ЭВРИСТИЧЕСКИЙ МЕТОД ИЗВЛЕЧЕНИЯ КОРНЯ СЛОВА НА РУССКОМ ЯЗЫКЕ

Белорусский государственный университет информатики и радиоэлектроники
г. Минск, Республика Беларусь

Третьяков Ф. И.

Серебряная Л. В. — канд. техн. наук, доцент

Одним из наиболее эффективных методов полнотекстового поиска является морфологический разбор слова. Выделение главной морфемы слова — корня — дает возможность произвести поиск с более высокой точностью.

На сегодняшний день существует большое количество неупорядоченной текстовой информации. Поэтому поиск необходимой информации по ключевому слову является одной из важнейших задач.

Существует множество способов найти информацию. Одним из самых популярных является поиск по полному совпадения текста. Однако, он может давать коллизии. К примеру, пользователь задает слова для поиска «кошка». Поэтому в тексте, либо совокупности тексте будут выделяться слова которые состоят из «кошка..» и любого продолжения. Такой поиск самый банальный и совершенно очевидно, что он отсекает такие результаты, которые могли бы быть полезны для пользователя, к примеру: «кошачий», «кот». В таком случае необходим более сложный поиск — семантический [1].

Настоящая работа посвящена поиску в текстах на русском языке. В русском языке удобно использовать корень слова в качестве входного параметра функции поиска, потому как корень слова —есть его смысловая единица, а все образованные от него слова как правило несут похожий смысл. В любом случае, это увеличивает точность поиска.

Одним из способов выделения корня слова является стемминг [2]. Первая публикация на заданную тему датируется 1968 годом. На сегодняшний день созданы различные реализации алгоритмов стемминга. Они применяются для решения различных задач интеллектуальной обработки текстовой информацией.

Для решения задачи поиска информации используется специальный алгоритм стемминга под названием — стеммер [2]. Он может выделять корень слова (стем).

Однако, стеммер может допускать ошибки.

Ошибки стемминга 1-го рода: стем дает слишком большое обобщение и поэтому сопоставляется с грамматическими формами более чем одной словарной статьи. Это самая многочисленная группа ошибок стемминга. Ошибки стемминга 2-го рода — усечение формы дает слишком длинный стем, которые не сопоставляется с некоторыми грамматическими формами этого же слова. К таким ошибкам приводит стремление разработчика стеммера найти компромисс с ошибками 1-го рода в случае, когда при словоизменении меняется основа слова. Ошибки стемминга 3-го рода — стем построить невозможно из-за изменения в корне слова, которое оставляет единственную букву в стеме. Либо модель словоизменения подразумевает использование приставок.

Для создания эвристического стеммера необходимы словари окончаний, формы причастий и деепричастий, суффиксов и приставок. По данным словарям и будет эвристически определяться часть речи. Суть алгоритма сводится к определению части речи для слова по его окончанию используя словари окончаний. Порядок определения задается уникальностью окончания данной части речи. К примеру, окончания причастий невозможно спутать ни с чем другим, поэтому, стемминг начинается именно с них. Далее стемминг будет проходить по следующему алгоритму.

1. Происходит поиск окончаний причастий и деепричастий в слове. Если оно есть, то удаляем его и переходим к шагу 3.
2. Ищется окончания прилагательных, глаголов или существительных. Если нашли их, удаляем.
3. Если слово оканчивается на «и», удаляем его.
4. Находим в слове следующую последовательность с его начала: гласная-согласная. Все буквы после этого сочетания будут блоком *n*. Если ее нет или блок *n* пустой, переходим к шагу 7.
5. Ищем в блоке *n* блок *m* — это блок, который следует после конструкции гласная-согласная. Если его нет, или он пустой, переходим к шагу 7.
6. Ищем в блоке m части слова «ост» и «ость». Если находим, удаляем.
7. Если слово имеет окончание «ейш» или «ейше» то удаляем его.
8. Если на конце получается удвоенное «н», удаляем второе.
9. Если на конце слова «ь», удаляем его.

Стеммер позволяет сделать поиск в тексте на русском языке более осмысленным и логичным. Минусами является сложность архитектуры модуля и пониженная точность.

Список использованных источников:

1. Третьяков, Ф. И. Методы обработки текстовой информации [Текст] / Ф. И. Третьяков, Л. В. Серебряная // VI международная научно-практическая конференция «Актуальные вопросы методики преподавания математики и информатики».— Биробиджан, 2011.— С. 175–181.
2. TextMining. Глубинный анализ текста. Из цикла лекций «Современные Internet-технологии» для студентов 5-го курса кафедры Компьютерных технологий физического факультета Донецкого национального университета. ДонНУ, кафедра КТ, проф. В. К. Толстых