УДК 519.92+002.6

ОТ ТЕМАТИЧЕСКОГО ПОИСКА К ИНФОРМАЦИОННОМУ МОДЕЛИРОВАНИЮ

© В.В. Зубец

Ключевые слова: потребитель информации; информационный поиск; стратегии информационного моделирования области знания.

В работе дан краткий обзор технологий информационного поиска, используемых в настоящее время. Сформулированы основные принципы предлагаемой стратегии информационного моделирования области знания.

Развитие компьютерных технологий в последние десятилетия привели к тому, что слово «информационный» стало одним из самых популярных как в средствах массовой информации, так и в научных изданиях. Появилась концепция «информационного общества», идущего на смену постиндустриальному. Действительно, внедрение информационных и компьютерных технологий позволило значительно повысить производительность труда в развитых странах и, в конечном счете, улучшить качество жизни людей. Тем не менее, специалисты знают, что одна из основных проблем информатики так и не решена. Эта проблема заключается в разрешении противоречия между ограниченными возможностями отдельного человека по переработке информации и возрастающими потоками информации. Проблема обострилась в XX в., когда количество различных источников информации стало возрастать угрожающими темпами. Появление в конце XX в. Интернета еще более усугубило ситуацию, поскольку, наряду с ростом количества источников информации, обострилась проблема оценки их достоверности. Обозначенную проблему различные авторы пытались решать разработкой эффективных технологий информационного поиска. Что же из этого получилось?

Методы информационного поиска начали разрабатывать с момента появления первых документов, т. е. истории этого вопроса тысячи лет. Как только возникли первые библиотеки, возникла проблема поиска в них книг. В течение тысячелетий были разработаны и усовершенствованы методы такого поиска. Уточним, что далее речь будет идти о наиболее общем случае информационного поиска – тематическом поиске. В основе такого поиска лежит использование иерархических классификационных систем. Суть таких систем заключается в том, что вся Область знаний делится на крупные предметные области, которые, в свою очередь, подразделяются на более мелкие, те – на еще более мелкие и т. д. Каждой области знания и подобласти присваивается свой индекс. Возникает разветвленная древовидная система, позволяющая классифицировать все источники информации. Примерами таких систем являются международная универсальная десятичная классификация (УДК) и национальная библиотечно-библиографическая классификация (ББК). После того, как все источники информации проиндексированы, стратегия поиска заключается в том. что потребитель информации определяет предметную область, которая его интересует, просматривает соответствующий раздел каталога библиотеки, сначала выбирает релевантные документы, а затем среди них пертинентные документы и изучает их. На этом процесс поиска не останавливается. В процессе изучения документов потребитель уточняет свой запрос, т. е. уточняет границы предметной области, изучает библиографические списки, размещенные в документах, и вновь проводит поиск. Процесс осуществляется циклически до тех нор, пока не будут найдены и изучены все релевантные документы. Кроме систематического каталога, основанного на классификационных системах, появились предметные каталоги, построенные по алфавитному принципу. Они значительно расширили возможности поиска. Таким образом, в недрах библиографии возникла стройная система классификации и поиска документов [1–3]. Эта система логична и эффективна в том случае, если количество источников информации сравнительно невелико, как это было вплоть до XX в.

В XX в. произошел «документальный взрыв», т. е. бурный, многократный рост количества источников информации. Если в начале XX в. на удвоение общей суммы знаний требовалось 50 лет, то в конце XX в. – всего год [4]. Изменилась и структура размещения документов. Если раньше в развитых странах существовало небольшое количество библиотек, аккумулирующих почти все издаваемые в то время книги и журналы (их издавалось немного), то, в связи с бурным ростом количества изданий, система превратилась в распределенную, т. е. фонды большинства библиотек стали значительно различаться. Возникла необходимость поиска в фондах нескольких библиотек. Библиотекари пытались справиться с данной ситуацией, были созданы службы межбиблиотечного обмена (МБА), международного межбиблиотечного обмена (ММБА), централизованной библиотечной сети (ЦБС). Появились библиографические указатели, издания, классифицирующие и описывающие документы вне зависимости от места их хранения. Были созданы мощные исследовательские институты, такие как ВИНИТИ и ИНИОН, которые стали разрабатывать целую систему библиографических указателей. В результате количество указателей стало таким большим, что потребовался указа-

тель библиографических указателей. И это ведь только национальные указатели. Значит, нужен следующий уровень – международный указатель национальных указателей предметных указателей. И так далее. То есть мощные информационные центры более или менее справились с индексацией возросшего потока документов, но при этом для конкретного потребителя информации система стала очень сложной, просмотр толстых реферативных журналов и других указателей потребовал большого количества времени.

В связи с возникновением в середине XX в. электронной вычислительной техники и развитием автоматизированных поисковых систем появились новые технологии информационного поиска. Поиск релевантных документов стали автоматизировать с помощью ЭВМ, т. е. создавать информационно-поисковые системы (ИПС). Вычислительная машина просматривает документы с гигантской скоростью, экономя время потребителя информации. Начался бум ИПС. Были разработаны как теоретические основы ИПС, так и реальные практические системы [5–12]. Нельзя отрицать положительную роль этих разработок в ускорении научно-технического прогресса. Результаты этих разработок сейчас активно используются в информационных структурах всемирной компьютерной сети – Интернете. И все-таки, проблема осталась. Потоки документов выросли настолько, что компьютер стал заваливать потребителя сотнями и тысячами релевантных документов. Количество их всё растет.

В конце XX в. всемирная компьютерная сеть в принципе решила проблему быстрого доступа к документу. В условиях распределенной мировой системы хранения информации Интернет позволяет пользователю быстро и дешево получить доступ к электронной копии документа, не перемещаясь самому и не перемещая документ. Конечно, это значительно экономит время и средства. Но это в идеале. В реальности потребитель информации сталкивается с тем, что уровень шума при поиске в Интернете многократно возрастает по сравнению с поиском в библиотеках. Дело в том, что основной массив информации, размещенной в Интернете, не проходит никакой экспертизы на достоверность, в отличие от документов, содержащихся в фондах библиотек, где практически вся информация проходит ту или иную экспертизу. Кроме того, информация, размещенная на сайтах, нестабильна по времени, уязвима для ее искажения. Не решен также вопрос авторских прав, что препятствует размещению в сети высококачественной информации, с одной стороны, и способствует засилию рекламы, дезинформации, различного рода афер и т. д., с другой. В результате, при попытке провести информационный поиск в Интернете потребитель получает еще больше ссылок, чем при традиционном поиске, но при этом достоверность их не поддается оценке, да и доступ к первичному документу часто отсутствует.

На фоне неоспоримых успехов новых информационных технологий, достигнутых в различных областях общественной деятельности, возникло немало мифов информатизации [13]. Один из таких мифов заключается в том, что последовательная работа по формализации процессов информационного поиска и рост мощи компьютеров позволят разрешить основное противоречие информационного общества, о котором говорилось выше.

Рассмотрение истории развития технологий поиска информации создает ощущения какого-то непреодолимого тупика. Как бы ни совершенствовались технологии поиска, потребитель все равно оказывается один на один с таким количеством документов, которое он не в состоянии переработать и усвоить. Несмотря на то, что авторы приводят описание различных технологий поиска информации [14], по сути в их основе лежит одна и та же стратегия: найти все релевантные документы и изучить их. Эта стратегия нашла свое отражение в государственном стандарте в виде определения коэффициента полноты поиска. Государственный стандарт определяет этот коэффициент как величину, численно равную отношению количества найденных релевантных документов (документов соответствующих запросу) к общему количеству релевантных документов, имеющихся в информационном массиве [15]. Такой простой и ясный показатель хорош для оценки эффективности запросов в замкнутых и закрытых информационных системах, но мало что дает для реальной, практической работы исследователя. Дело в том, что если под информационным массивом понимать весь массив научных документов, существующих в мире на определенную тему, то определить их количество просто нереально. Но каждый исследователь имеет дело с открытой информационной системой, изменяющейся во времени. Поэтому количество релевантных документов не поддается точной оценке.

Изучать все релевантные документы не только нереально, но и совершенно не нужно. Представьте себе, что студенту-медику сказали бы, что для того чтобы стать врачом, необходимо изучить абсолютно каждую клеточку человеческого тела. За всю свою жизнь он успел бы изучить половину ногтя мизинца, так и не став врачом. Абсурд? Конечно. Зачем изучать каждую из миллиона одинаковых клеток ногтя, достаточно изучить одну абстрактную клетку и распространить эти знания на другие. И вообще, медик изучает организм на разных уровнях, не только на клеточном, но и на уровне отдельных органов, отдельных систем и т. д. Это позволяет справиться с гигантским массивом информации об организме. Точно так же и потребителю информации вовсе не нужно изучать все релевантные документы (клетки тела знания), т. к. степень совпадения их содержания по показателю пертинентности часто очень велика.

Обратим внимание на потребителя информации. Именно потребитель информации ставит цель поиска, и только он может определить, достигнута ли она. Между тем, изучение многочисленной литературы по ИПС приводит к удивительному выводу: главное лицо информационного поиска включается в ИПС в виде маленького черного ящика с одной единственной функцией – генерация запроса. Он генерирует запросы, а сложная ИПС, включающая в себя системы и подсистемы, специальные поисковые языки, компьютеры и информационных посредников, коммуникационные каналы и т. д., проводит поиск и выдает массив релевантных документов. Потребитель перерабатывает этот массив, выдает новый запрос, и процесс повторяется. С точки зрения анализа эффективности ИПС по поиску

релевантных документов, здесь все правильно. Но с точки зрения потребителя – это очень далеко до его цели. Так в чем же его цель?

Посмотрим на процесс тематического поиска с конца. Что является целью потребителя информации в общем случае? Под потребителем информации будем понимать лишь того реального человека со всеми его физиологическими и социальными свойствами, целью которого является изучение, познание некоторой области знания. Схема познания такова: имеется объективная реальность, эта реальность отражается в научном знании человечества, материальными носителями которого являются документы. Потребитель, изучая документы, строит свою индивидуальную модель области знаний, отражающую общечеловеческую модель. Потребитель обладает ограниченной скоростью (и весьма ограниченной) по переработке документов и ограниченным временем для построения модели. Это означает, что для решения задачи построения реальной (с точки зрения потребителя информации) модели области знания необходимо изучить весьма ограниченное количество релевантных документов. Именно так выглядит ситуация в реальной жизни. Если это так, то первая проблема, которая стоит перед потребителем, заключается вовсе не в поиске всех релевантных документов, а в поиске документов с наиболее качественным содержанием. Именно качество содержания документа играет важнейшую роль для потребителя, а релевантность – это одно из необходимых свойств документа. Необходимых, но недостаточных. В теории ИПС существует понятие пертинентности, т. е. соответствия содержания документа потребности. Потребитель из релевантных документов выбирает пертинентные. Но и пертинентные документы обладают различным качеством содержания.

Что мы будем понимать под качеством содержания документа с точки зрения информационного моделирования? Качество содержания документа – это такая характеристика документа, которая позволяет потребителю информации максимально оптимизировать процесс моделирования области знания. Понятие качества содержания документа включает в себя множество свойств: это и пертинентность, и информационная емкость, и достоверность, и понятность, и множество других свойств. Этот вопрос слабо исследован и требует дальнейших разработок. Понятие качества содержания документа кажется очень абстрактным и трудно оцениваемым. Тем не менее, можно предложить некоторые подходы к его оценке.

Так как для окончательного построения модели области знания необходим набор документов, то можно ввести понятие качества содержания набора документов. Таким образом, если для построения модели в одном случае потребовалось 100 документов, а в другом 200 документов, то качество первого набора в два раза выше, чем второго, и соответственно среднее качество содержания документов из первого набора в два раза выше, чем второго. Конечно, этот пример демонстрирует лишь принцип оценки качества содержания, однако он задает некоторый критерий, позволяющий разрабатывать более четкие процедуры и технологии.

Критерии качества содержания документа можно разделить на две группы: объективные и субъективные. Объективные критерии не зависят от свойств самого потребителя информации. Это вид издания, год, наличие рекомендации или грифа авторитетной организации (министерство, учебно-методический совет) и т. д. Хотя в последнее время появились исследования на эту тему [16], следует признать, что этот вопрос разработан недостаточно. Что же касается субъективных критериев качества документа, то они практически не исследованы.

Субъективные критерии качества содержания имеют большое практическое значение в процессе моделирования для очень большой группы потребителей. Что это за группа?

Даже если потребителю отводится не роль мелкого винтика в ИПС, а статус специалиста, то подразумевается, что это очень квалифицированный специалист как в своей предметной области, так и в области информационного поиска. Таких специалистов немного, они есть, но многократно больше потребителей информации неквалифицированных. Неквалифицированных – не означает плохих. Вспомним систему образования. Любой учащийся, студент, аспирант представляет именно эту категорию специалистов. Целью обучения и является получение высокой квалификации. Оставляя в стороне обучение конкретной специальности, следует отметить, что в области информационного поиска оно практически отсутствует. Нельзя же считать таким обучением четырехчасовое (за все пять лет обучения) знакомство с основами библиотечно-библиографических знаний студентов – будущих журналистов на первом курсе. Опрос этих студентов на пятом курсе показывает, что они все уже забыли [17]. На занятиях по информатике, в основном, идет освоение компьютерных офисных технологий, и здесь достигнуты хорошие результаты, разработаны отличные методики, учебные пособия, подготовлены преподаватели. Но все понимают, что этого недостаточно для формирования информационной культуры специалиста. Появились новые интересные курсы, такие как «Социальные коммуникации» [18], в котором, в частности, говорится о том, что для потребителя информации важна не просто информация, а ее смысл. Другой автор, анализируя проблемы ИПС [8], приходит к выводу о необходимости разработки модели потребителя информации. Все понимают, что необходимо обучать методам информационного поиска, но как? Это сложный вопрос, вызывающий множество дискуссий и требующий дальнейшей научной разработки.

Сформулируем основные принципы предлагаемой стратегии информационного моделирования области знания.

– Целью потребителя информации является изучение некоторой области знания, т. е. построение ее информационной модели. Тематический, документальный поиск является одним из средств такого моделирования.

– В процессе моделирования должны главенствовать интеллектуальные технологии оценки качества содержания и ранжирования источников информации и документов. Формализованные методы поиска должны быть подчинены интеллектуальным.

– Потребитель информации – лицо реальное, физическое, требующее учета его субъективных свойств, в т. ч. и необходимости обучения. Таким образом, использование современных педагогических средств –

неотъемлемая часть процесса информационного моделирования.

Надо сказать, что многие специалисты интуитивно пользуются этими принципами на практике. В неявном виде они присутствуют и в некоторых учебных пособиях [19]. Наверное, пора более четко сформулировать эти негласные правила, которыми пользуются опытные специалисты, прежде всего для того, чтобы повысить квалификацию обучающихся и еще неопытных потребителей информации.

ЛИТЕРАТУРА

1. Библиотековедение. Общий курс / под ред. К.И. Абрамова. М.: Книжная палата, 1988.
2. Основы библиотечно-библиографических знаний / Л.Р. Брауде, В.Н. Роженс, О.В. Чеснокова. М.: Высш. шк., 1987.
3. Библиографическая работа в библиотеке: организация и методика / под ред. О.П. Коршунова. М.: Книжная палата, 1990.
4. Информатика. Учебник. 3-е изд., перераб. / под ред. Н.В. Макаровой. М.: Финансы и статистика, 2001. 768 с.
5. Научно-техническая информация. Источники. Поиск. Использование / под ред. А.А. Фомина. М., 1977.
6. *Попов И.И.* Автоматизированные информационные системы. М.: Изд-во Рос. Экон. Акад., 1999. 103 с.
7. *Максимович Г.Ю., Романенко А.Г., Самойлюк О.Ф.* Информационные системы. М.: Изд-во Рос. Экон. Акад., 1999. 198 с.
8. *Захаров В.П.* Информационные системы (документальный поиск). СПб., 2002. 188 с.
9. *Хеннингер М.* Эффективные стратегии поиска в Internet. М.: Хеннингер, 1998.
10. *Байков В.Д.* Интернет: поиск информации и продвижение сайтов. СПб.: БХВ-Петербург, 2000. 288 с.
11. *Аверченков В.И., Рощин СМ., Трифанов Ю.Т.* Информационный поиск в Интернете / под общ. ред. В.И. Аверченкова. Брянск: ВГТУ, 2002. 304 с.
12. *Ростовцева Т.В.* Поиск информации в Интернет. СПб., 2002. 45 с.
13. *Зубец В.В., Ильин А.А.* Мифы информатизации. К постановке проблемы // Психолого-педагогический журнал Гаудеамус. 2003. № 2 (4). С. 181-187.
14. *Ростовцева Т.* Сравнительный анализ традиционных и автоматизированных технологий информационного поиска // Информационные ресурсы России. 2003. № 6 (76). С. 9-12.
15. ГОСТ 7.73-96. Поиск и распространение информации. Термины определения. М., 1996.
16. *Гречихин А.А., Древе Ю.Г.* Вузовская учебная книга: Типология стандартизация, компьютеризация. М.: Логос; Моск. гос. ун-т печати, 2000. 255 с.
17. *Зубец В.В.* Особенности информационной культуры студентов // Актуальные проблемы информатики и информационных технологий: материалы рос. науч.-практ. конф. (Сентябрь 2003). Тамбов: Изд-во ТГУ, 2003. С. 56-57.
18. *Соколов А.В.* Социальные коммуникации. М.: ИПО Профиздат, 2001. 224 с.
19. *Эко У.* Как написать дипломную работу. Гуманитарные науки: учеб.-метод. пособие / пер. с ит. Е. Костюкович. М.: Кн. дом «Университет», 2001. 240 с.

Поступила в редакцию 7 июля 2009 г.

Zubets V.V. From thematic search towards information modeling. In the work the short review of technologies of information search being used now is given. Main principles of the offered strategy of informational modeling of the knowledge area are formulated.

Key words: information consumer; information search; strategies of the knowledge area informational modeling.