УДК 004.822

РАСШИРЕНИЕ ПОИСКОВЫХ ЗАПРОСОВ С ИСПОЛЬЗОВАНИЕМ СЕМАНТИЧЕСКОЙ СЕТИ С УЧЕТОМ КОНТЕКСТА ПОИСКА

Рогаленко Н.А. (Университет ИТМО)
Научный руководитель – старший преподаватель Цопа Е.А.
(Университет ИТМО)

В докладе рассматриваются проблемы повышения пертинентности результатов информационного поиска. В рамках работы для увеличения числа результатов поиска, удовлетворяющих пользователя, предлагается расширять исходные поисковые запросы близкими по значению терминами, извлеченными из семантической сети, с учетом предметной области, по которой производится поиск. Основная часть доклада посвящена разработке подходов к автоматическому определению контекста поискового запроса и поиску семантически близких терминов для расширения, а также созданию программной системы, осуществляющей расширение исходного запроса.

Введение

С постоянным увеличением объема информации как в сети интернет, так и в корпоративных и иных информационных системах, как никогда становится актуальной проблема поиска среди множества страниц и документов тех элементов, которые удовлетворяют поисковому запросу пользователя. При этом критически важно обеспечить высокое качество поиска. Несмотря на существенный прогресс в этой области, среди всего объема страниц и документов в коллекции, среди которой производится поиск, не всегда удается найти результаты, удовлетворяющие пользователя, что может быть вызвано, например, несоответствием запроса предметной области поиска, использование слишком коротких запросов, несоответствие лексики запроса и документов корпуса [1].

Для повышения качества поиска используется расширение поисковых запросов (query expansion), данный процесс представляет собой комплекс методов для анализа пользовательского поискового запроса и добавления в него новых терминов и понятий, близких по смыслу к терминам исходного запроса, что позволяет получить больше релевантных результатов в процессе поиска. Для решения проблем несоответствия лексики пользователя терминам документов, а также учета предметной области и контекста, в котором происходит поиск, предлагается использовать подход к расширению запросов с использованием семантической сети как источника данных для расширения. Благодаря графовой структуре сети и наличию семантических отношений между понятиями, представляющими собой вершины графа [2], появляется возможность находить подграфы понятий, относящихся к одной предметной области, а также выстраивать иерархию предметных областей, позволяя выбирать для расширения те понятия, которые семантически близки к понятиям исходного запроса на заданных уровнях контекста.

Основная часть

Первостепенной задачей, решаемой в рамках данной работы, стало определение подхода к автоматическому выявлению контекста текстового документа. Был предложен подход, предполагающий последовательный обход текста окном заданной длины, внутри которого производился поиск наиболее важных (часто используемых) терминов. Найденным терминам ставились в соответствие понятия семантической сети. Далее для каждого понятия производился переход вверх по иерархии понятий семантической сети (от частных понятий к более общим), вплоть до корневого. При совпадении вышестоящих узлов считалось, что найден общий контекст, при совпадении родительского узла найденного понятия,

представляющего собой контекст, с родительским узлом другого понятия считалось, что найден контекст более высокого уровня. Подобный метод позволил проследить цепочку вложенности предметных областей, позволяя переходить от более частных к более общим в зависимости от специфики поиска. Движение же по тексту позволило проследить, как меняется контекст по ходу текста.

Следующим шагом стала разработка метода определения семантической близости понятий в найденном подграфе контекста. В процессе выполнения данного этапа было установлено, что в общем виде задачу решить невозможно, поскольку понятие семантической близости индивидуально для каждой предметной области. В связи с этим был предложен механизм задания правил близости для отдельной предметной области. Правила представляют собой набор переходов от заданного узла по указанным семантическим отношениям на указанную глубину; понятия, найденные в результате этих переходов, считаются близкими к исходному.

После этого с учетом ранее определенных методов и подходов была разработана программная система для расширения поисковых запросов, которая работает следующим образом. В первую очередь производится первичный поиск по исходному запросу пользователя, в результате которого извлекается корпус релевантных документов. Далее этот корпус с исходным запросом передаются модулю определения контекста, и на основе корпуса релевантных документов определяется контекст поиска. Следующим этапом производится обход терминов исходного запроса и для каждого термина ищутся семантически близкие понятия согласно правилам, определенным для найденной предметной области. Найденные таким образом понятия добавляются в исходный запрос, образуя запрос расширенный, на основе которого происходит повторный поиск.

Выводы

В конечном итоге была разработана система, позволяющая расширить поисковой запрос семантически близкими терминами с учетом контекста результатов поиска и предметной области. Методом экспертной оценки было установлено, что в результате повторного поиска на основе расширенного с помощью полученной системы запроса удалось увеличить пертинентность результатов информационного поиска. Кроме того, был предложен подход к определению контекста произвольного текста, представленного на естественном языке, который также может быть использован и в других прикладных задачах, отличных от информационного поиска.

Список использованных источников:

1. Azad H. K., Deepak A. Query expansion techniques for information retrieval: a survey //Information Processing & Management. – 2019. – Т. 56. – №. 5. – С. 1698-1735.
2. Клименков С.В., Николаев В.В., Харитонова А.Е., Гаврилов А.В., Письмак А.Е., Покид А.В. Применение семантической сети для хранения слабоструктурированных данных //Инженерный вестник Дона. — 2020. — №. 2 (62).