automation of knowledge, semantic technologies

Abstract

Any organization accumulates a huge number of information in the information systems automating her activity. As a rule, data in different systems are structured without uniform concept, poorly connected, badly give in to search and use out of in advance provided routine business processes. As a result the saved-up information becomes inefficiently used asset for which support considerable means, but which brings quite limited benefit are spent. Potential of use of this information is huge, but his disclosure requires the combination of administrative will and motivation, vision of IT architecture and adequate technological means to an opportunity which isn't often found in practice to realize the ambitious project.

DOI: 10.34706/DE-2018-01-12