Создайте образ

Creating an Image

Створіть образ

Васильев В. В., Хливненко Л. В.
Воронежский государственный университет, Воронеж, Россия

Сороколетова Н. В.
Белгородская государственная универсальная научная библиотека, Белгород, Россия

Valery Vasiliev and Lyubov Khlivnenko
Voronezh State University, Voronezh, Russia

Nataliya Sorokoletova
Belgorod State Universal Scientific Library, Belgorod, Russia

Васильєв В. В., Хливненко Л. В.
Воронезький державний університет, Воронеж, Росія

Сороколєтова Н. В.
Бєлгородська державна універсальна наукова бібліотека, Бєлгород, Росія

Дается краткая характеристика современному состоянию подходов пользователя Интернет, разработчиков ИПС и авторов электронных публикаций к формированию поискового образа и поискового запроса документов. На основе проведенного анализа, предлагается вариант повышения эффективности выполнения пользовательских запросов, реализованный в Белгородской государственной универсальной научной библиотеке.

Briefly described are the state-of-the-art approaches of Internet users, developers of information retrieval systems and authors of e-publications to the formation of a retrieval image and document retrieval request. Based on the analysis data, the authors propose a method of enhancing the efficiency of performing user requests, which has been implemented in Belgorod State Universal Scientific Library.

Подається коротка характеристика сучасного стану підходів користувача Інтернет, розробників ІПС та авторів електронних публікацій до формування пошукового образу та пошукового запиту документів. На основі проведеного аналізу запропоновано варіант підвищення ефективності виконання запитів користувачів, який реалізовано в Бєлгородській державній універсальній науковій бібліотеці.

В поисках необходимой информации можно провести много времени и так и не найти её. С увеличением объема информации проблема поиска не исчезла, а наоборот осложнилась. Казалось бы, получив «сверхумный» инструмент – компьютер, люди должны были бы вздохнуть спокойно, но слова Д. Бриллюэна емко определяют сегодняшнюю ситуацию: «Вам нужна какая-то информация? Нажмите кнопку – и машина ответит… Нелепость! Подобное представление вырастает из убогой научно-популярной беллетристики». Давайте проанализируем различные подходы к поиску и подготовке к поиску информации в глобальной информационной сети – Интернет.

По отношению к систематизации и поиску информации можно выделить три основных подхода: подход пользователя, подход разработчика информационно-поисковой системы и подход автора ВЕБ-документа.

Пользователь ставит перед собой цель найти нужную ему информацию. Более точно эта цель определяется как поиск не просто релевантных документов, т.е. соответствующих запросу, а как поиск документов действительно необходимых пользователю, т.е. пертинентных. И чем больше список релевантных документов и меньше в них пертинентных, тем сильнее раздражается пользователь.

Снизить раздражение пользователей взялись разработчики информационно-поисковых систем (ИПС). Которые постоянно стремятся улучшить системы. Степень эффективности ИПС чаще всего определяется двумя показателями – коэффициентом точности, который определяют как отношение числа релевантных документов в выдаче к общему объёму выдачи, и коэффициентом полноты –

отношение числа релевантных документов в выдаче к общему числу релевантных документов в массиве ИПС. Неэффективность работы ИПС можно охарактеризовать коэффициентом шума (шумом или информационным шумом) – это множество документов в выдаче, не соответствующих запросу. ИПС опирается на базу данных о документах, находящихся в зонах действия ИПС. База данных формируется с помощью специальных программ-роботов, которые занимаются сбором и автоматическим индексированием документов. Под индексированием понимают составление списка ключевых слов для документа и сохранения этого списка (поискового образа документа) и связи с самим документом в базе индекса. Для индексирования используются различные алгоритмы. Основными типами индексирования являются координатное и фактографическое. В первом случае документ описывается в целом, а во втором индексируются конкретные факты, как отдельные поисковые единицы. Механизм индексирования работает в двух направлениях – с одной стороны собирает документы, с другой выдает их по запросу пользователя. Координатное индексирование задает частотный поиск по ключевым словам. Каждый документ представляется в базе индекса набором наиболее часто встречающихся терминов, которые составляют поисковый образ документа. Запрос пользователя преобразуется к тому же виду. При этом выбрасываются «стоп-слова», т.е. слова, вручную занесенные в базу данных как запрещенные, оставшиеся слова приводят к норме с помощью морфологических анализаторов. Существует множество модификаций частотного поиска, использующие дополнительные параметры текста для уточнения запроса и словари синонимов. Но в случае частотного поиска документы и запросы представляются векторами терминов, а смысловые и синтаксические связи между терминами игнорируются, что ведет к увеличению шума. Используя существующие инструменты автоматического индексирования, в современную модель поисковой машины вместо алгоритмов на базе частотной модели внедряются методы более гибкого анализа текста, которые допускают учет семантики обрабатываемого текста. Разработчики ИПС постоянно совершенствуют не только алгоритмы индексирования, но и видоизменяют информационно-поисковый язык, на котором пользователь может задавать запросы. Идеальный результат работы ИПС – это полное совпадение поискового образа документа в индексной базе с поисковым запросом пользователя, причем не формального его выражения, а мыслеобраза. Очевидно, что это не достижимо. Поэтому ИПС используют формальный показатель соответствующий 25-60% совпадению ПОД с ПОЗ.

Фактически, и пользователи, и разработчики поисковых систем изначально имеют дело с авторским документом. Чаще всего автор не ставит перед собой цель найти свой документ. Автор желает самовыразиться или опубликоваться. На самом деле степень информационного шума могла бы быть значительно ниже, если бы авторы чаще оформляли документы, учитывая особенности формирования ПОД и ПОЗ. Так как большинство публикуемых документов имеют html-формат представления, то не лишним будет напомнить, в каких тегах выбирается роботами ключевая информация для создания ПОД. В первую очередь индексируются сведения из заголовка страницы, расположенные в теге <title>. Следующие ключевые слова выбираются из тега <meta name=«keywords» http-equiv=«keywords» content=«список ключевых слов»>. Причем роботы отбрасывают из списка те слова, которые были найдены в <title>. Нет смысла их дублировать. Далее индексируются тексты заголовков <H1>, <H2> и в последнюю очередь анализируется полный текст документа. Это должен учитывать автор и по возможности нормировать формулировку ключевых слов и заголовков, а также не засорять их «стоп-словами» и широкоупотребляемыми словами. Для корректного отображения аннотации на авторский документ в поисковом списке, нужно заполнять тег <meta name=«description» http-equiv=«description» content=«аннотация»>. В случае отсутствия такого описания в качестве аннотации будет взят первый абзац страницы. Полезно заполнять теги <meta http-equiv=«Refresh» content=«n; url=URL»>, где n – секунды до перезагрузки документа, и <meta http-equiv=«Content-type» content=«text/html; charset=windows-1251»>, позволяющий указывать кодировку текста.

Поставим простой эксперимент: посмотрим на состояние указанных выше тегов для WEB-страниц, указанных в списке литературы к данной статье. «Плюсом» будем отмечать тот элемент старицы, который оформлен корректно, «минусом» – отсутствие исследуемого тега. Получим следующую таблицу:

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
title	+	+	-	+	+	+	+	+	+	-	+	+	+	+	+	+
keywords	-	-	-	-	-	-	-	-	-	-	+	-	+	-	-	+
H	-	-	+	+	+	+	-	+	-	-	-	-	-	-	+	-
description	-	-	-	-	-	-	-	-	-	-	+	-	+	-	-	+
Refresh	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Content-type	-	-	+	+	-	-	-	+	+	-	-	-	+	-	+	+

Вот такая картина. Остается только уповать на разработчиков ИПС, которые в очередной раз должны выдумать что-то из ряда вон выходящее, чтобы информация все-таки была найдена.

Поисковый образ документа можно регулировать не только с помощью html-тегов. Не секрет, что библиотечные специалисты создают метаинформацию (информацию об информации) на информационные ресурсы различного типа: печатные, электронные и другие. Создание метаинформации или библиографического описания подчиняется правилам государственного стандарта. Для библиотечных ИПС существует стандарт описания машиночитаемой библиографической записи, который позволяет к метаданным привязывать полные тексты документов. С появлением нового вида информации правила создания ее метаданных тоже меняются. В целом, библиотечные системы дают для пользователей Интернет ПОД высокого качества.

В Белгородской государственной универсальной научной библиотеке формирование полнотекстовых электронных ресурсов осуществляется с учетом особенностей требований пользователей, информационно-поисковых систем и государственных библиотечных стандартов. Каждый электронный текст сопровождается метаданными, созданными библиографами в АБИС «OPAC-Global», и метаданными, входящими в состав html-кода электронного документа. Таким образом, пользователь имеет возможность разыскать информацию, собранную в полнотекстовых базах данных единого информационного пространства библиотек Белгородской области (http://opac.bgunb.ru), как с помощью общедоступных поисковых систем Интернет, так и с помощью поискового аппарата автоматизированной библиотечной информационно-поисковой системы.

Подводя итоги сказанному, хочется отметить, что совершенствование результатов поиска информации зависит от состояния подхода и пользователя, и разработчиков ИПС, и авторов. Пользователь совершенствуется, повышая свою культуру поиска, изучая возможности ИПС. Разработчики изучают потребности пользователей и особенности авторов, видоизменяют алгоритмы формирования и установления взаимного соответствия ПОД и ПОЗ. Авторы прибегают к использованию программных и государственных стандартов и готовят документы для корректного индексирования. Так поступательно мы продвигаемся вперед. Невольно вспоминается высказывание И.-В. Гёте «Эти добрые люди и не подозревают, каких трудов и времени стоит научиться читать. Я сам на это употребил 80 лет и все не могу сказать про себя, чтобы вполне достиг цели».

Литература

1. Бельтикова, Н.В., Кузина, И.А., Храмцов, П.Б. Заголовк HTML-документа: [Электронный документ].- (http://webclass.polyn.kiae.su/ classes/head/pod.htm). Проверено 3.04.2006.
2. Березовский, А. М. Об одной модели формирования поискового образа документа и поискового образа запроса: проблемы и решения: [Электронный документ].- (http://www.gpntb.ru/libcom4/eng /index.cfm?n=tez/doc1/doc4). Проверено 3.04.2006.
3. Бороздин, Д.С., Коротков, А. Е., Попов, Ю.А. Проект: метапоисковая интеллектуальная машина: [Электронный документ].- (http:// molod.mephi.ru/Data/600.htm). Проверено 3.04.2006.
4. Введение в HTML: [Электронный документ].- (http://www.intuit.ru/department/internet/htmlintro/2/htmlintro_2.html). Проверено 3.04.2006.
5. ГОСТ 7.73-96 Поиск и распространение информации: [Электронный документ].- (http://www.minskcom.com/content/view/49/31/). Проверено 30.01.2006.

6. ГОСТ 7.66-92. Индексирование документов. Общие требования к координатному индексированию. – (Соотв. ИСО 5985-85). – Утв. 1992. – (Система стандартов по информации, библиотечному и издательскому делу): [Электронный документ].- (http://gsnti-norms.ru/norms/common/doc.asp?0&/norms/stands/7_66.htm). Проверено 3.04.2006.
7. Информационно-поисковые системы: [Электронный документ].- (http://www.ergeal.ru/archive/cs/ppo/1-4.htm). Проверено 3.04.2006.
8. Козлов, А.В. Системы поиска информации в Интернет-ресурсах: [Электронный документ].- (http://nit.miem.edu.ru/cgi-bin/article?id=57). Проверено 3.04.2006.
9. Коршунов, О. П. Библиографоведение. Общий курс. Основы теории библиографии: Учебник для библиотечно-информационных факультетов вузов: [Электронный документ].- (http://www.libs.ru/doc/corshunov/24.htm). Проверено 3.04.2006.
10. Куглер, В.М. Некоторые аспекты индексации и поиска документов на основе вложенных многоуровневых структур: [Электронный документ].- (http://www.dialog-21.ru/Archive/2003/Kugler.pdf). Проверено 3.04.2006.
11. Кузина, И. Новое поколение поисковых машин: [Электронный документ].- (http://www.osp.ru/cw/1997/32/91.htm). Проверено 3.04.2006.
12. Куклина, О.Г. Компьютерные технологии поиска документальной информации: [Электронный документ].- (http://www.rusnauka.com/Inftehn/18.html). Проверено 3.04.2006.
13. Куршев, Е. П., Осипов, Г. С., Рябков, О. В., Самбу, Е. И., Соловьева, Н. В., Трофимов, И. В. Интеллектуальная метапоисковая система: [Электронный документ].- (http://www.dialog-21.ru/archive_article.asp?param=7618). Проверено 30.01.2006.
14. Рубин, А. Поисковые системы для web-сервера: [Электронный документ].- (http://www.networkmagazine.ru/cgi-bin/materials.pl?issue=101999&article=74). Проверено 3.04.2006.
15. Чурсин, Н. Выход в автоматизации?: [Электронный документ].- (http://orel.rsl.ru/nettext/russian/chursin/populjarn_informat/ pi06.htm). Проверено 3.04.2006.
16. Щербак, С.С. Применение агентной технологии в поисковых информационных средах: [Электронный документ].– (http://www.ontolib.com/materials/publications/first/first.html). Проверено 30.01.2006.