ОСНОВНЫЕ ПРОБЛЕМЫ ОБРАБОТКИ ТЕКСТОВ В ИНТЕГРИРОВАННЫХ КОРПОРАТИВНЫХ ИНФОРМАЦИОННЫХ СИСТЕМАХ

Аджит Пратап Сингх Гаутам, Шаронова Н.В.
Национальный технический университет
"Харьковский политехнический институт",
г. Харьков, ул. Пушкинская, 79/2, тел. 707–63–60,
e-mail: nvsharonova@mail.ru

Перед автоматизированными системами извлечения знаний из текста сегодня встают насущные практические задачи, появление которых стимулировано развитием Интернета, содержащего огромное количество текстовой информации – реальные элементы утилитарного знания, полученные людьми в результате их деятельности. К таким задачам, как показывает анализ литературных источников, относятся: поиск и извлечение элементов знания, явно присутствующих в текстовой коллекции в виде: а) утверждения; б) факта; порождение сложного знания путем обработки элементов знания следующими способами: а) генерация нового знания как цепочки логического вывода из элементарных утверждений и/или фактов; б) эксплицирование обобщенного знания, скрытого в совокупности частных утверждений и/или фактов [1, 2].

Большой объем информации, циркулирующей в интегрированных корпоративных информационных системах, а также ее низкая структурированность обуславливают повышение роли такой процедуры, как извлечение информации из текста. В Украине также ведутся разработки по интеграции данных, в том числе и семантической интеграции, и интеграции, основанной на онтологии.

Современные процедуры извлечения информации, использующие методы обработки текстов на естественном языке, как правило, направлены лишь на решение узкого класса задач (отбор ограниченного набора тем (вопросов, проблем), а зачастую и только на одну тему). Это привело к тому, что разработанные эффективные процедуры обработки текста невозможно применить в качестве универсального метода обработки текста.

Традиционно в системах анализа текстов для представления знаний используется четыре типа моделей: продукционная, формально-логическая, фреймовая и семантико-сетевая модели. На базе этих моделей описываются решения и основные перспективы их использования [2]. В литературе представлен систематизированный анализ методов, позволяющих обнаруживать и извлекать из текста конструктивные элементы. Исследования таких ученых, как Ю.П. Шабанов-Кушнаренко, М. Ф. Бондаренко, Н. Ф. Хайрова и др. доказывают, что именно использование лингвистических методов позволяет существенно улучшить качество автоматического анализа текста. Следует отметить перспективность использования алгебрологического аппарата для решения задач анализа текста. При этом ядром таких систем могут стать интегрированные интеллектуальные системы, включающие элементы искусственного интеллекта,

основанные на методах и средствах теории интеллекта, развиваемых на кафедре интеллектуальных компьютерных систем НТУ «ХПИ» [2, 4]. Существующее многообразие частных задач обработки текстовой информации позволяет сгруппировать их в следующие крупные классы, связанные с анализом текстовой информации [5]:

1. Распознавание именованных элементов (сущностей), например, имён людей, названий организаций, географических названий и пр.
2. Разрешение анафоры и кореференций: поиск связей, относящихся к одному и тому же объекту.
3. Выделение терминологии: нахождение для данного текста ключевых слов и словосочетаний (коллокаций).
4. Автореферирование: выделение из текста смысловой, эмотивной, оценочной и пр. информации.
5. Корпусная лингвистика: создание и использование корпусов текстов.
6. Создание электронных словарей, тезаурусов, онтологий.
7. Автоматический перевод текстов.
8. Автоматическое извлечение фактов из текста.
9. Построение систем управления знаниями.
10. Создание вопросно-ответных систем.
11. Информационный поиск.

Информационные системы представляют широкий класс программного обеспечения, используемого различными предприятиями для автоматизации их работы. Поскольку объем обрабатываемой информации огромен, уже в каждой организации существует, как правило, несколько информационных систем. Часто в этих системах обрабатывается одна и та же информация. В связи с этим возникает проблема интеграции данных из различных систем. Под интеграцией данных понимается процесс объединения данных из различных источников для получения их согласованного представления, а в широком смысле – процесс организации регулярного обмена данными между различными информационными системами предприятия [5].

Список литературы:

1. Ермаков А.Е. Автоматизация онтологического инжиниринга в системах извлечения знаний из текста / А.Е.Ермаков // Труды Международной конференции Диалог 2008. – С.137.

2. Лингвотехнологии идентификации знаний в информационных системах : монография / О. В. Канищева, Н. В. Шаронова. – Saarbrücken, Deutschland : LAP LAMBERT Academic Publishing, 2013. – 173 с. – На рус. яз.

3. *Бондаренко М. Ф.* Мозгоподобные структуры: Справочное пособие. / М. Ф. Бондаренко, Ю. П. Шабанов-Кушнаренко. Том первый. Под редакцией акад. НАН Украины И.В. Сергиенко. – К.: Наукова думка, 2011. – 460 с.

4. *Хайрова Н.Ф., Шаронова Н.В.* Лингвистические технологии экстракции и идентификации знаний // Тези доповідей Міжнародної науково-технічної конференції "Інтелектуальні технології лінгвістичного аналізу" (м. Київ, 22-23 жовтня 2013 р.). – К.: НАУ, 2013. – С. 7.

5. *Павленко М.А.* Анализ методов решения задачи извлечения информации из текстов / М.А.Павленко // Системи обробки інформації. – 2013. – Т.1.– С.29.