Стемминг (Stemming)

Разделы: Алгоритмы

Стемминг представляет собой процесс нахождения основы для заданного слова. Аналогично «лемматизации» позволяет проанализировать все словоформы одного слова как единый элемент, что значительно повышает качество анализа. Алгоритмы стемминга используются многими поисковыми системами при поиске информации согласно введенному пользователем ключевому запросу.

Алгоритм стемминга может реализовываться несколькими способами:

- поиском по заданному словарю (аналогично лемматизации);
- усечением окончаний;
- удаление приставок, суффиксов и окончаний;
- комбинацией нескольких вариантов.

Поиск по словарю является сложным процессом с точки зрения затрат вычислительной мощности (нужно каждое слово в тексте сопоставить с таблицей значений). При этом он обеспечивает высокое качество стемминга, т.к. ошибка будет заключаться только в отсутствующих в словаре словах, что разрешается его регулярным обновлением и актуализацией.

Более быстрым и простым решением будет выглядеть удаление окончаний или удаление приставок, окончаний и суффиксов. При этом качество стемминга будет зависеть от заложенных в алгоритм способов морфологического разбора слова (определения части речи и выделения «добавок» к слову). Недостаток подхода очевиден: на английском языке, например, слово «eat» и «ate» может считаться разными вариантами, хотя являются двумя формами одного и того же слова.

Комбинация нескольких вариантов позволяет найти компромисс между вычислительными затратами и качеством стемминга. Он позволяет осуществлять проверку по словарю не всех слов в тексте, а лишь отдельных словоформ, требующих более подробного разбора. По этой причине большинство современных алгоритмов стемминга используют данный подход.