УДК 519.95

ПРОБЛЕМА ПЕРТИНЕНТНОСТИ СОВРЕМЕННЫХ ИНФОРМАЦИОННО-ПОИСКОВЫХ СИСТЕМ

© С.Е. Жуликов, О.В. Жуликова

Ключевые слова: информационно-поисковая система; математическая модель; поиск; релевантность; индексация; пертинентность.

Рассматриваются проблемы организации поиска в информационных системах, в т. ч. системах Интернета. Наиболее значимой, кроме проблем индексации, обновления и борьбы со спамом, авторы считают оценку эффективности информационно-поисковых систем с точки зрения пертинентности результатов поиска.

Информационная система – это организационно-упорядоченная взаимосвязанная совокупность средств и методов информационных технологий, используемых для хранения, обработки и выдачи информации в интересах достижения поставленной цели. Такое понимание информационной системы предполагает использование в качестве основного технического средства переработки информации ЭВМ и средств связи, реализующих информационные процессы и выдачу информации, необходимой в процессе принятия решений задач из любой области.

Информационная система является средой, составляющими элементами которой являются компьютеры, компьютерные сети, программные продукты, базы данных, люди, различного рода технические и программные средства связи и т. д. Хотя сама идея информационных систем и некоторые принципы их организации возникли задолго до появления компьютеров, однако компьютеризация в десятки и сотни раз повысила эффективность информационных систем и расширила сферы их применения.

Информационно-поисковые системы ориентированы на решение задач по поиску информации, документа или факта во множестве источников информации (документов).

Для того чтобы осуществлять поиск документов, содержащихся в базе данных поисковых систем, используется математическая модель, позволяющая упростить процесс обнаружения нужных документов по введенному пользователем поисковому запросу и процесс определения релевантности всех найденных документов этому запросу. Чем больше документ соответствует данному запросу, чем он релевантнее, тем выше он должен стоять в поисковой выдаче.

Основная задача, выполняемая математической моделью любой поисковой системы, – это поиск документов, веб-страниц в своей базе обратных индексов, соответствующих данному поисковому запросу, и сортировка этих найденных документов в порядке убывания их релевантности поисковому запросу. Использование простой логической математической модели, когда документ будет являться найденным, если в нем встречается искомая фраза, не является корректным в силу огромного количества таких документов, выдаваемых на рассмотрение пользователю.

Поисковая система должна не только предоставить список всех документов, веб-страниц, на которых встречаются слова из поискового запроса. Она должна предоставить этот список документов в такой форме, когда в самом начале этого списка будут находиться наиболее соответствующие запросу пользователя документы, тем самым информационно-поисковая система должна осуществить сортировку найденных страниц по релевантности.

Неидеальностью любой математической модели поисковых систем и пользуются оптимизаторы, влияя теми или иными способами на ранжирование документов в поисковой выдаче зачастую в пользу продвигаемого ими сайта. Математическая модель, используемая всеми поисковыми системами, относится к классу векторных математических моделей. В этой математической модели используется такое понятие, как вес документа по отношению к заданному пользователем запросу.

В этой связи представляется интересным обсудить некоторые проблемы, возникающие при работе с информационными поисковыми системами (на примере информационно-поисковых систем Интернета).

Первая и основная проблема, которая встает перед любой поисковой системой, – это постоянно растущий размер индексной базы поисковых систем. Индексную базу нужно где-то хранить, а в связи с тем, что размер коллекций поисковых систем постоянно растет, то и места для ее хранения требует все больше и больше. Проблема эта будет стоять перед поисковыми системами всегда, и решать ее можно только за счет увеличения количества серверов в дата-центрах.

Например, Яндекс использует для хранения своей индексной базы на данный момент уже около десятка дата-центров по несколько тысяч серверов в каждом дата-центре. При этом Яндекс до недавнего времени индексировал только русскоязычный Интернет, и лишь сейчас он выходит на мировой уровень и начинает индексировать отличные от русского языка документы. Поисковая система Google занимает одну из лидирующих позиций по сборке компьютеров, в то время как

все эти компьютеры идут на личные нужды Google – их используют в дата-центрах Google для хранения индексной базы.

Второй основной проблемой, стоящей перед поисковыми системами, является борьба с дубликатами в поисковой выдаче. Ведь если выкинуть из индексной базы все дубликаты, то поисковая выдача от этого не ухудшится, а вот место, требуемое для хранения уменьшившейся индексной базы, позволит сэкономить немалые средства. Борются с дубликатами поисковые системы как с помощью их удаления из поисковой выдачи, так и превентивными мерами, предписывающими веб-мастерам самим бороться с дубликатами на своих собственных сайтах. Если веб-мастера будут игнорировать это требование поисковых систем, то к их проектам, возможно, будут применены различные санкции в виде наложения фильтров, вылета страниц ресурса из индекса и прочих репрессивных действий.

Еще одной проблемой, с которой довольно успешно борются современные поисковые системы, – это спам в поисковых выдачах. Такой спам попадает в поисковую выдачу при использовании веб-мастерами черных методов поисковой оптимизации, при которых спам попадает в топ поисковых выдач по каким-либо запросам, а при переходе на данные ссылки из поисковой выдачи пользователь отсылается на совершенно другой ресурс.

Также перед поисковыми системами стоит задача не только хранения постоянно расширяющейся индексной базы, но и проблема ее обновления, для того чтобы она соответствовала реальной действительности. Нужно не только индексировать новые документы в сети, но и обновлять индексы уже ранее проиндексированных документов.

Одна из глобальных проблем, стоящих перед поисковыми системами, – это удовлетворение информационных потребностей пользователя: понять, что он хочет увидеть в результатах поисковой выдачи, вводя тот или иной поисковый запрос. Понимание поисковыми системами намерений пользователя позволит поисковой системе сформировать наиболее подходящую для этого случая поисковую выдачу, тем самым удовлетворив запросы пользователя. Удовлетворенный пользователь опять вернется именно в эту поисковую систему, т. к. она хорошо «понимает», что он хотел получить в ответ на свой запрос.

Эта проблема актуализирует оценку результата поиска с точки зрения релевантности и пертинентности.

Существует множество определений релевантности. Например, ГОСТ 7.73-96 гласит: «релевантный: соответствие полученной информации информационному запросу» [1]. Таким образом, релевантность определяется исключительно математическими моделями поиска конкретной поисковой системы.

В том же ГОСТе говорится: «пертинентность; пертинентный: соответствие полученной информации информационной потребности» [1], т. е. пертинентность определяет степень соответствия между ожиданиями пользователя и результатами поиска.

Из-за субъективности пертинентности добиться точного совпадения нельзя: любая поисковая система настраивается на информационные нужды усредненного, а не конкретного пользователя. Для удовлетворения нужд конкретного пользователя и были придуманы разные способы коррекции результатов поиска по релевантности. Фактически, учет цитируемости документа в других документах является одним из таких способов.

Недостаточная пертинентность поисков обусловлена следующими основными причинами:

– излишняя декомпозиция запросов: информационная потребность пользователя раскрывалась в виде серии из 7–10 очень конкретных запросов;

– обслуживание по чрезмерно широким запросам: на один запрос абонент получает от сотен тысяч до сотен миллионов документов, веб-страниц, хотя непосредственно соответствует запросу только малая часть информации.

Существующие информационно-поисковые системы изначально проектировались для обеспечения именно релевантности, а не пертинентности выборки по отношению к формальным запросам, и в этом их главная слабость в современных условиях. Низкий, а точнее говоря, неконтролируемый уровень пертинентности выборки с высоким уровнем ее релевантности порождает различные ситуации, допускающие более или менее общую типизацию.

Например, по запросу «президент» можно получить, кроме необходимых новостей, рекламу отеля «Президент», прайс-листы сигарет «Президент» и т. п.

Степень эффективности информационно-поисковых систем может быть определена сравнением реальной действующей системы с идеальной моделью. Идеальная модель может быть определена (как это было сделано основоположником научно-технической информатики К. Муэрсом) так: «Это система, которая из документального фонда выдает ровно те и все те документы, которые бы отобрал сам пользователь, если бы он мог внимательно прочитать каждый из них» [2]. Наиболее популярны два отношения, это:

– *коэффициент точности* – отношение числа релевантных документов в выдаче к общему объему выдачи;

– *коэффициент полноты* – отношение числа релевантных документов в выдаче к общему числу релевантных документов в массиве.

Кроме того, используются другие коэффициенты. Так, множество, содержащее документы выдачи, не соответствующие запросу, называется шумом (информационный шум). Относительное количество шумовых документов в выдаче называется коэффициентом шума. Множество, содержащее релевантные документы, не выданные пользователю, называется потерями. Отношение числа «потерянных» документов к общему числу релевантных документов в массиве может быть названо *коэффициентом потерь*, или *коэффициентом молчания*. Коэффициенты потерь и шума не являются самостоятельными показателями эффективности поиска.

Очевидно, что чем выше коэффициенты полноты и точности, тем эффективность поиска выше. Следует заметить, что полнота и точность поиска зависит не только от поисковой системы, но и от типа запросов. По одним типам запросов система может проводить поиск лучше, а по другим – хуже. Также эффективность может зависеть от представления о реальной потребности в получении той или иной информации.

Накопленные к настоящему времени колоссальные объемы информации в совокупности с непрерывно увеличивающимися темпами ее роста определяют ак-

туальность и значимость исследований в области информационного поиска. Бурное развитие сетевых технологий, в т. ч. и Интернета, способствует значительному увеличению доступных информационных ресурсов и объемов передаваемой информации. Зачастую это разнородная, слабо структурированная и избыточная информация, обладающая высокой динамикой обновления.

Проблема низкой пертинентности информационного поиска обусловлена сложностью формализации информационной потребности пользователя в поисковый запрос, используемый поисковыми системами [3].

ЛИТЕРАТУРА

1. ГОСТ 7.73-96. Межгосударственный стандарт. Система стандартов по информации, библиотечному и издательскому делу. Поиск и распространение информации. Термины и определения. М.: ИПК Изд-во стандартов, 1997. 19 с.
2. *Федотов А.М.* Проблемы поиска информации: история и технологии. URL: http://www-sbras.nsc.ru/win/elbib/data/dinamo_cat/1.pdf. Загл. с экрана.
3. *Терехов А.А.* Разработка методов и инструментальных средств повышения пертинентности поиска в современных информационных средах: автореф. дис. ... канд. тех. наук. Рязань: Рязан. гос. радиотехн. ун-т, 2010. 16 с.

Поступила в редакцию 23 ноября 2012 г.

Zhulikov S.E., Zhulikova O.V. PERTINENCE PROBLEM OF MODERN INFORMATION SEARCH SYSTEMS

The problems of the organization of search in information systems, including the Internet, are considered. Except indexing, updating, and to combat spam problems the authors consider evaluation of the effectiveness of information retrieval systems in terms of pertinence results.

Key words: information search system; mathematical model; search, relevance; indexing; pertinence.