Российская академия наук
Институт программных систем
Исследовательский центр
искусственного интеллекта (ИЦИИ ИПС РАН)

ИНВЕСТИЦИОННОЕ ПРЕДЛОЖЕНИЕ

Технология интеллектуальной обработки текста на естественном языке

Бизнес-концепция: совместная разработка технологий, инструментальных средств или конечных приложений анализа текста (например, базы структурированной текстовой информации; сервисы по обработке бизнес-информации в рамках предприятия; персональные приложения для обработки текстовых и частично структурированных документов).

Потенциальные клиенты: различные консалтинговые компании, деятельность которых требует постоянного анализа больших объемов слабоструктурированной информации на естественном языке.

Полезные партнеры: IT-компании, занимающиеся разработкой ПО для лингвистического анализа текстов или накоплением знаний.

Необходимые инвестиции и сроки: 23 млн. рублей на 2 года.

Автоматическая обработка текстов: проблемы и перспективы

В современном мире информация — один из важнейших ресурсов; от полноты, точности и своевременности получения информации зависит успех любого предприятия. Накопленные объемы текстовой и прочей информации стали настолько велики, что для ее восприятия и дальнейшей обработки необходима предварительная структуризация информации.

Проблемой поиска и анализа информации исследовательские коллективы занимаются уже давно, и в настоящее время существует достаточное количество технологий поиска и управления документами. Методы компьютерной лингвистики (информационный поиск, извлечение структурированной информации из текста) предоставляют такие средства, но не всегда в достаточном объеме. Ситуация осложняется тем, что многие методы компьютерной лингвистики требуют значительных вычислительных мощностей.

Разработка технологий автоматической аналитической обработки текстов из различных предметных областей — актуальное и перспективное направление исследований.

От текста - к модели предметной области

Создание специализированных приложений анализа текстовой информации, относящейся к единственной предметной области, экономически неоправданно создание и дальнейшая модификация такого приложения требует крупных трудозатрат.

Таким образом, для эффективной автоматизации задач анализа текста требуется подход, обеспечивающий следующие возможности:

- адаптация к предметной области;
- глубокий лингвистический анализ;
- использование различных видов ресурсов знаний;
- возможность использования в качестве компонента более крупных систем;
- использование методов машинного обучения;
- параллельная обработка данных.

Технология интеллектуальной обработки текста на естественном языке

Опыт ИЦИИ

В ИЦИИ ИПС РАН на протяжении многих лет создаются и совершенствуются технологии и программные средства интеллектуального анализа текстовой информации, ее структуризации, поиска описаний конкретных ситуаций, отношений, фактов. Более того, разработан ряд инструментальных средств и технологий для построения приложений такого рода (в частности, системы INEX и ИСИДА-Т). Гибкие средства конфигурирования и настройки на предметную область позволяют в сжатые сроки создавать прикладные системы аналитической обработки текста с заданными параметрами производительности и точности работы. Возможность параллельной обработки документов (в том числе и на кластерной архитектуре) позволяет создавать поточные высокопроизводительные приложения реального времени для анализа текстов.

Технологии и построенные на их основе инструментальные средства опираются в своей работе на знания, как предметные (онтологии, наборы правил), так и лингвистические (словари разного рода). В процессе работы систем ресурсы знаний могут пополняться.

Приложения технологий анализа текста

Спектр областей применения разрабатываемых технологий очень широк:

- структурирование и визуализация текстовой информации, имеющей коммерческую ценность;
- Мониторинг сообщений в средствах массовой информации;
- автоматизированное пополнение реляционных баз данных на основе информации, содержащейся в неструктурированном тексте;
- поисковые системы;
- системы автоматического аннотирования (реферирования) документов;
- системы классификации документов;
- автоматические средства выявления зависимостей между объектами предметной области.

Помимо прочего, наличие таких технологий позволит сократить временные и финансовые затраты на разработку практически любых приложений обработки текста.

Коммерциализация решения

Для развития и коммерческого использования требуются следующие действия:

- Развитие программных средств (всестороннее тестирование и оптимизация, улучшение пользовательского интерфейса).
- Усовершенствование технологической базы для создания надежного конкурентного преимущества.
- Изучение конкретных рынков сбыта, модификация программных средств и разработка продукта (продуктов) под нужды крупных пользователей.
- Изучение перспектив продажи базовых технологических средств.
- Разработка маркетинговой стратегии для создан-

В число возможных конечных коммерческих продуктов входят:

- сервисы, обеспечивающие доступ к базе структурированной текстовой информации в определенной предметной области (новости, правовая, научная информация);
- сервисы в рамках предприятия, ориентированные на обработку бизнес-информации, значимой для организации (возможна интеграция с системами документооборота);
- персональные Настраиваемые приложения для обработки текстовых и частично структурированных документов (возможна интеграция с механизмами локального поиска).

152020 Россия, Ярославская область, г. Переславль-Залесский, ИПС РАН
Телефон/Факс: +7 (48535) 98064 E-mail: psi@botik.ru
Веб-сайт: http://www.botik.ru/PSI

Переславль-Залесский, 2007