ТЕХНОЛОГИИ ИЗВЛЕЧЕНИЯ ЗНАНИЙ НА ОСНОВЕ ФАКТОГРАФИЧЕСКИХ ДАННЫХ ЭВМ В СТРУКТУРИРОВАННЫХ КОНТЕНТАХ

Дорошенко А.Ю.
Национальный технический университет "Харьковский политехнический институт",
г. Харьков, ул. Пушкинская, 79/2, тел. 707–63–60,
e-mail: doroshenkoanastasiia@gmail.com

В работе обсуждаются новые технологии извлечения знаний на основе фактографических данных, предлагается подход построения семантической сети, позволяющий работать с большими структурированными корпусами текста.

Цель работы – сравнение наиболее актуальных средств интеллектуального анализа структурированных контентов на основе построения семантической сети фактографической информации ЭВМ с помощью алгебры предикатов.

В ближайшем будущем наиболее востребованными станут системы с максимально автоматизированными ETL-процессами структурирования контента (extract, transfer, load- «извлечение, преобразование, загрузка»). Важной чертой таких систем будет функция оперативного анализа информации, полученной по запросу для выбора дальнейшего направления исследования документов (автопилотирование направления исследования), выполняемой с помощью методов интеллектуального анализа текста [4].

К актуальным средствам интеллектуального анализа текстов относятся технологии выделения фактографической информации об объектах с учетом анафорических ссылок на них (ссылочные местоимения на объект, поименованный в тексте); нечеткий поиск; тематическое и тональное (точное и полное) рубрицирование; кластерный анализ хранилищ и подборок документов; выделение ключевых тем; построение аннотаций; построение многомерных частотных распределений документов и их исследование с помощью OLAP-технологий; использование методов анализа текста для определения направления исследования больших подборок документов и извлечения новых знаний [1,2].

Факты выделяются из предложений, содержащих упоминания объектов или ссылки на них. Технология выделения фактов основана на использовании специальных семантико-лингвистических методов, которые дают возможность получить точность и полноту фактов, сравнимую с экспертными. Зачастую факты содержат информацию о взаимосвязях объектов и классифицируются как прямые (имеется факт о связи двух объектов); нечеткие (нет фактов); общего места и времени (для пары различных фактов различных объектов); косвенные, или транзитивные (через общий третий объект-связь у пары фактов различных объектов); рефлективные (между парой атрибутов досье, связанных семантически). Если в одном из них появляется факт с определенным объектом-

связью, то в симметричном атрибуте для объекта-связи также появляется этот факт.

Все эти свойства необходимы в системах аналитической разведки, немыслимых без следующих сервисов: автоматическое выявление прямых и косвенных связей объекта; автоматическое выявление связей объектов по месту и времени; типизация связей, представленных различной лексикой; формирование групп объектов, связанных между собой общностью фактов; построение карты связей объектов для различных типов связей, визуализация и фильтрация связей; поиск оптимальных связей между заданными объектами; построение многомерных частотных распределений фактов. Сегодня системы извлечения фактов являются наиболее эффективным инструментом выделения нужной для принятия решений информации, заменяя ее поиск [3].

Заключение. На основе анализа и сравнения наиболее актуальных средств интеллектуального анализа структурированных контентов, была построена модель семантической сети ЭВМ с использованием фактографических данных и алгебры предикатов. Поиск в семантической сети определенного факта, дает возможность считать, что подсеть установлена, после чего производится извлечение сущностей и их маркировка ролями, заданными в соответствующих узлах лингвистических описаний [1,4]. Таким образом, результатом семантического поиска на основе интеллектуального анализа структурированного контента, является имя факта и набор указателей на сущности семантической сети с указанием соответствующих им ролей в лингвистическом описании.

Список литературы

1. Алисейко З. А. Использование алгебры предикатов и предикатных операций для формализации декларативной и процедурной составляющих знаний / З. А. Алисейко, В. И. Булкин, О. В. Канищева, Н. В. Шаронова // Біоніка інтелекту. – Харків : ХНУРЕ, 2006. – № 1(64). – С. 59-63.

2. Бондаренко М. Ф. О мозгоподобных ЭВМ / М. Ф. Бондаренко, З.В. Дударь, И.А. Ефимова, В.А. Лещинский, С.Ю. Шабанов-Кушнаренко // Радиоэлектроника и информатика. – Харьков : ХНУРЭ, 2004. – № 2. – С. 89-105.

3. Булкин В.И. Математические модели знаний и их реализация с помощью алгебропредикатных структур / В. И. Булкин, Н.В. Шаронова: монография. – НТУ «ХПИ», МЭГИ. : Донецк, 2010. – 304 с.

4. Sussna M. Word sense disambiguation for free-text indexing using a massive semantic network // The Second International Conference on Information and Knowledge Management. – 1993. – P. 67-74.