ОБОЗРЕНИЕ
ПРИКЛАДНОЙ И ПРОМЫШЛЕННОЙ
Том 31 МАТЕМАТИКИ Выпуск 1
2024

И. О. Игнатьев, **С. Ю. Катышев** (Москва, Лаб. ТВП).
Формирование цепочек преобразований и их применение.

УДК 519.233.33 DOI https://doi.org/10.52513/08698325_2024_31_1_1

Резюме: В представленной работе рассматривается процесс формирования цепочек преобразования поисковых запросов и варианты их применения. Данный труд находится на пересечении таких научных направлений как теория информации и компьютерная лингвистика и является несомненно одной из популярных сфер исследований в последнее время.

Ключевые слова: запросы, расстояние Ливенштейна.

Поисковые запросы сети интернет разделяются на две группы: состоящие из одного слова или слова с предлогом и развернутые запросы, состоящие из словосочетаний или предложений. Предметом исследования данной работы являются короткие запросы. После разделения их на группы схожих, можно сформулировать три задачи. Первая — проанализировать каждую группу и выявить закономерности, вторая — сформировать цепочки преобразований и третья — с помощью их смоделировать запросы.

Существует множество исследований по данной теме. Одной из значительных работ, которая заслуживает внимания и изучения, является статья [1], в которой проводится анализ существующих методов исправления запросов, разработанных компаниями «Яндекс» и «Google». Рассмотрены методы исправления ошибок и на основе эмпирических данных выбран наилучший.

Для выявления закономерностей необходимо построить цепочки преобразования словоформ. Сначала разделим запросы одной группы на подгруппы, находящиеся на разных расстояниях Левенштейна от начальной словоформы, вокруг который и была выстроена группа (Рис. 1).

Рис. 1

Далее были найдены цепочки преобразований и найдена их численность.

Рис. 2.

Для моделирования поисковых запросов на основе цепочек преобразований, предлагается два варианта генерации. Применение целых цепочек, в зависимости от их ценности, коррелирующих с частотой их встречаемости. Комбинирование цепочек, основанное на байесовских оценках: цепочки, в которых присутствуют уже примененные преобразования, теряют вес, остальные равномерно прибавляют.

Проведённые исследования подтвердили эффективность предложенного подхода построения цепочек преобразований и его применимость в различных областях.

СПИСОК ЛИТЕРАТУРЫ

1. *Алдошин М. В.*, *Андросов А. Ю.*, *Бородащенко А. Ю.*, *Зуева Ю. Г.* Модифицированный алгоритм исправления ошибок в информационно-поисковых запросах. — Издания Тульского государственного университета. Технические науки, 2020.
2. *Карпенко М. П.*, *Протасов С. В.* Некоторые методы очистки словаря запросов поиска. — Компьютерная лингвистика и интеллектуальные технологии. М.: Бекасово: РГГУ, 2011, с. 326–338.
3. *Левенштейн В. М.* Двоичные коды с исправлением выпадений, вставок и замещений символов. — Доклады Академии Наук СССР, 1965.

Поступила в редакцию
13.XII.2024

UDC 519.233.33 DOI https://doi.org/10.52513/08698325_2024_31_1_1

Ignatev I. O, ***Katyshev S. Yu.*** (Moscow; TVP Laboratory). **Formation of transformation chains and their application.**

Abstract: The presented paper examines the process of forming chains of conversion of search queries and their application options. This work is located at the intersection of such scientific fields as information theory and computational linguistics and is undoubtedly one of the most popular areas of research in recent times.

Keywords: queries, Livenstein distance.