УДК 004

Автоматизированный анализ открытых источников по нормоконтролю в Google Scholar с использованием языка Python

Кобец Е.А., Кумпан В.В., Басакина А.А, факультет ИКТ, Университет ИТМО
Горлушкина Н.Н., к.т.н., доцент, с.н.с., факультет ИКТ, Университет ИТМО

Введение

При проведении исследований приходится сталкиваться с поиском необходимой информации в различных источниках. В настоящее время этот поиск осуществляется, как правило, в базах данных Интернета. Во время ручного поиска в наукометрических БД ученому-исследователю необходимо из всего массива открытых данных найти необходимые материалы (статьи и кейсы). При этом количество затрачиваемого времени оказывается достаточно ощутимым. А с возрастанием объема информации возрастает и количество затрачиваемого времени, что является проблемой в наш динамичный век.

Целью работы является автоматизация процесса поиска информации для анализа предметной области.

В рамках исследования открытых источников проводились аналитические работы с использованием методов логического и сравнительного анализа, а также с привлечением методов группировки и зависимости в таких базах данных (БД) как: Google Scholar (scholar.google.ru), Федеральный институт промышленной собственности (www1.fips.ru), Научная электронная библиотека eLIBRARY.RU (elibrary.ru), научная электронная библиотека Киберленинка (cyberleninka.ru), библиографическая и реферативная база данных Scopus (scopus.com).

Для поиска научных материалов (статьи и кейсы) с использованием вышеперечисленных методов и их реализация с использованием языка Python - обеспечило качественное моделирование алгоритма для автоматизированного анализа открытых источников, с выгрузкой всех релевантных источников в отдельную таблицу для последующего самостоятельного более детализированного и предметного анализа из полученного имеющегося массива информации. Также, на основании дополнительно заданных критериев и весов, и с использованием настраиваемого фильтра, можно определить наилучшие для ознакомления и изучения материалы.

В итоге использование указанного подхода выполненного на языке программирования Python, позволило автоматизировать поиск и составить выборку релевантных источников в БД Google Scholar (scholar.google.ru), а также добиться детализированного результата.

Для тестирования предлагаемого подхода автоматизированного поиска информации в наукометрических базах были выбраны «ключи», связанные с областью нормоконтроля, однако предлагаемая разработка может быть использована для анализа и в других областях знаний. Ключами называются ключевые понятия, характеризующие заданную предметную область.

Изначально был выполнен анализ работы с базами данных (БД) (Google Scholar (scholar.google.ru), Федеральный институт промышленной собственности (www1.fips.ru), Научная электронная библиотека eLIBRARY.RU (elibrary.ru), научная электронная библиотека Киберленинка (cyberleninka.ru), библиографическая и реферативная база

данных Scopus (scopus.com). Этот анализ позволил определить алгоритм поиска как в ручном, так и автоматизированном режимах.

Описание ручного поиска позволило разработать краулер по поиску информации. Его программная реализация выполнена с использованием языка Python.

Алгоритм работы заключается в следующем.

1. Совместно с экспертами подбираются «ключи» на русском языке для ручного поиска, в нашем случае они относились к области нормоконтроля.
2. После этого, «ключи» были переведены также на английский язык, и этот перевод был согласован с экспертами.
3. Далее был выполнен автоматизированный поиск по теме работы в рамках заданных «ключей» по БД с использованием краулера.
4. В итоге, результаты поиска информации по нормоконтролю были автоматически выгружены в таблицу. Таблица содержит все выявленные источники и их характеристики. В качестве характеристик преведены следующие сведения (в колонку A заносится «ссылка на источник информации»; в колонку B заносится результат по «длительности разработки»; в колонку C заносится результат по «актуальности работы»; в колонку D заносится результат «через сколько был первый успех»; в колонку E заносится результат «чем выражен результат», в колонку F заносится результат «была ли реализация», в колонку G заносится результат «количество страниц / сайт»).
5. Далее можно проводить анализ источников информации по заданной предметной области.

Основной результат

Предложены способ и средство для автоматизированного поиска в открытых источниках информации для исследования предметной области, в основе которых лежат заранее сформированные экспертами «ключи» и подготовленные для поиска в Google Scholar (scholar.google.ru). Разработанное средство было выполнено на основе языка Python. Получена выборка релевантных источников в БД Google Scholar (scholar.google.ru) по теме нормоконтроля. Итогом работ является составленная автоматически выборка релевантных источников и визуализация этих результатов на основании дополнительно заданных критериев и весов.

Вывод

Использование указанного подхода позволило автоматизировать поиск требуемой информации в БД Google Scholar (scholar.google.ru), на его основе составить выборку релевантных источников относительно изучаемой области: нормоконтроль. Однако, предлагаемый подход может быть использована для поиска и анализа информации и в других областях знаний.

Проведенный эксперимент показал, что предложенный подход позволяет значительно сократить затрачиваемое количество времени на поиск и анализ полученной информации.