Приложение 2

Резюме проекта (НИР), выполненного в рамках ФЦП «Исследования и разработки по приоритетным направлениям развития научно-технологического комплекса России на 2007 – 2013 годы»

Номер контракта: 07.514.11.4109
Тема: «Исследование и разработка методов автоматического создания фактографических информационных ресурсов на базе анализа текстовых документальных материалов»
Приоритетное направление: Стратегические компьютерные технологии и программное обеспечение
Критическая технология: Технологии обработки, хранения, передачи и защиты информации
Период выполнения: 26 октября 2011г. – 4 сентября 2012г.
Плановое финансирование проекта: бюджет - 4,4 млн. руб.,
внебюджет – 1,1 млн. руб.
Исполнитель: ИПС им. А.К. Айламазяна РАН
Ключевые слова: обработка естественного языка, семантический анализ, извлечение информации, онтология, язык описания правил, отождествление извлеченной информации, фактографическая база данных, ресурс знаний

1. Цель исследования, разработки

Задача проекта — получение научно-технического задела в области создания фактографических информационных ресурсов посредством анализа текстом на русском языке.

Целью НИР являлось получение значимых научных результатов, позволяющих переходить к созданию новых видов научно-технической продукции и усовершенствованию существующих технических решений с целью повышения уровня автоматизации при работе с текстовыми документальными материалами.

В результате выполнения НИР были разработаны подходы и алгоритмы, а также экспериментальное программное обеспечение для извлечения фактографической информации из текстов на русском языке. Спецификой проекта было использование формализованных предметных знаний в процессе извлечения фактов и отождествления информации при множественном ее упоминании в тексте. Это позволило добиться высокой точности выявления фактографического материала довольно сложной многогранной структуры. Точное выявление сложных, многогранных фактов и объектов позволяет применять результаты НИР для построения фактографических информационных ресурсов в широком спектре предметных областей. Таким образом, становится возможным создание новой информационно-аналитической продукции в тех предметных областях, для которых ранее точных методов извлечения информации не существовало.

2. Основные результаты проекта

Выполнен аналитический обзор современной научно-технической, нормативной, методической литературы в области методов и систем извлечения информации из текстов; формализмов представлений знаний и онтологий.

На основании аналитического обзора сделаны выводы по наиболее перспективным направлениям исследований, в числе которых на первом месте стоит использование предметных знаний на этапах выявления и отождествления извлеченной информации.

Проведены патентные и маркетинговые исследования.

Проведены теоретические и экспериментальные исследования методов извлечения информации из текстов под управлением онтологии, методов отождествления извлеченной информации при множественной ее номинации в тексте, методов создания фактографических информационных ресурсов на базе анализа текстовых документов, методов представления знаний в задачах извлечения информации из текстов.

Исследования показали, что для представления знаний в процессе создания программной системы для извлечения информации целесообразно опираться на атрибутоцентрический подход и фасетную организацию концептуальных иерархий. Это упрощает процесс создания и модификации концептуальной модели предметной области. Извлечение информации можно осуществлять при помощи языка описания целевых контекстов извлечения. Целесообразно обращаться к модели знаний при помощи функций непосредственно из правил извлечения информации. Такой подход более экономен к вычислительным ресурсам по сравнению с традиционным подходом, предполагающим привнесение семантической информации (обычно, семантических категорий) до поиска и извлечения целевой информации. Для отождествления информации был разработан декларативный подход, предполагающий описание на специальном языке условий, при выполнении которых пара фактов (или их участников) должна быть отождествлена. Эти условия выражаются в терминах модели предметных знаний. Все известные аналоги осуществляли отождествление лишь на базе статистического материала, что не позволяло им достичь приемлемого уровня точности отождествления.

Эксперименты подтвердили, что разработанные методы извлечения и отождествления информации позволяют достичь качественных характеристик, заявленных в техническом задании.

Проведены обобщение и оценка полученных результатов. Разработаны рекомендации по использованию результатов проведенных НИР в реальном секторе экономики и проект ТЗ на проведение ОКР по теме «Разработка фактографического информационного ресурса, наполняемого и актуализируемого путем анализа информации новостных сайтов на базе методов извлечения и отождествления информации под управлением онтологии».

Новизна полученных решений заключается в разработке методов извлечения и отождествления информации на базе предметных знаний, содержащихся в онтологии, для русского языка. Предложены новые подходы к концептуализации предметных знаний: атрибутоцентрический подход, фасетная организация таксономий, определение атрибутов на верхних уровнях иерархии. Разработан новый способ использования предметных знаний на стадии микросинтаксического анализа текста. Разработаны новые подходы и алгоритмы для отождествления извлеченной из текста информации при множественной ее номинации. Предложены оригинальные подходы к вскрытию фоновых предметных знаний для обогащения результатов извлечения.

Сопоставление с аналогичными работами, определяющими мировой уровень развития технологий. Для анализа русского языка на настоящий момент не существует аналогов, решающих задачу извлечения и отождествления информации из текстов на базе онтологий. Единственным близким решением может считаться система OntosMiner (компании Ontos AG), в которой онтология может выступать в качестве контейнера для извлекаемой информации. Для иностранных языков существуют близкие по технологическому уровню решения: система SProUT и система KIM. Детальное сопоставление результатов, полученных в ходе НИР, с данными системами по количественным показателям невозможно, так как для них данная информация не приводится. По функциональным возможностям эти системы сопоставимы с полученными в ходе НИР результатами. Обе системы используют онтологии на различных уровнях анализа текста с целью повышения точности и полноты извлечения.

3. Охраноспособные результаты интеллектуальной деятельности (РИД), полученные в рамках исследования, разработки

Программа для ЭВМ «Фактография», свидетельство о регистрации № 2012617809 от 29.08.2012, РФ.

4. Назначение и область применения результатов проекта

Результаты проведенной НИР могут быть использованы для проведения ОКР, направленных на создание фактографического информационного ресурса, наполняемого и актуализируемого путем анализа информации новостных сайтов на базе методов извлечения и отождествления информации под управлением онтологии.

Областями применения технологий семантического анализа и структурирования текстовой информации могут быть информационная поддержка бизнеса (business intelligence) и управление знаниями (knowledge management); маркетинговые исследования; финансовая аналитика; военная и коммерческая разведка и мониторинг; информационная поддержка органов государственной власти; работа библиотек, издательств и СМИ.

5. Эффекты от внедрения результатов проекта

Разрабатываемые методы и подходы должны обеспечивать повышение уровня автоматизации при работе с текстовыми документальными материалами, что приведет к повышению производительности труда в различных областях, связанных с аналитической обработкой документов.

6. Формы и объемы коммерциализации результатов проекта

Возможные формы коммерциализации полученных результатов: заключение лицензионных договоров, заключение договоров уступки прав на РИД.

На основе полученных РИД могут быть созданы информационно-аналитические программные системы для различных предметных областей, например, для автоматического анализа вакансий и резюме в кадровом бизнесе, проведения маркетинговых исследований в интернет-ресурсах, для мониторинга определенных событий или сбора статистики по заданным классам событий в новостных источниках.