УДК 004.655

МЕТОД ПОИСКА СЛОВОФОРМ В СТРУКТУРИРОВАННЫХ ТЕКСТАХ С ИСПОЛЬЗОВАНИЕМ «СМЫСЛОВЫХ ЕДИНИЦ» ИЗ ЛИНГВИСТИЧЕСКОЙ ОНТОЛОГИИ

Ночевной Д.С. (ИТМО)**, Бессмертный И.А.** (ИТМО),
Научный руководитель – доктор технических наук, профессор Бессмертный И.А.
(ИТМО)

Введение. Одним из ключевых научно-практических направлений в области современных информационных технологий, требующих автоматизации, являются обработка, хранение и поиск необходимой информации в больших объемах текстовых данных. Человеку не под силу вручную находить соответствующую требованиям информацию за очень маленький промежуток времени, потому что во многих случаях для этого необходимо анализировать множество несвязанной информации. Таким образом, появляется необходимость в инструментах автоматизированного поиска информации и формирования удобного хранилища знаний на ее основе.

Формализовать знания в определенной предметной области и систематизировать их для быстрого доступа всегда являлось важной и актуальной задачей. Во многих научных областях такие системы знаний принято использовать для того, чтобы помочь пользователям легко и оперативно получать требуемую информацию. В итоге, можно выделить проблему качества (пертинентности) поиска нужной информации в больших массивах данных. Целью данной работы стало повышение пертинентности извлекаемой информации из структурированных источников.

В рамках заданной цели выделены следующие задачи.

1. Анализ существующих методов извлечения данных и семантических отношений из структурированных источников.
2. Разработка программного модуля для преобразования текстов на естественном языке в структурированный унифицированный формат.
3. Разработка языка запросов с поддержкой использования лингвистической онтологии для извлечения данных с помощью «смысловых единиц».

Основная часть. В настоящее время существует немало готовых решений, позволяющих делать преобразования текстов на естественном языке в структурированный формат [1, 2]. Поэтому в данной работе была выбрана библиотека, наиболее подходящая по критериям, после чего были реализованы дополнительные шаги для получения требуемого унифицированного формата документов.

В разработанный язык запросов была также добавлена возможность использования «смысловых блоков». Поисковый запрос с использованием «смысловых блоков» или «смысловых единиц» позволяет выполнять поиск, оперируя одновременно десятками или даже сотнями возможных словоформ, которые могут иметь заданное или более узкое (гипонимы) значение. Этот прием значительно повышает полноту поиска с использованием специальных выражений, и после этого остается лишь отсеивать неверные результаты, постепенно повышая точность алгоритма.

Проведено сравнение средней точности нескольких поисковых систем, включая поисковую систему на основе разработанного алгоритма. Следует отметить, что средняя точность поисковых систем была вычислена на основе нескольких тестовых запросов в различных предметных областях, в том числе законодательство, музыка, образование, науки об окружающей среде [3]. Исходя из полученной информации, был сделан вывод, что поисковая система на основе разработанного языка запросов позволяет достичь более высокой точности.

Выводы. В итоге, был проведен анализ существующих методов извлечения данных и семантических отношений из структурированных источников. Также был разработан конвертер на языке Java для преобразования текста на естественном языке в унифицированный формат, используемый для поиска в тексте информации с помощью специальных регулярных выражений. Кроме того, было рассмотрено решение задачи по добавлению в язык запросов поддержки использования лингвистической онтологии для извлечения данных с помощью «смысловых единиц». В результате исследования был разработан метод поиска словоформ в структурированных текстах, отличающийся использованием «смысловых единиц» из лингвистической онтологии, обеспечивающий более высокую пертинентность извлечения требуемой информации.

Список использованных источников:

1. Обработка текста – NLPub [Электронный ресурс]. – 2018. – URL: https://web.archive.org/web/20220812210142/https://nlpub.ru/Обработка_текста#Синтаксический_анализ (дата обращения 25.01.2024).
2. Мухамедиев Равиль Ильгизович, Сымагулов Адилхан, Кучин Ян Игоревич, Абдуллаева Сабина, Абдолдина Фарида Наурузбаевна Облачные сервисы для обработки текстов на естественном языке // Современные информационные технологии и ИТ-образование. 2018. №4. URL: https://cyberleninka.ru/article/n/oblachnye-servisy-dlya-obrabotki-tekstov-na-estestvennom-yazyke (дата обращения: 25.01.2024).
3. Brophy J., Bawden D. Is Google enough? Comparison of an internet search engine with academic library resources //Aslib proceedings. – Emerald Group Publishing Limited, 2005. – Т. 57. – №. 6. – С. 498-512.