МАРКЕТИНГ

Овчинников Степан Александрович,
к. т. н., ст. преподаватель кафедры программного
обеспечения автоматизированных систем,
e-mail: stepan@volgograd.ru
Белков Сергей Владимирович,
магистрант кафедры программного обеспечения
автоматизированных систем
Волгоградского государственного технического университета,
e-mail: serg-vlg@ya.ru

РОЛЬ ОПРЕДЕЛЕНИЯ ТЕМАТИКИ ВЕБ-САЙТА ДЛЯ ПОИСКОВОЙ ОПТИМИЗАЦИИ ВЕБ-САЙТА БИЗНЕСА В РОССИЙСКОМ СЕГМЕНТЕ СЕТИ ИНТЕРНЕТ

Role of defining the web-site topics for search engine optimization of business web-site in the Russian segment of the Internet

В статье приведены обзор поисковой оптимизации веб-сайтов и факты, доказывающие ее эффективность для повышения конкурентоспособности бизнеса. Рассмотрены основные факторы ранжирования веб-сайтов поисковыми системами, полученные в ходе проведенной научно-исследовательской работы. Предложен способ решения задачи определения тематики веб-сайта. Приведена архитектура разработанной автоматизированной системы определения тематики веб-сайта, а также основные этапы работы алгоритма определения тематики веб-сайта.

Практическая значимость. В работе дано решение актуальной прикладной задачи.

Апробация работы. Разработанное решение прошло тестирование и апробацию на ряде веб-сайтов.

The overview of search engine optimization of websites has been given in the article. The article presents a method of solving the problem of determining the subject site and the facts, showing evidence of its effectiveness to improve business competitiveness. The main factors of Web sites' ranking by search engines were considered in this article. Those ones were got as a result of the scientific research. The author describes the process of developing an automated system to determine the subject of the web site and the main stages of the algorithm to determine the subject of the web site.

The practical significance. The solution of important applied problem has been considered in this article.

The validation of work. The developed solution has been tested and validated on a number of websites.

Ключевые слова: определение тематики, поисковая оптимизация, поисковое продвижение, интернет-реклама, целевой трафик, классификация объектов, факторы ранжирования, повышение конкурентоспособности, онлайн-бизнес, seo.

Keywords: identifying theme, search engine optimization, search process, internet advertising, targeted traffic, data classification, ranking factors, improving competitiveness, online business, seo.

Бизнес все активнее использует веб-технологии. Неоспоримым фактом является бурный рост пользователей Интернета и объемов средств, направляемых на интернет-рекламу и интернет-маркетинг. Веб-сайт для фирмы стал важным атрибутом успешной деятельности, так как он не только расширяет возможности клиента, ищущего услуги или товар, но и позволяет фирме заявить о себе наибольшему числу потенциальных клиентов. По итогам 2009 года статистикой был отмечен рост количества пользователей, приходящих на сайт именно через поисковые системы[1]. Для создания посещаемого веб-сайта недостаточно сделать его красивым и удобным с точки зрения пользователя, необходимо принять ряд мер поисковой оптимизации, направленных на то, чтобы сайт можно было найти в лидирующих поисковых системах Рунета по целевым запросам. Стоит отметить, что зачастую пользователи просматривают только первые страницы поисковой выдачи, что приводит к конкуренции между веб-сайтами в популярных тематиках. Этим и обусловлена востребованность поисковой оптимизации на рынке услуг. Многие бизнесмены понимают, что интернет-реклама способна повысить конкурентоспособность их бизнеса, и поэтому созданию веб-сайта и его поисковому продвижению отводят одну из ключевых ролей.

Под поисковой оптимизацией веб-сайтов понимается изменение HTML-кода, текста, структуры и внешних факторов сайта с целью повышения его позиций в выдаче поисковой системы. Одним из ее методов является увеличение количества и качества обратных ссылок (внешний фактор ранжирования). Данный метод в большинстве случаев использует способ размещения платных ссылок на других веб-сайтах, что является с точки зрения поисковых систем искусственным повышением рейтинга сайта, страниц сайта в индексах информационно-поисковых систем. Поисковые системы совершенствуют алгоритмы определения платных ссылок. В частности, сотрудники поисковой системы Яндекс в статье «Использование методов категоризации текстовых привязок и анализа графов для идентификации платных ссылок» показали часть алгоритма идентификации оп-

[1] Liveinternet – статистика. [Электронный ресурс]. – Режим доступа: www.liveinternet.ru/stat/ru/searches.html?slice=ru (дата обращения 25.01.2010)

тимизированных тематических запросов[2]. Из данного алгоритма следует, что учитывается тематика веб-сайта, на котором размещается ссылка.

Задача определения тематики сайта является востребованной как в области информационного поиска, так и в области поисковой оптимизации веб-сайтов, что доказывает актуальность и значимость исследований в области классификации объектов на основе тематической близости текстового признака.

Классификация объектов является одним из развивающихся направлений в области интеллектуальных систем[3]. Информация в виде текстов на большинстве сайтов представлена на естественном языке, но стоит отметить необходимость ее отделения от html-разметки. Под задачей классификации понимается формализованная задача, в которой имеется множество объектов, которые разделены на классы. В настоящий момент проводится значительное количество исследований в области классификации данных[4], в области поисковой оптимизации веб-сайтов: Белов А. А., Волович М. М. «Автоматическое распознавание тематики сверхкоротких текстов». В основе большинства методов классификации лежит извлечение метаданных из текстов.

В ходе проведенных исследований факторов, влияющих на ранжирование веб-сайтов, в рамках научно-исследовательской работы было установлено, что тематика веб-сайта играет важную роль и непосредственно относится к внешним (ссылочным) факторам. Это также подтверждено сотрудниками поисковой системы Яндекс[5]. На рисунке 1 приведена обобщенная схема факторов ранжирования веб-сайтов.

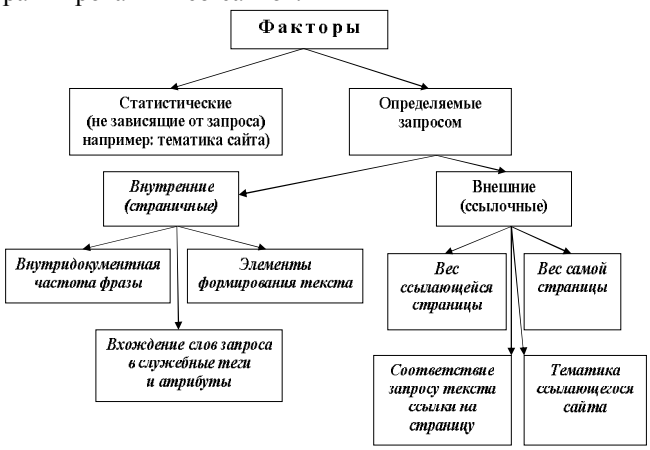

Рис. 1. Факторы ранжирования веб-сайтов

Был поставлен следующий эксперимент:

Дано: 2 группы по 10 веб-сайтов тематики «Образование». Домены без истории (ранее не регистрировались), на веб-сайтах отсутствует внешняя ссылочная масса.

Цель: сформировав внешнюю ссылочную массу, установить влияние «тематики», как одной из составляющих факторов ранжирования.

Суть: на первую группу на ссылочных биржах были закуплены ссылки с веб-сайтов различной тематики, на вторую – только с веб-страниц, имеющих тематику «Образование».

Результат: экспериментальным путем на группе сайтов было получено, что если рассматривать только фактор «тематика веб-сайта», при этом приняв другие факторы фиксированными, можно установить, что полное соответствие тематик: ссылающего сайта, веб-страницы ссылающегося сайта текста ссылки и сайта-акцептора – будет максимально влиять на соответствие запросу веб-сайта.

В рамках научно-исследовательской работы был разработан алгоритм определения тематики текста страницы веб-сайта по каталогу с фасетной рубрикацией. Схема основных этапов алгоритма определения тематики, приведена на рисунке 2.

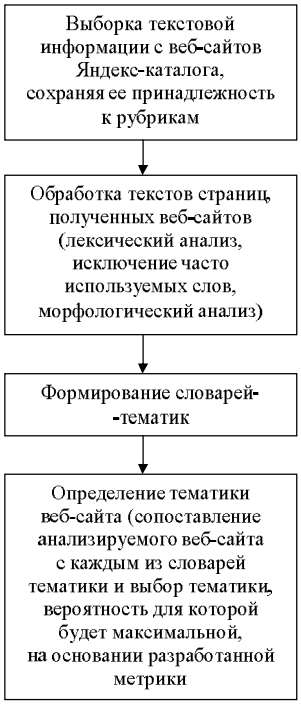

Рис. 2. Этапы определения тематики веб-сайта

В разработанный прототип программного средства вошли следующие модули: модуль обработки Яндекс-каталога, модуль обработки текста страниц веб-сайтов, модуль, формирующий словари тематик, модуль-классификатор выбранного сайта, определяющий его тематику.

Для построения словарей тематик был выбран Яндекс-каталог, имеющий более 700 рубрик и 121940 веб-сайтов. Все сайты, находящиеся в каталоге, классифицируются вручную экспертами специальной службы компании Яндекса. Формально рубрика или группа рубрик каталога определяют тематику отнесенных к ним веб-сайтов. Разработанный модуль обработки Яндекс каталога фактически реализует функционал робота, выкачивающего исходные тексты веб-сайтов из каталога, при этом сохраняя их принадлежность к рубрике. Работу модуля обработки текста страниц веб-сайтов можно разделить на следующие этапы: лексический анализ, ис-

[2] Конференция WWW 2009 в Мадриде. [Электронный ресурс]. – Режим доступа: http://www2009.org/proceedings/pdf/p1105.pdf (дата обращения 03.04.2010)

[3] Поспелов Г. С. Искусственный интеллект – прикладные системы / Г. С. Поспелов, Д. А. Поспелов. – М. : Знание, 1985. – 48 с.

[4] Боридко В. С. Извлечение, очистка и загрузка данных из первичных информационных систем в централизованное хранилище данных / В. С. Боридко, К. Ю. Колыбанов // Об организации системы мониторинга материально-технического оснащения учреждений профессионального образования : сб. трудов научно-практической конференции. – СПб. : Изд-во МИТХТ им. М. В. Ломоносова, 2005.

[5] Конференция WWW 2009 в Мадриде. [Электронный ресурс]. – Режим доступа: http://www2009.org/proceedings/pdf/p1105.pdf (дата обращения 03.04.2010)

ключение часто используемых слов и морфологический анализ.

Модуль автоматической тематической классификации выполняет сопоставление анализируемого сайта с полученными ранее словарями тематик и выбирает наиболее подходящую в соответствии с вычисленными вероятностями.

На рисунке 3 приведена архитектура разработанного прототипа программного средства.

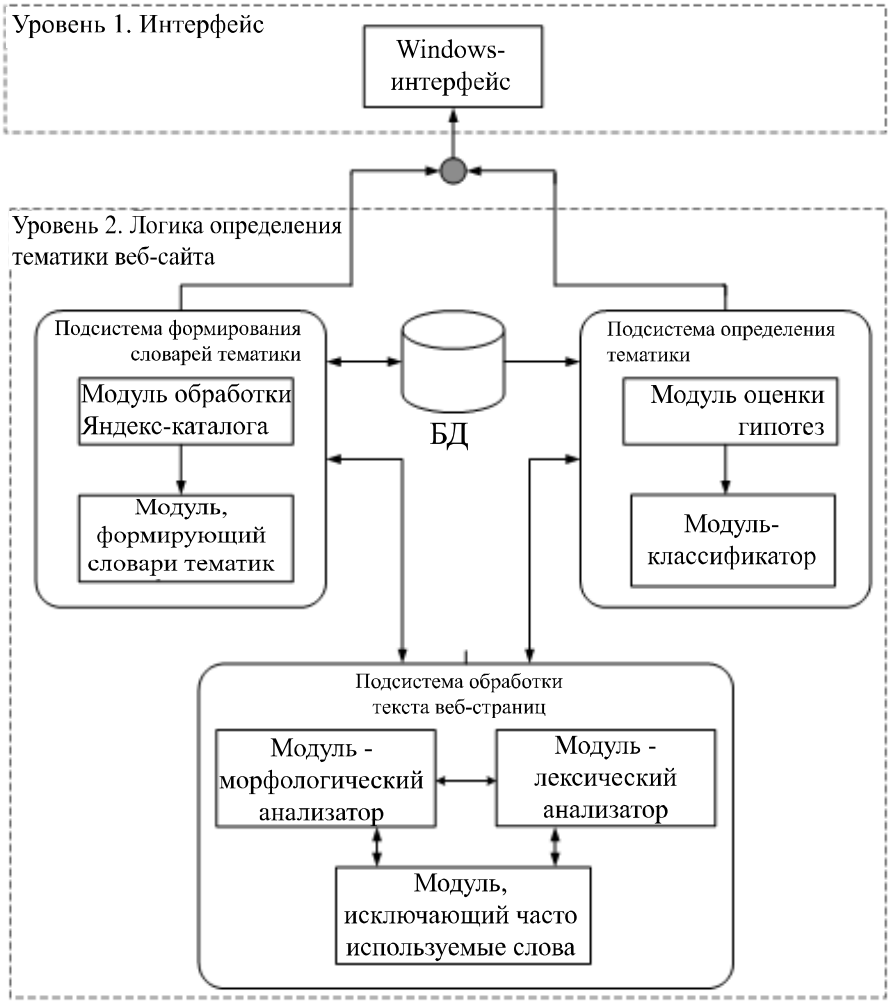

Рис. 3. Архитектура системы определения тематики веб-сайта

Таким образом, в ходе научно-исследовательской работы были поставлены и решены следующие задачи:

1. Проведено исследование методов, задач и подходов классификации.

2. Проведено исследование общей модели ранжирования веб-сайтов поисковыми системами.

3. Разработаны и реализованы алгоритмы в программном комплексе автоматической системы определения тематики веб-сайтов.

4. Подготовлена заявка в федеральный орган исполнительной власти по интеллектуальной собственности – на государственную регистрацию программы для ЭВМ.

ЛИТЕРАТУРА:

1. Liveinternet – статистика. [Электронный ресурс]. – Режим доступа: www.liveinternet.ru/stat/ru/searches.html?slice=ru (дата обращения 25.01.2010)

2. Конференция WWW 2009 в Мадриде. [Электронный ресурс]. – Режим доступа: http://www2009.org/proceedings/pdf/p1105.pdf (дата обращения 03.04.2010)

3. Поспелов Г. С. Искусственный интеллект – прикладные системы / Г. С. Поспелов, Д. А. Поспелов. – М. : Знание, 1985. – 48 с.

4. Боридко В. С. Извлечение, очистка и загрузка данных из первичных информационных систем в централизованное хранилище данных / В. С. Боридко, К. Ю. Колыбанов // Об организации системы мониторинга материально-технического оснащения учреждений профессионального образования : сб. трудов научно-практической конференции. – СПб. : Изд-во МИТХТ им. М. В. Ломоносова, 2005.

5. Автоматическое распознавание тематики сверхкоротких текстов. [Электронный ресурс]. – Режим доступа: http://www.dialog-21.ru/dialog2007/materials/html/05.htm (дата обращения 03.04.2010)

6. Городецкий Б. Ю. Компьютерная лингвистика : моделирование языкового общения (Вступительная статья) / Б. Ю. Городецкий // Новое в зарубежной литературе. – М. : Прогресс. Вып. XXIV (Компьютерная лингвистика). – С. 5-31.

7. Поляков А. А. Стандартизация и сертификация средств информационных технологий в сфере образования / А. А. Поляков [и др.] // Индустрия образования : сб. статей. Выпуск 1 ; Минобразование России, ГОСНИИСИ. – М., 2001. – С. 148-152.