УДК 004.021

# ОБЗОР МЕТОДОВ ПОВЫШЕНИЯ ЭФФЕКТИВНОСТИ ПОИСКОВЫХ СИСТЕМ. ИССЛЕДОВАНИЕ АЛГОРИТМА СТЕММИНГА

*Безуглый Е.Н. Аноприенко А.Я.*
*Донецкий национальный технический университет*

## Введение

Классические механизм поиска документов по индексу достаточно очевидны [1]: пользователь выражает его информационные потребности путем перечисления и ввода в поисковую систему соответствующих словоформ, после чего система преобразует введенный запрос к своему внутреннему языку запросов и выполняет этот запрос. Существуют несколько методов преобразования запроса из натурального языка в язык запросов. Очевидно при этом, что и сама система, принимающая такие запросы должна оперировать специальным языком запросов. Рассматриваются основные 3 метода формирования запросов.

## 1 Словарный метод

Основной идеей метода является создание большого словаря всех слов и форм, объединенных по признаку общего (первоначального) слова. При поиске алгоритм преобразования запросов выполняет поиск по словарю и выдает общие случаи появления таких слов или дополняет запрос различными словоформами.

В системах с динамически поддерживаемыми словарями время поиска при увеличении объема базы документов сначала также увеличивается (т. к. пропорционально увеличивается объем словаря и, соответственно, объем индекса), а затем так же, как в системах со статическими словарями, перестает зависеть от

объема базы документов. Это объясняется тем, что с некоторой границы объема базы документов словарь системы уже набирает практически полный набор словоформ, присущих конкретной предметной области, и вероятность появления в новом документе слова, которого еще не было в словаре системы, резко падает.

Недостатки такого метода очевидны: большой объем словаря (и соответственно, требуемого дискового пространства), большое время поиска по словарю (и соответственно, большие временные затраты). При этом не учитывается факт эволюции натурального языка, т.е. словарь работает с большим, но ограниченным набором слов, что является также недостатком (хотя и разрешимым).

## 2 Алгоритмы Стемминга

Повышению эффективности поиска способствует морфологический разбор документов и запросов. Помимо существенного уменьшения индекса системы и отказа от словаря, морфологический разбор повышает и эффективность поиска, так как не реагирует на несущественные с точки зрения смыслового содержания грамматические различия искомого текста документов и запросов [2, 5].

Алгоритм стемминга представлен в виде таблицы переходов конечного автомата [2]. Каждое правило – это префикс, окончание и последние несколько букв неизменяемой основы. Правила были порождены автоматически следующим образом: некоторое количество полнотекстовой информации было разбито на слова, которые подавались на вход морфологического анализатора основанного на словаре. Для всех распознанных анализатором словоформ выделялись их точная основа, отделенное окончание с двумя последними буквами основы либо формировалось новое правило, или увеличивался вес существующего правила. Затем правила были проранжированы по вероятности встречи в текстах, и маловероятные (с вероятностью менее одной десятитысячной) были отброшены. К алгоритму также было

добавлено специальное правило, гласящее, что неизменяемая основа должна содержать как минимум одну гласную.

Результат работы алгоритма – все допустимые варианты выделения формальной основы поданного на вход слова.

### 3 Тезаурусы и WordNet

В современной лингвистике тезаурус – это особая разновидность словарей общей или специальной лексики, в которых указаны семантические отношения (синонимы, антонимы, паронимы, гипонимы, гиперонимы и т.п.) между лексическими единицами. Тезаурусы в электронном формате являются одним из действенных инструментов для описания отдельных предметных областей и создания информационно-поисковых систем. Как правило, тезаурусы ограничены областью применения, поскольку они рассчитаны только на одну определенную (конечную) сферу использования или для полностью мертвых языков.

Расширением и усовершенствованием такого рода словарей является WordNet. Он содержит все слова (как правило, из энциклопедического источника), упорядоченных в дерево по смыслу. WordNet - это семантическая сеть, разработанная в Принстонском университете, и выпущенная вместе с сопутствующим программным обеспечением под «некопилефтной» (от слова «копилефт» - своеобразного антонима термина «копирайт») свободной лицензией [6].

Узлами сети являются «синсеты» — множества синонимов, имеющих общий смысл, и список кратких общих определений. Между синсетами существуют связи: «синонимы», «гипонимы-гиперонимы» и т. п.

Преимущество такого подхода очевидны: быстрая навигация по дереву, поиск синсетов, а также вычисление логического расстояния между синсетами (на подобии, как в векторной модели поиска).

К недостаткам словаря можно отнести то, что он занимает

определенное дисковое пространство (около 30 Мб), хотя это и преодолимая проблема (весь объем словаря можно разместить в оперативной памяти, при этом, уменьшив время доступа и поиска нужного синсета), и другая проблема состоит в отсутствии такого WordNet для большинства языков. Для русского языка существует WordNet, переведенный с английского языка (http://wordnet.ru/). Для остальных языков существуют только тезаурусы.

## 4 Исследование эффективности использования алгоритма стемминга

Исследование проводилось при следующих условиях: набор документов (на основе сайтов магистрантов ДонНТУ), проиндексированный в одном случае с применением алгоритма

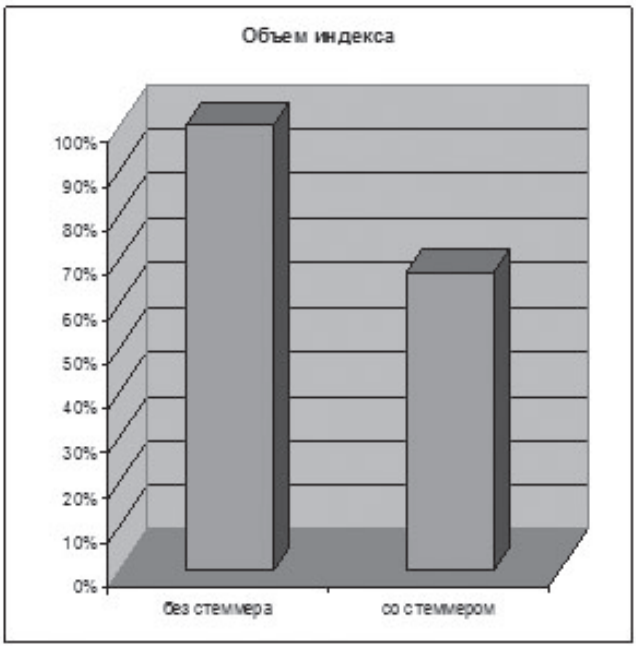


Рисунок 1 - Диаграмма зависимости объема занимаемого пространства жесткого диска для индексных файлов

стемминга, в другом – без каких-либо специальных алгоритмов. Программы индексирования и поиска одинаковые, отличие лишь в использовании функции стеммера.

На рис.1 приведена диаграмма зависимости объема занимаемого пространства жесткого диска для индексных файлов в первом случае созданных без стемминга, во втором случае со стеммером. Как видно разница объемов составляет около 30%, что для больших объемов данных может быть существенно.

Другим важным показателем является зависимость количества найденных файлов от количества слов в запросе. Результат исследования приведен на рис.2. При этом последовательно задавались запросы: 1) «исследования», 2) «научные исследования», 3) «научные исследования для», 4) «научные исследования для статьи».

Видно, что для программы с алгоритмом стеммера количество найденных документов растет с количеством слов в

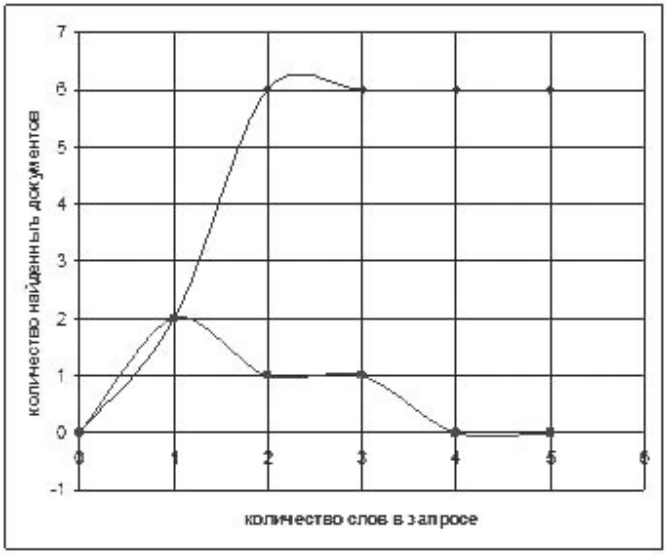


Рисунок 2 - Зависимость количества найденных файлов (релевантных) от количества слов в запросе

запросе, чего нельзя сказать об альтернативном алгоритме. Эти данные свидетельствуют о росте таких показателей как точность и полнота поиска, что напрямую влияет на эффективность поиска.

**Литература**

[1] Frakes, W.B. & Baeza-Yates, R (1992) Information Retrieval: Data Structures and Algorithms, Englewood Cliffs, NJ, Prentice Hall, 504 p. (http://www.scribd.com/doc/13742235/Information-Retrieval-Data-Structures-Algorithms-William-B-Frakes).

[2] Popovic, M. and Willett, P., (1992) The effectiveness of stemming for natural language access to Slovene textual data, Journal of the American Society for Information Science, 43(5), 384-390.

[3] Hull, D.A. & Grefenstette, G. (1996) A Detailed Analysis of English Stemming Algorithms, Xerox Technical Report.

[4] Коваленко А. Стемка — морфологический анализ для небольших поисковых систем // «Системный администратор» № 1, Октябрь 2002 (http://www.samag.ru/cgi-bin/go.pl?q=articles;n=10.2002).

[5] Сегалович И., Маслов М. Русский морфологический анализ и синтез с генерацией моделей словоизменения для не описанных в словаре слов. Казань, ООО «Хэтер», 1998. Т. 2. С. 547-552.

[6] Материалы сайта WordNet (www.wordnet.princeton.edu).