СИСТЕМНЫЙ АНАЛИЗ И УПРАВЛЕНИЕ

УДК 025.4

ИНФОРМАЦИОННЫЙ ПОИСК В ИНФОРМАТИКЕ И БИБЛИОТЕКОВЕДЕНИИ

Л.В. ГРЕКОВА

Белгородский государственный институт искусств и культуры

e-mail: grekova@belkult.ru

Статья посвящена вопросам становления теории информационного поиска в отечественной информатике и библиотековедении и содержит ретроспективный обзор научных исследований, освещающих проблематику поиска документальных источников. Рассматриваются терминология теории информационного поиска, основы работы информационно-поисковых систем.

Ключевые слова: информационный поиск, информатика, библиотековедение, информационно-поисковые системы, информационно-поисковые языки, лингвистическое обеспечение информационно-поисковых систем.

Вторая половина XX века характеризуется стремительным ростом объемов различных видов информации. Статистика мирового книжного рынка за последние пятьдесят лет показывает, что число ежегодно выпускаемых книг увеличивается на 15–20 тыс. названий. В 1960 г. по данным ЮНЕСКО опубликовано 332 тыс. названий книг, в 1970 г. – 521 тыс., в 1980 г. – 715 тыс., в 1990 г. – 842 тыс., в 2000 г.– 1,25 млн. Это далеко не все выходящие книги, а только те, которые поступают в продажу. Статистические сведения о числе выходящих журналов можно получить из «Международной библиографии периодических изданий»: 13-е изд. библиографии (1969–70) содержит 40 тыс. названий, 15-е изд. (1971–72) – 55 тыс., 17-е изд. (1973–74) – 60 тыс., 19-е изд. (1980) – 62 тыс., 21-е изд. (1982) – 63 тыс., 24-е изд. (1988) – 100 тыс., 30-е изд. (1995) – 120 тыс., 39-е изд. (2001) – 164 тыс. [3].

Показательны цифры, которые приводил А. И. Черный еще в 1975 году: ежегодно публикуется около 3 млн. журнальных статей по естественным наукам и технике, 70-75 тыс. книг, примерно 450 тыс. описаний к патентам и авторским свидетельствам, а также сотни тысяч других научных документов. Если просматривать статьи со скоростью 30 публикаций в час, работая по 40 часов в неделю, то для просмотра, например, статей по химии, которые были опубликованы лишь в 1972 году, потребовалось бы около 5 лет [9, 9].

В условиях «информационного бума» в 50-60 гг. XX века проблема поиска информации выходит за пределы учреждений, которые традиционно накапливали, систематизировали и хранили документы (архивы, библиотеки, издательства и типографии, книготорговые организации), и начинает активно разрабатываться в рамках новой научной дисциплины – *информатики*.

Как точно заметил Р. С. Гиляревский, «информатика, заявившая о себе в середине XX столетия, принесла с собой не только новую и получившую распространение терминологию... Новым, действительно новым, оказался более широкий подход к явлениям и принципам. Понятие информационно-поисковая система объединило многие предметы, которые прежде рассматривались изолированно, например, библиотечные фонды и ката-

логи, различного вида... указатели, справочники энциклопедии, автоматизированные поисковые системы и т.п. Это дало возможность выявить общие принципы их построения, найти общие критерии их эффективности» [4, 11].

В России первое научное издание в области информатики опубликовано в 1965 году. Это монография А. И. Михайлова, А. И. Черного и Р. С. Гиляревского «Основы научной информации», созданная во Всесоюзном институте научной и технической информации (ВИНИТИ). В издании формулируется предмет и метод теории научной информации, дается характеристика различных видов документов как источников научной информации, освещаются методы и формы аналитико-синтетической переработки документов, излагаются основные принципы информационного поиска.

В 60-80 гг. XX века ряд научных исследований в области информационного поиска оформился в следующие издания: Э. С. Бернштейн «Вопросы теории поисковых систем» (1962), Д. Г. Лахути «Вопросы теории поисковых систем» (1963), Ф. У Ланкастер «Информационно-поисковые системы: характеристики, испытания и оценка» (1972), Г. Г. Белоногов, В. И. Богатырев «Автоматизированные информационные системы» (1973), Ю. И. Шемакин «Тезаурус в автоматизированных системах управления и обработки информации» (1974), А. И. Черный «Введение в теорию информационного поиска» (1975), П. И. Никитин «Автоматизированные системы обработки и поиска документальной информации» (1977), Ч. Миндоу «Анализ информационных систем» (1977), Дж. Солтон «Динамические библиотечно-информационные системы» (1979), А. В. Соколов «Информационно-поисковые системы» (1981).

Анализ перечисленных выше изданий позволил определить терминологию теории информационного поиска. Термин «информационный поиск» впервые введён в научный оборот американским математиком Кельвином Муэрсом в 1951 году. Основоположники отечественной информатики определяют *информационный поиск* как последовательность логических операций, конечной целью которых является выявление по заданным признакам всех документов, содержащих требуемую информацию (с последующей выдачей самих документов или их копий), или выдача фактических данных, представляющих собой ответы на заданные вопросы [6, 248].

В ГОСТе 7.73-96 «Поиск и распространение информации. Термины и определения» информационный поиск – это действия, методы и процедуры, позволяющие осуществлять отбор определенной информации из массива данных.

Выделяют два основных вида информационного поиска: документальный и фактографический. *Документальный поиск* – это информационный поиск, цель которого – нахождение в информационном фонде документов, соответствующих полученному запросу. *Фактографический поиск* – информационный поиск, при котором отыскиваемая информация имеет характер конкретных фактических сведений. Например, на запрос «Суда на подводных крыльях, выпускаемые промышленностью СССР», результатом документального поиска будут книги, статьи, информационные материалы, содержащие сведения об этих судах; результатом фактографического поиска будет перечень судов с их техническими характеристиками [8, 12].

Задача информационного поиска сводится к тому, чтобы, не прочитывая текстов документов, по каким-то внешним описательным признакам выбрать из множества такие, которые удовлетворяют информационную потребность и соответствуют информационному запросу. Для этого каждый документ снабжается *поисковым образом документа* (ПОД) – характеристикой, в которой кратко и однозначно выражается основное смысловое содержание документа. В виде такой же краткой и однозначной записи – *поискового предписания* – должен быть сформулирован информационный запрос. Процедура информационного поиска состоит в сопоставлении ПОДов с поисковыми предписаниями и при их формальном совпадении считается, что документ соответствует информационному запросу [6, 250].

В поисковый образ документа включают результаты его аналитико-синтетической переработки. ГОСТ 7.0-99 «Информационно-библиотечная деятельность, библиография. Термины и определения» определяет *аналитико-синтетическую переработку* как процесс преобразования документов в процессе их анализа и извлечения необходимой информации, а также оценка, сопоставление, обобщение и представление информации в

виде, соответствующем запросу и раскрывает виды аналитико-синтетической переработки документов: библиографирование, аннотирование, реферирование, индексирование.

Информационный поиск в документальных массивах производится с использованием *информационно-поисковых систем* (ИПС).

В ГОСТе 7.73-96 «Поиск и распространение информации. Термины и определения» информационно-поисковая система – совокупность справочно-информационного фонда и технических средств информационного поиска в нем. В исследованиях по теории информационного поиска приводятся по сути аналогичные определения ИПС:

– «некий комплекс, охватывающий документы, запросы, формализованные описания этих документов и запросов, механизм, позволяющий сравнивать эти описания, и человека» [5];

– «некоторая совокупность или комплекс связанных друг с другом отдельных частей, предназначенный для выявления в каком-либо множестве элементов информации (документов, сведений и т.д.), которые отвечают на информационный запрос, предъявленный системе» [6];

– «совокупность информационно-поискового языка, правил обработки, поиска и выдачи информации, программы, а также технических средств, с помощью которых осуществляется процесс хранения, поиска и выдачи информационных материалов» [7].

Примерами традиционных информационно-поисковых систем являются библиотеки, архивы, музеи, других хранилища информации. С развитием информационных компьютерных технологий получили широкое распространение автоматизированные информационно-поисковые системы: электронные каталоги различного назначения; электронные справочники и словари; электронные библиотеки; информационно-правовые базы данных; поисковые системы сети Интернет.

В структуре реально действующих ИПС выделяют следующие основные элементы:

– информационно-поисковый массив (т.е. определенное множество документов, снабженных поисковыми образами, среди которых разыскиваются необходимые документы);

– логико-семантический аппарат (т.е. информационно-поисковые языки – один или два, правила индексирования и критерий выдачи);

– технические средства (т.е. устройства, которые необходимы для записи и хранения поисковых образов, для хранения самих документов, а также для осуществления процесса сопоставления поисковых образов документов с поисковыми предписаниями);

– люди, взаимодействующие с системой (т.е. те, кто пользуется данной ИПС и обслуживает ее – осуществляет индексирование документов и информационных запросов, выбирает стратегию поиска, а также выполняет другие интеллектуальные операции, без которых невозможен информационный поиск) [9, 18-19].

А.И. Михайлов, А.И. Черный, Р.С. Гиляревский предлагают рассматривать ИПС в абстрактном виде. В понятие абстрактной ИПС не включаются средства ее технической реализации [6, 250].

ГОСТ 7.73-96 «Поиск и распространение информации. Термины и определения» определяет *информационно-поисковый массив* как упорядоченную совокупность документов, фактов или сведений о них, предназначенную для информационного поиска.

В теории информационного поиска под документом понимается любой записанный на каком-либо материальном носителе осмысленный текст, который обладает определенной логической завершенностью и содержит сведения о его источнике и/или создателе. По этому определению документом является не только книга, статья, описание к авторскому свидетельству или патенту и т.д., но и отдельные фрагменты такого текста – глава, раздел, абзац и т.п. [9]. Различают первичные и вторичные документы.

Первичный документ – это документ, содержащий в зафиксированном на материальном носителе исходную информацию, полученную в процессе исследований, разработок, наблюдений, анализа или других видов познавательной человеческой деятельности, независимо от ее характера или тематики, оформленный в установленном порядке, имеющий в соответствии с действующим законодательством юридическую силу [2].

Содержание и форма первичных документов определяется замыслом автора и чаще всего не может быть унифицирована. Вместе с тем, автоматизация информационных про-

цессов, расширение сферы человеко-машинной коммуникации и вовлечение в процессы обмена информацией все большего числа участников предъявляет свои требования к языковым средствам, используемым для фиксации, передачи, хранения и поиска информации. Усиливается регламентация порядка изложения содержания отдельных типов документов, например, основные требования к оформлению организационно-распорядительных документов изложены в ГОСТе Р 6.30-97 «Унифицированная система организационно-распорядительной документации. Требования к оформлению документов», отчеты о научных работах оформляются в соответствии с ГОСТом 7.32-2001 «Отчёт о научно-исследовательской работе.

В документах наряду с элементами естественного языка (слова, словосочетания и т.п.) широко используются цифровые и алфавитно-цифровые коды и индексы, которые выступают эквивалентами наименований понятий на естественном языке. Так, например, в экономике, статистике, управлении для унификации информации используются классификаторы технико-экономической и социальной информации.

Первичные документы различаются в зависимости от материального носителя (формы), способов распространения и содержания. При всем многообразии документов, прежде всего, это *издания* – документы, предназначенные для распространения содержащейся в них информации, прошедшие редакционно-издательскую обработку, самостоятельно оформленные, имеющие выходные сведения. Номенклатура изданий зафиксирована в ГОСТе 7.60-2003 «Издания. Основные виды. Термины и определения».

Вторичный документ – формализованный документ, полученный в результате аналитико-синтетической переработки одного или нескольких первичных документов. Примерами вторичных документов являются справочные и энциклопедические издания, рефераты и реферативные издания, библиографические пособия, каталоги и картотеки [2].

Завершая рассмотрение информационно-поискового массива ИПС, подведем некоторый итог. Информационно-поисковый массив ИПС содержит первичные и вторичные документы и разделен на два массива – пассивный и активный. Пассивный массив (первичные документы) образуют сами документы. Активный массив (вторичные документы) содержит поисковые образы документов и адреса хранения этих документов в пассивном массиве. Именно в активном массиве ведется информационный поиск, т.е. сопоставление хранящихся в нем поисковых образов документов с поисковыми предписаниями, поступающими в ИПС.

Аналогичный подход к организации информационных массивов в виде первичных и вторичных документов используется в работе информационно-поисковых систем сети Интернет. Вторичные документы – поисковые образы web-страниц хранятся в базе данных индекса (Index database). Программы сканирования сети (роботы-индексировщики) просматривают web-страницы в сети Интернет, автоматически приписывают им ключевые слова и помещают ключевые слова в базу данных индекса. Обычно роботы используют для отбора ключевых слов следующие источники: гипертекстовые ссылки, заголовки, заглавия, аннотации, списки ключевых слов, полные тексты документов, а также метаданные, указанные с помощью метатегов title, description и keywords.

Информационно-поисковый язык (ИПЯ) – формализованный искусственный язык, предназначенный для индексирования документов, информационных запросов и описания фактов с целью последующего хранения и поиска [13].

В теории информационного поиска ИПЯ по праву отводится определяющая роль, «ибо от эффективности применяемого ИПЯ и логики поиска в решающей степени зависят его результаты» [9].

Становление и развитие ИПЯ относятся к началу 60-х гг. прошлого века. Вопросам структуры, назначения, типологии ИПЯ, практике их использования посвящена обширная Список литературы, опубликованная именно в этот период, вплоть до середины 1980-х годов. Одной из первых отечественных монографий, полностью посвященных проблемам типологии и конструирования искусственных языков, включая ИПЯ, является книга В.А.Московича «Информационные языки» (1971). В ней на основании анализа исследования нескольких сот информационных языков общего и специального назначения впервые в отечественной практике была приведена характеристика основных разновидностей языков, используемых при информационном поиске, дан анализ их особенностей и отли-

чительных черт. Практически одновременно с понятием ИПЯ в публикациях появляется термин «*лингвистическое обеспечение*» (ЛО).

В России системная разработка лингвистического обеспечения ИПС велась, начиная с 1960-х гг., по нескольким направлениям. В 1965 году было начато проектирование лингвистического обеспечения Государственной автоматизированной системы научно-технической информации (ГАСНТИ). В результате к концу 1980-х гг. в состав ЛО ГАСНТИ входило до 200 тезаурусов и рубрикаторов по всем отраслям народного хозяйства. Кризис 1990-х гг. в системе НТИ России совпал со сменой поколений компьютеров, что в совокупности привело к почти полной утрате достижений того времени. В настоящее время из общесистемных языковых средств ЛО ГАСНТИ поддерживается Государственный рубрикатор научно-технической информации и Универсальная десятичная классификация (УДК).

Параллельно с ГАСНТИ велось создание комплекса языковых средств автоматизированных систем организационно-экономического управления разного уровня, получившего название «Единая система классификации и кодирования технико-экономической информации» (ЕСКК ТЭИ). Научный уровень этих разработок был несколько ниже, чем в ГАСНТИ, зато масштабы работ гораздо шире. В результате была создана система общероссийских классификаторов, число которых к концу 1980-х гг. достигло 35, а их общий объем превысил 3 млн. позиций.

В 1960-1980 гг. в состав лингвистического обеспечения библиотечных систем входили созданные в конце XIX – начале XX века Десятичная классификация Дьюи, УДК, язык библиографического описания, язык предметных рубрик. В 1960-1968 годах опубликована Библиотечно-библиографическая классификация, разрабатываема с 30-х годов XX века. Появление в российских библиотеках электронных каталогов способствовало возобновлению научных исследований в области ИПЯ. В 1980 годы появляются публикации, посвященные использованию кодовых иерархических классификаций, языка предметных рубрик и дескрипторных ИПЯ в условиях автоматизированного информационного поиска.

В 1990-е гг. в России бурно развивались коммерческие и негосударственные информационные системы. В результате были сделаны первоклассные разработки в области ЛО. Среди них следует отметить поисковые машины с применением морфологического анализа (Яндекс, Рамблер и др.), системы навигации и поиска правовой информации (Консультант Плюс, Гарант, Кодекс и др.), системы оптического распознавания текстов (ABBYY FineReader, OCR CuneiForm), системы распознавания устной речи, системы машинного перевода и др.

В настоящее время наиболее продвинутыми являются средства ЛО коммерческих автоматизированных ИПС. Однако коммерческие компании не занимаются научными исследованиями, которые можно тиражировать в практику. В целом можно констатировать отсутствие в отечественной науке конца XX – начале XXI вв. системных разработок в области развития информационно-поисковых языков [1].

Кроме ИПЯ в логико-семантический аппарат ИПС входят правила индексирования и критерий выдачи.

Под *правилами или методикой индексирования* понимают совокупность приемов и правил образования поисковых образов документов или поисковых предписаний, т. е. приемов и правил перевода с естественного языка на искусственный – информационно-поисковый язык. Основной задачей методики индексирования является обеспечение единообразия подходов к созданию поисковых образов документов.

В целом общие требования к систематизации и предметизации документов установлены ГОСТом 7.59-2003 «Индексирование документов. Общие требования к систематизации и предметизации»; требования к координатному индексированию содержатся в ГОСТе 7.66–92 «Индексирование документов. Общие требования к координатному индексированию»; процесс аннотирования и реферирования регламентируются ГОСТом 7.9-95 «Реферат и аннотация. Общие требования».

Чтобы понять содержание термина «критерий выдачи», необходимо проводить четкое различие между такими понятиями, как «информационная потребность» и «информационный запрос». Информационный запрос – это словесное выражение определенной информационной потребности, которая далеко не всегда бывает правильно осознана и точно сформулирована человеком, испытывающим такую потребность. Всем людям в разной сте-

пени свойственна способность сразу адекватно выражать свои информационные потребности в виде информационного запроса. Поэтому реальная ИПС может обеспечить отыскание лишь таких документов, которые отвечают на информационный запрос в том виде, в каком он сформулирован. Документ, центральный предмет или тема которого формально соответствует информационному запросу, называется *релевантным*. Документ, соответствующий информационной потребности, называется *пертинетными*. Понятия релевантности и пертинентности не эквивалентны: они пересекаются, но не совпадают друг с другом. Как правило, факт пертинентности документа может быть установлен лишь после прочтения полного текста этого документа.

Совокупность признаков, на основании которых определяется релевантность документов по отношению к информационному запросу и принимается решение о выдаче или невыдаче данного документа в ответ на поставленный информационный запрос, называется *критерием выдачи* [9, 116].

Простейшим критерием выдачи является полное совпадение поискового образа документа с поисковым предписанием или полное вхождение последнего в поисковый образ. Однако практический опыт свидетельствует о том, что такой критерий выдачи не обеспечивает достаточной полноты информационного поиска, а иногда и его точности. Поэтому широко применяется критерий выдачи, основанный на частичном совпадении поискового образа документа с поисковым предписанием.

Подведем итого. Основы современной теории информационного поиска были заложены в результате научных исследований, проведенных в период с конца 1950-х до начала 1980-х гг. в рамках информатики. В этот период публикуются основополагающие труды в этой научной области. Начиная с конца 80-х годов прошлого века, проблема информационного поиска активно разрабатывалась библиотечными специалистами. В результате большинство понятий теории информационного поиска стандартизировано в ГОСТах Системы стандартов по информации, библиотечному и издательскому делу. В последние два десятилетия отдельные вопросы информационного поиска исследуются фрагментарно в рамках кандидатских диссертаций. С сожалением приходится отмечать отсутствие концептуальных научных работ, обобщающих опыт реализации автоматизированных информационно-поисковых систем, в том числе ИПС сети Интернет.

Список литературы

1. Антопольский, А. Б. Лингвистическое обеспечение электронных библиотек [Электронный ресурс] // Российский научно-электронный журнал «Электронные библиотеки». – 2002. – № 2. – Режим доступа: http://www.elbib.ru. – Загл. с экрана.

2. Воройский, Ф. С. Информатика. Новый систематизированный толковый словарь-справочник. – 3-е изд., перераб. и доп. – М., 2003.

3. Гиляревский, Р. С. Основы информатики. – М., 2003.

4. Гиляревский, Р. С. К проблеме совместимости ИПЯ различных типов // НТИ. Сер.2. – 1978. – № 1.

5. Ланкастер, Ф. У. Информационно-поисковые системы. – М., 1972.

6. Михайлов, А. И. Основы научной информации / А. И. Михайлов, А. И. Черный, Р. С. Гиляревский. – М., 1965.

7. Никитин, П. И. Автоматизированные системы обработки и поиска документальной информации. – М., 1977.

8. Соколов, А. В. Информационно-поисковые системы. – М., 1981.

9. Черный, А. И. Введение в теорию информационного поиска. – М., 1975.

INFORMATION SEARCH IN COMPUTER SCIENCE AND LIBRARY SCIENCE

L.V. GREKOVA

Belgorod State Institute of Arts and Culture

e-mail:grekova@belkult.ru

The article is devoted to the matters of becoming the theory of information search in domestic computer science and library science and includes retrospective overview of scientific researches which report the problems of documentary sources' search. In the article the terminology of information search theory, work fundamentals of information search systems, are considered.

Keywords: information search, computer science, library science, information search systems, information search languages, linguistic support of information search systems.