

# Использование промышленной технологии анализа текстов на естественных языках в работе ситуационных центров

*О. В. Ена*

Первый заместитель генерального директора ЗАО «Авикомп Сервисез»

Важным направлением деятельности современного ситуационного центра является сбор и обработка неструктурированных источников информации — отчётов, сводок, статей электронных СМИ, новостей Интернет.

При расширении области информационного охвата для мониторинга и увеличении динамики поступления новостей и сводок большое значение приобретает применение эффективных и производительных автоматизированных средств обработки и обобщения неструктурированных источников.

Это позволяет дополнить оперативные и консолидированные отчеты сведениями об объектах внимания, извлеченными из текстовых документов, выявить неявные связи между объектами внимания, сформировать пресс-портреты людей и организаций на основе информации документарных хранилищ и сети Интернет.

С ростом доли информации, представленной в неструктурированном виде как в России, так и за рубежом (по оценкам аналитических агентств доля неструктурированной информации составляет более 80 %) увеличивается количество государственных и межгосударственных программ, направленных на стандартизацию, формирование методологических основ и внедрение новых технологий анализа неструктурированной информации.

Так или иначе, все программы связаны с новым направлением развития информационных технологий — семантическими технологиями. В рамках данного направления международный консорциум W3C формирует набор определений, стандартов и принципов работы с текстами на естественных языках.

В качестве примеров можно привести программу немецкого правительства Theseus по созданию механизмов семантического поиска, а также использование семантики для лучшего связывания различных слоев геоинформационных данных в работе ситуационного центра НАТО.

В нашей стране работы в области извлечения знаний из текстов выполняются в рамках различных государственных и ведомственных программ.

В рамках инновационных проектов Федеральной целевой программы «Исследования и разработки по приоритетным направлениям развития научно-технологического комплекса России на 2007 - 2012 годы» компания «Авикомп Сервисез» по заказу Роснауки создала промышленную технологию, которая позволяет выполнять анализ большого количества текстов и обрабатывать миллионы текстовых документов.

Данная технология в этом году проходит приемочные испытания. Уже в настоящее время достигнуты промышленные режимы работы системы.

Система позволяет обрабатывать в параллельном режиме 30 тысяч документов-первоисточников, общее количество документов, обрабатываемых в рамках одной тематической области, достигает 25 миллионов оригинальных документов. Система позволяет хранить не менее 1 млрд. фактов по заданной тематической области.

Программно-техническая архитектура системы задействует более 40 высокопроизводительных физических серверов, развернутых в рамках датацентра Курчатовского института.

В чём же особенность и ключевые отличия технологии «Авикомп Сервисез»?

Данная система способна в автоматическом режиме для разных языков извлекать из обычного текста объекты внимания — люди, организации, политические партии. Причём программное обеспечение с одинаковыми характеристиками точности извлекает из текста как хорошо известных персон — губернаторов, ключевых лиц субъектов Федерации, так и абсолютно неизвестных людей.

Одно из ключевых отличий решения — помимо объектов внимания система автоматически извлекает из текстов семантические отношения между объектами — «работает», «высказывается», «участвует в выборах», и т.д.

Объекты и отношения выделяются на основе правил семантики. В естественном языке, особенно таком гибком как русский, одно отношение может быть представлено большим количеством вариантов. Поэтому на каждый тип объекта и каждый тип отношения профессиональные лингвисты компании разрабатывают от 300 до 500 семантических правил.

В том числе в рамках данной технологии определяется тональность — позитивные и негативные высказывания персон в отношении других объектов внимания.

После обработки для каждого текста генерируется формализованное представление в виде объектов и связей. Уже на этом этапе можно решать качественно новые аналитические задачи, например, определение схожести документов, и основных тем документа. По информационной насыщенности объекта — количеству связей, входящих и исходящих из данного объекта можно определить, что данный объект — суть главная тема документа. Документ может называться по-другому, в нём может упоминаться большое количество других объектов, но форма графа, большое количество связей объекта (родился, учился, купил, участвовал и пр.) безошибочно говорят о том, что данный объект — главная тема документов.

Существенно более широкий спектр приложений можно применить при решении наукоёмкой задачи глобальной идентификации объектов: необходимо отождествить идентичные объекты или, проще говоря, сказать, что объекты, найденные в разных документах одинаковые. Это дает возможность объединить отдельные схемы документов в одну и давать аналитику на всем информационном пространстве обработанных источников. Такое объединение тоже уникальная область созданной промышленной технологии, так как без семантических отношений между объектами качественно ее выполнить невозможно.

Чем больше обрабатывается информации, тем более насыщенными становятся объекты. В итоге формируется огромный граф, содержащий сведения обо всех фактах, объектах внимания и семантических отношениях на всем информационном пространстве обработанных источников.

На данном графе можно применять и разрабатывать разнообразные аналитические инструменты, ведь основная работа — формализация и консолидация данных уже сделана.

Указанная промышленная технология может успешно применяться в работе ситуационных центров разной тематической и организационной направленности.

Так, в рамках сотрудничества «Авикомп Сервисез» и компании «Голлард» данные решения по обработке текстов используются для расширения области информационно-аналитического охвата предлагаемых компанией «Голлард» полнофункциональных ситуационных центров.

Обработывая с помощью данной технологии коллекции документов сети Интернет, файловых хранилищ, транскрипции телетрансляций и радиопередач, можно для разных категорий пользователей под разными углами зрения представлять следующие сведения:

- анализ объёма присутствия объекта внимания в прессе;
- анализ репутации (соотношения негативной и позитивной информации, в том числе, в случае неявного представления);
- мониторинг высказываний и заявлений объекта внимания и об объекте внимания в СМИ;
- построение информационных портретов (досье) объектов внимания на основе информации из корпоративных банков данных и сведений из сети Интернет;
- автоматическое формирование дайджестов по интересующим связям объектов внимания на основе информационных ресурсов сети Интернет (например, «Негативные высказывания в отношении ключевых персон субъекта Федерации», «Выступления Д. Медведева о состоянии приоритетных национальных проектов в ходе поездок по субъектам Федерации в 2009 году»);
- оперативный мониторинг появления информации на заданные пользователем темы.

Форма представления результатов выдачи может быть самая разнообразная. В том числе:

- фрагмент видеостены ситуационного центра, содержащий визуальные элементы для оперативного информирования и мониторинга изменения обстановки;

- набор аналитических отчётов, доставляемых по расписанию, для экспертов ситуационных комнат;
- важной особенностью данного технологического решения является полная поддержка стандартов и открытость интерфейсов доступа ко всем элементам технологического цикла обработки информации. Это позволяет легко интегрировать визуальные и технологические компоненты системы в разнообразные аналитические приложения, уже функционирующие в ситуационном центре.

В настоящее время с применением данной технологии в реальном режиме времени обрабатываются все новостные сообщения высокорепутационных изданий российской и англоязычной прессы — Коммерсант, Газетару, ВВС, Guardian, Washington Post и других (всего более 200 информационных ресурсов). На всём информационном пространстве обработанных источников предоставляется набор семантических сервисов.

Следует отметить, что решения компании «Авикомп Сервисез» не ограничены только общественно-политическим анализом, помимо этого на данный момент реализованы средства обработки текстов для русского, английского, немецкого и французского языков для экономической, медицинской, нанотехнологической и других предметных областей.

Учитывая семилетний опыт компетенции в области извлечения знаний из текстов и перспективность семантических технологий, компания ведёт работу по расширению учебных программ и разработке программ курсов повышения квалификации для подготовки специалистов в области новых дисциплин, связанных с семантическими технологиями — инженеров по знаниям, прикладных лингвистов, аналитиков и программистов.

Специалисты компании могут оказывать методическую поддержку сотрудникам ситуационного центра в части принципов и способов извлечения знаний из неструктурированной информации.

В заключение, следует отметить, что предлагаемая компанией «Авикомп Сервисез» мощная высокопроизводительная, соответствующая всем мировым стандартам промышленная технология анализа текстов на естественных языках позволит не только существенно увеличить область информационного охвата для мониторинга ситуации, но и использовать широкий спектр качественно новых методов и средств для расширения спектра аналитических приложений ситуационного центра ведомства или субъекта Федерации.