ОПТИМИЗАЦИЯ WEB-СТРАНИЦ ДЛЯ ПОИСКОВЫХ СИСТЕМ В INTERNET

Ас. каф. ИИ Паю Сергей Владимирович, ст. ИСПР-04-1 Панасенко Вячеслав Юрьевич.
Харьковский национальный университет радиоэлектроники
61166, Харьков, пр. Ленина, каф. Искусственного интеллекта, тел. (057) 702-13-06
E-mail: usverster@gmail.com

The given work is devoted to the modern researches in the field of search engines optimization and intellectual analysis of text information. This work describes modern tendencies in web optimization and promoting web-sites. Research of search engines methods and algorithms, development of algorithm for intellectual analysis of text information on html-pages.

На данный момент сеть Интернет набрала огромный темп развития, что связано со множеством аспектов. Одним из важнейших показателей развития сети является количество ее основных элементов — веб-сайтов. В мае 2008 года размер глобальной сети Интернет достиг более150 000 000 веб-сайтов. Соответственно, количество страниц составляет число еще в десятки раз большее.

Проблема поиска в сети Интернет является одной из самых актуальных в области информационных технологий. Огромное количество существующих веб-сайтов, большой и все время увеличивающийся темп расширения сети (как увеличение количества пользователей, так и количества веб-сайтов) усложняют процесс поиска информации в Интернет. Таким образом в современной сети Интернет поиск стал не обычной процедурой, а неким видом искусства. Как в любом другом искусстве здесь так же тяжело добиться качественного результата, а тем более — массового.

Решение этой задачи частично ложится на поисковые системы (движки). Разрабатываются новые методики, усложняются алгоритмы индексирования страниц, расширяется база веб-страниц. Среди многообразия поисковых движков и систем передовые позиции занимают всего две крупных компании — Яндекс.Ру (www.yandex.ru) и Google (www.google.com). Первая является лидером среди поисковых сервисов в русскоязычном сегменте сети Интернет, а поисковая система компании Google считается лидером мирового масштаба.

Основной принцип функционирования поисковых систем (движков) заключается в анализе всего многообразия сайтов и сопоставлении каждому из них соответствующих ключевых запросов (слов, фраз). Для этого в первую очередь служит анализ html-заголовков каждой страницы. Основным пунктом при анализе html- заголовка является проверка тега <meta> и списка слов в атрибуте “keywords” (ключевые слова).

Оптимизация сайтов под поисковые движки занимает одно из основных мест в процессе поддержки и продвижения сайта. Этот факт объясняется тем, что в наше время Интернет стал не просто сетью с большим скоплением информации, а огромной рыночной площадкой, где любой желающий может разместить свой «товар». Многие предприниматели, почти все предприятия и каждая уважающая себя компания имеет собственного представителя в Интернет — свой веб-сайт. Разные по сложности, дизайну и функциям — все они преследуют одну главную цель — заявить о себе как можно большему количеству пользователей, которые являются потенциальными клиентами.

По статистике на большинство запросов пользователя поисковый сервис выдает более 10 страниц с результатами. Рядовой пользователь редко заходит дальше второй-третьей страницы, поэтому становится очевидной необходимость занимать высокие позиции в рейтинге поисковых систем, для успешной работы бизнеса.

Существует много методов оптимизации, продвижения и раскрутки сайтов, но одним из важнейших этапов является составление списка ключевых слов характерных, как для сайта в целом, так и для отдельных его страниц в частности. Именно ключевые

слова ищет поисковая система и именно их необходимо выделить для качественного продвижения сайта в рейтинге поисковой системы.

В данной работе рассматривается создание алгоритма построения списка ключевых слов для заданной веб-страницы. Алгоритм должен быть разработан с учетом современных тенденций в области оптимизации для поисковых систем, алгоритмов функционирования поисковых систем, поддерживать разные языки и обладать гибкостью и устойчивостью к шумам.

Алгоритм оптимизации веб-страницы для поисковой системы разрабатывался на основе поставленной задачи и в ходе разработки дополнялся для повышения эффективности работы и соответствия результатов реальному тексту веб-страницы.

Основной принцип функционирования поисковых систем (движков) заключается в анализе всего многообразия сайтов и сопоставлении каждому из них соответствующих ключевых запросов (слов, фраз). Для этого в первую очередь служит анализ html-заголовков каждой страницы. Основным пунктом при анализе html- заголовка является проверка тега <meta> и списка слов в атрибуте “keywords” (ключевые слова).

Для правильной работы алгоритма необходимо определить критерии по которым будут оцениваться слова, содержащиеся в обрабатываемом документе (веб-странице). В качестве максимально простого, удобного и понятного критерия был выбран «вес» слова в тексте. По мере работы алгоритма веса будут изменяться с целью достижения максимально точного результата.

Первоначально на вход алгоритм получает веб-страницу в формате HTML. Необходимо получить из входного документа полный текст этой страницы с учетом всех тэгов присутствующих в этом документе. На основании проведенных исследований была разработана таблица тэг – вес (Таблица 1), необходимая для правильной работы алгоритма, учета значимости слов и соответствия выходных данных реальной информации.

Таблица 1 — Соответствие между тэгом и весом слова

Тэги	Вес
<title>	90
<meta>, <dfn>, <h1>	40
<h2>	35
<h3>	30
<h4>	25
<h5>, <big>	20
<h6>	15
<b>, <u>, <i>, , <plaintext>	10
<a>,	5
<small>	-5
<code>, <script>, <applet>, <s>, <strike>, <noscript>	-10

После первого прохода мы получаем пары <текст> - <вес>. Для очистки и разбиения на пары применяется морфологический анализатор. В результате получается список из пар: слово в начальной форме и его вес. Для того чтобы список был наиболее

полным применяется словарь аббревиатур — все сокращения встречающиеся в тексте расшифровываются. Потом используется словарь стоп-слов, для очистки списка от слов, которые при индексации поисковой системой не учитываются. Затем происходит сворачивание списка с суммированием весов с целью получить список из уникальных слов. Для понижения веса часто встречаемых слов их вес делится на соответствующий коэффициент из частотного словаря (Частотный словарь — это такой вид словаря, в котором слова характеризуются с точки зрения степени их употребительности в совокупности текстов, представительных либо для языка в целом, либо для отдельного функционального стиля, либо для одного автора.). Результирующий список обрабатывается с помощью словаря синонимов для повышения весов слов попавших в вершину списка (при условии, что у них есть синонимы в списке). На выходе получается список из пар – ключевое_слово — вес.

В качестве недостатков, выявленных в результате тестирования на страницах-примерах, можно отметить то, что при анализе и составлении списка ключевых слов для художественных произведений первые места в списках занимают имена собственные. Это обусловлено тем, что имена собственные не содержатся в частотном словаре, используемом в программе. Так же в результатах могут встречаться слова, которых в тексте нет. Это обусловлено особенностью морфологического анализатора — если он не находит начальную форму слова, то выдает предположение о начальной форме такого слова.

Данный алгоритм реализован в приложении Web_optimizer на языке Java. Разработанное приложение использует внешнее приложение mystem (морфологический анализатор) от фирмы Яндекс. Используются словари стоп-слов, сокращений и аббревиатур, частотный словарь (английские и русские) и русский словарь синонимов. Все словари в хранятся в файлах формата TXT и подключаются на этапе выполнения программы.

В процессе выполнения работы был произведен анализ и рассмотрение основных проблем в предметной области оптимизации веб-страниц для поисковых систем в Интернет, поиск существующих методов решения проблем и разработка новых.

Была рассмотрена необходимость и целесообразность решения изученных проблем с учетом современных направлений и тенденций в области развития информационных технологий в целом и сети Интернет в частности.

Результатом стала разработка алгоритма, позволяющего бороться с описанными проблемами и решающего поставленные задачи. На основе алгоритма был разработан программный продукт Web_optimizer, который имеет ряд положительных сторон, но не лишен недостатков. Для качественной работы алгоритма были тщательно подобраны различные вспомогательные словари, а так же использованы уже существующие прикладные пакеты и программы.

Исследования результатов работы программы Web_optimizer показали, что разработанный алгоритм работает правильно и качественно. Для повышения релевантности результатов необходимо дальнейшее расширение и дополнение используемых словарей.

В дальнейшем планируется расширение программного продукта и усовершенствование разработанного алгоритма для составления ключевых фраз, которые бы максимально соответствовали возможным запросам пользователей для некоторого сайта.