УДК 004

А. Г. Кумпель – магистрант кафедры компьютерной математики и программирования
В. В. Бураков (д-р техн. наук) – научный руководитель

ГЕНЕРАЦИЯ КЛЮЧЕВЫХ СЛОВ В «СИСТЕМЕ ОЦЕНКИ СТЕПЕНИ ПОИСКОВОЙ ОПТИМИЗАЦИИ»

По мере развития всемирной сети Интернет появилась необходимость в обеспечении быстрого и качественного поиска информации внутри нее. Были разработаны поисковые машины, назначением которых являлось облегчение этой задачи. Но одновременно с ними образовалась отрасль поисковой оптимизации, одной из негативных черт которой стало использование ее как препятствия для работы систем поиска с целью «нечестного» продвижения своего сайта выше в результатах поиска. Старый механизм указания ключевых слов в заголовках страниц уже не предоставлял достоверной информации, и возникла необходимость непосредственного анализа содержимого страницы для ее последующего индексирования.

Одним из методов (и наиболее действенным) стало использование некоторых закономерностей, подмеченных Джорджем Ципфом (George K. Zipf), которые он опубликовал в 1949 году. Пять лет спустя знаменитый математик Беноит Мандельброт (Benoit Mandelbrot) внес небольшие изменения в формулы Ципфа, добившись более точного соответствия теории практике. Хотя некоторые исследователи и подвергают исследования Ципфа острой критике, без учета подмеченных им закономерностей сегодня не способна работать ни одна система автоматического поиска информации.

Ципф заметил, что длинные слова встречаются в тексте реже, чем короткие. На основе этой закономерности Ципф вывел два закона.

Первый из них связывает частоту появления того или иного слова в каком-то тексте (она называется частота вхождения слова) с рангом этой частоты. Если к какому-либо достаточно большому тексту составить список всех используемых в нем слов, а затем расположить эти слова в порядке убывания частоты вхождения в данном тексте и пронумеровать в возрастающем порядке, то для любого слова произведение его порядкового номера в этом списке (ранга) и частоты его вхождения в тексте будет величиной постоянной

$$F * R = C,$$

где F – частота появления слова в тексте; R – ранг слова (наиболее часто употребляемое слово получает ранг 1, следующее – 2 и т.д.); C – константа.

Ципф экспериментально определил, что C ≈ 0,1 (что может изменяться в зависимости от языка, на котором написан текст). [1, 2] Графическое изображение закона Ципфа представлено на рис.1

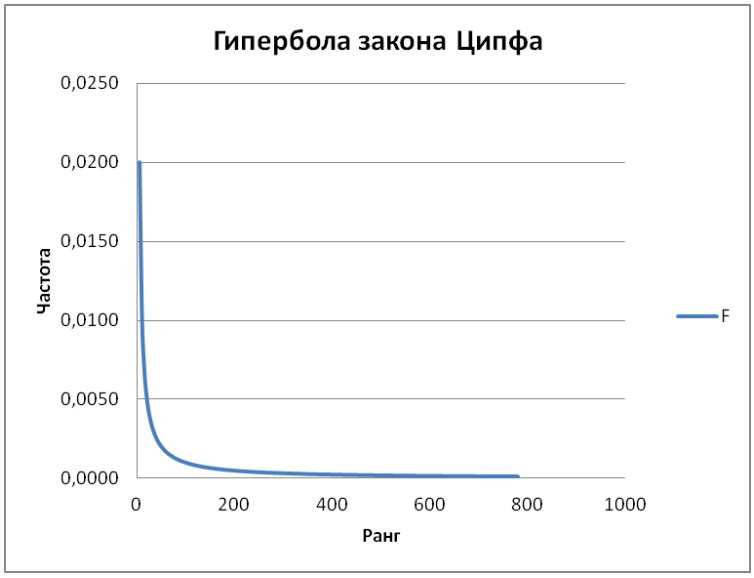

Рис. 1. Гипербола закона Ципфа

Если наиболее распространенное слово встречается в тексте 100 раз, то следующее по распространенности встретится не 99 и не 90, а примерно 50 раз (статистика не гарантирует точных цифр). Это также означает, что самое популярное слово в английском языке (the) употребляется в 10 раз чаще, чем слово, стоящее на десятом месте, в 100 раз чаще, чем сотое, и в 1000 раз чаще, чем тысячное.

Значение вышеупомянутой постоянной в разных языках различно, но внутри одной языковой группы она остается неизменной. Так, например, для английских текстов постоянная Ципфа равна приблизительно 0,1. Для русского языка постоянная Ципфа равна примерно 0,06 – 0,07 [1].

Мандельброт немного модифицировал формулу Ципфа:

$$F = C * R^{\frac{-1}{a}},$$

где a – коэффициент, характеризующий богатство словарного запаса; чем больше значение a, тем богаче словарный запас текста, поскольку кривая зависимости частоты появления каждого слова от его ранга убывает медленнее, и, например, редкие слова появляются чаще, чем при меньших значениях a.

Второй закон Ципфа констатирует, что частота и количество слов, входящих в текст с этой частотой, связаны между собой. Если построить график, отложив по одной оси (оси X) частоту вхождения слова, а по другой (оси Y) - количество слов, входящих в текст с данной частотой, то получившаяся кривая будет сохранять свои параметры для всех без исключения созданных человеком текстов.

Как поисковые машины могут использовать законы Ципфа?

Для того чтобы ответить на этот вопрос, вернемся к рис.1. Можно предположить, что наиболее значимые для текста слова лежат в средней части представленного графика. Слова, которые *встречаются слишком час*то, – это предлоги, местоимения и т.д. (в английском, немецком и некоторых других языках - еще и артикли).

Редко встречающиеся слова также в большинстве случаев не несут особого смыслового значения, хотя иногда, наоборот, весьма важны для текста (об этом будет сказано чуть ниже). Каждая поисковая система решает, какие слова отнести к наиболее значимым, по-своему, руководствуясь общим объемом текста, частотными словарями и т.п. Если к числу значимых слов будут отнесены слишком многие, важные термины будут забиты «шумом» случайных слов. Если диапазон значимых слов будет установлен слишком узким, за его пределами окажутся термины, несущие основную смысловую нагрузку.

Для того чтобы безошибочно сузить диапазон значимых слов, создается словарь «бесполезных» слов, так называемых *стоп-слов* (а словарь, соответственно, называется *стоп-лист).* Например, для английского текста стоп-словами станут артикли и предлоги the, a, an, in, to, of, and, that... и др. Для русского текста в стоп-лист могли бы быть включены все предлоги, частицы и личные местоимения: на, не, для, это, я, ты, он, она и др.

Исключение стоп-слов из индекса ведет к его существенному сокращению и повышению эффективности работы. Однако некоторые запросы, состоящие только из стоп-слов (типа «to be or not to be»), в этих случаях уже не пройдут. Неудобство вызывают и некоторые случаи полисемии (многозначности слова в зависимости от контекста). Например, в одних случаях английское слово «can» как вспомогательный глагол должно быть включено в список стоп-слов, однако как существительное оно часто несет большую содержательную нагрузку.

Процедура оптимального выбора ключевых слов, основанная на применении законов Ципфа, заключается в следующем: берут любой текст-источник, близкий к искомой теме, то есть «образец», и анализируют его, выделяя значимые слова. В качестве текста-источника может служить книга, статья, web-страница, любой другой документ. Анализ текста производится в следующем порядке:

1) стоп-слова удаляются из текста;

2) вычисляется частота вхождения каждого слова и составляется список, в котором слова расположены в порядке убывания их частоты;

3) выбирается диапазон частот, лежащий в середине списка, и из него отбираются слова, наиболее полно соответствующие смыслу текста;

4) составляется запрос к поисковой машине в форме перечисления отобранных таким образом ключевых слов, связанных логическим оператором OR(ИЛИ) Запрос в таком виде позволяет обнаружить тексты, в которых встречается хотя бы одно из перечисленных слов.

Отмеченные выше закономерности были использованы для проектирования и разработки «Системы оценки степени поисковой оптимизации», дополнительными возможностями которой кроме формирования таблицы ключевых слов являются:

- определение текущих индексов цитирования страницы;
- разбор информации meta-тегов, используемых поисковыми системами (meta-description и meta-keywords);
- выделение ссылочной массы;
- расчет плотности ключевых слов;
- сопоставление результатов анализа с допустимыми параметрами и представление результатов в графическом виде.

Система представляет собой web-приложение на “тонком клиенте”. Все вычисления производятся не серверной стороне.

Библиографический список

1. Alexander Gelbukh, Grigori Sidorov. *Zipf and Heaps Laws' Coefficients Depend on Language*. Proc. CICLing-2001, Conference on Intelligent Text Processing and Computational Linguistics, February 18–24, 2001, Mexico City. Lecture Notes in Computer Science N 2004, ISSN 0302-9743, ISBN 3-540-41687-0, Springer-Verlag, pp. 332–335. – URL: http://www.gelbukh.com/CV/Publications/2001/CICLing-2001-Zipf.htm
2. *K. E. Kechedzhy, O.V. Usatenko, V. A. Yampol'skii* - Rank distributions of words in additive many-step Markov chains and the Zipf law = Arxiv LANL. — 2004.; Phys. Rev. E. – 2005. – V. 72. – P. 046138(1)–046138(6).