УДК 681.3.016:026/027

Нестеров А. В.

БИБЛИОТЕЧНЫЕ БАЗЫ ДАННЫХ КРУПНОЙ БИБЛИОТЕКИ

Рассмотрены некоторые вопросы создания, ведения и использования библиотечных баз данных ГПНТБ СО АН СССР. Показано, что на современном этапе крупная библиотека должна создавать свои собственные библиотечные базы данных. Представлены некоторые аспекты проблемы глубокой обработки данных документально-информационного потока. Приведено описание библиотечных баз данных ГПНТБ СО АН СССР, а также смыслового ключа, используемого в них.

В рамках компьютерной системы коммуникации НТИ СО АН СССР ГПНТБ СО АН СССР создает библиотечные базы данных. Основное направление в их развитии — создание библиотечных баз данных документов, аккумулирующих информацию из различных источников. Сначала к ним были отнесены авторефераты как наиболее многочисленный поток (до 500 экз. в неделю), затем — материалы конференций (до 50 экз. в неделю), библиографические указатели и, наконец, обзоры. Кроме того, в библиотеке создана база данных отечественной и иностранной периодики, поступающей в ГПНТБ и научно-исследовательские учреждения СО АН СССР.

Такое направление отражает концепции разработчиков в отделе автоматизированных систем обработки и анализа информации в ГПНТБ СО АН СССР по проектированию автоматизированных библиотечно-информационных систем. Под библиотечными базами данных понимается не только конечный продукт в виде катушки с магнитной лентой, на которой имеются библиографические записи, но и технология их создания (генерация, ретрансляция), ведения (накопление, хранение в архиве, тиражирование) и потребления (широкий доступ читателей к библиотечным базам данных, многократное использование в служебных целях и наукоемкое применение историками, разработчиками баз данных и др.).

Особенностью базы данных авторефератов является наличие в ней фактографической информации справочного характера — шифра научной специальности и шифра хранения. Авторефераты представляют собой уникальный вид документа, концентрирующий основное содержание диссертации и представляющий обзор достижений в конкретной области, а также список работ автора. Кроме того, автореферат фиксирует общественную значимость проведенного исследования, а множество авторефератов по определенной тематике указывает на общественную необходимость работ в этом направлении.

На основе анализа базы авторефератов и массива авторефератов можно проводить различные исследования. Так, приведено описание некоторых из них в США [1] и в СССР [2—4].

За рубежом в этой области работают О. Персон, М. Гарет, П. Булет-Налин, С. Басседж, которые доказали практическую ценность базы данных Dissertation Abstracts, содержащей информацию о всех защищенных в США и Канаде докторских диссертациях за период с 1861 г. Интерактивный доступ к этой БД осуществляется через систему BRS Information Technologies, а список дескрипторов был получен с помощью справочника BRS Diss. Database Guide.

Национальная Академия наук США выпускает справочник «A century of Doctorates» с информацией о диссертациях. В области библиотековедения рекордное количество диссертаций приходится на 1970—1974 гг., а в области информатики наблюдается устойчивый рост начиная с 1970 г.

В США и других странах присуждение научных степеней рассматривается как один из основных показателей продуктивности ученого. Дж. Б. Вуд [1] считает, что защита диссертации означает общественное официальное признание, поэтому автореферат диссертации может служить индикатором роста научных знаний.

Аналогичные работы по созданию баз данных рефератов диссертаций ведет Британская библиотека, которая выпускает библиографический указатель дессертаций. В среднем в БД по диссертациям в Британскую библиотеку поступает около 6 тыс. документов в год, а общий фонд составляет 76 тыс. экз. Кроме того, в Британскую библиотеку поступает около 6 тыс. документов по диссертациям из США, общий объем диссертаций около 416 тыс.

ГПНТБ СО АН СССР начала накапливать базу данных «Автореферат» с 1986 г. В ней фиксируются все выпущенные в СССР авторефераты диссертаций:

ГПНТБ СО АН СССР получает обязательный экземпляр. Средний прирост базы данных за год составляет 20—25 тыс. документов.

Среди работ, которые проводят сотрудники ГПНТБ СО АН СССР с базой данных «Автореферат», можно выделить две. Первая — исследовательская, связанная с проектом «Экспертиза» по разработке, созданию и внедрению в ГПНТБ СО АН СССР автоматизированного рабочего места документально-информационного эксперта на ЭВМ ЕС-1841. В соответствии с этим проектом в ГПНТБ СО АН СССР был проанализирован массив авторефератов по информатике и близким к ней темам с середины 1960-х гг. до 1988 г. включительно на основе предметного и тематического поиска по базе данных «Автореферат» и карточного каталога ГПНТБ СО АН СССР по разделу ББК «Информатика». Результаты этой работы изложены в отчете о научно-исследовательской работе [4] и статье в сборнике [3].

Вторая работа с базой данных «Автореферат» заключается в создании метаинформационных клипов (специально сделанная выборка информации об информации, хранящейся в библиотеке). Она отличается от библиографических указателей тем, что кроме стандартного библиографического описания содержит справочную информацию о шифрах хранения, шифрах научной специальности, уровне и виде присваиваемой степени, организации, где прошла защита диссертации. ГПНТБ СО АН СССР ежемесячно поставляет на коммерческой основе метаинформационные клипы по научным специальностям, темам, предметам, терминам, городам, организациям, имеющим ученые советы (на магнитных лентах, дискетах пять и восемь дюймов для ЕС ЭВМ, СМ ЭВМ, ДВК, «Электроника-85», ЕС-1840), а также наборы каталожных карточек в виде тематических карточек (фрагментов каталога ГПНТБ СО АН СССР). И, наконец, база данных «Автореферат» доступна читателям ГПНТБ СО АН СССР на СМ ЭВМ в Новосибирске и на ЕС ЭВМ в Академгородке ежедневно в течение одного часа.

Рассмотрим технологический аспект библиотечных баз данных.

Крупная библиотека является мощным перерабатывающим документы предприятием с устойчивым технологическим циклом. Документальный поток, который обрабатывается в крупной библиотеке, является не стационарным, так как документы поступают в библиотеку из нескольких источников комплектования, неравномерно, в разном количестве. Особенностью неравномерности документального потока является изменение числа документов по видам и по тематике. Это приводит к неравномерной загрузке библиотекарей, занятых в отделах обработки.

В связи с этим модель крупной библиотеки состоит из двух элементов, один из которых выполняет функции буфера, предназначенного для интегрирования колебаний входного потока документов, а второй — для накопления обработанных документов.

Для анализа работы этих элементов важное значение имеет классификация поступающих документов и статистические данные по ее результатам. Такая информация влияет не только на технологию обработки документов в крупной библиотеке, но и на политику комплектования литературы. И, наконец, коммуникационные процессы между автором и читателем замыкаются в библиотеке, поэтому их эффективность зависит от глубины обработки документов.

Одним из основных подходов при создании компьютерной системы коммуникации НТИ СО АН СССР является глубокая компьютерная обработка данных из документально-информационного потока, поступающего в ГПНТБ СО АН СССР.

Под глубокой обработкой данных подразумевается извлечение из этих данных информации и знаний: информация есть используемые данные для принятия решения и ретрансляции, а знания — используемые данные генерации новых данных, информации и знаний. Извлечение информации и знаний из данных осуществляется целеустремленным индивидом или системой в том случае, если на основе этих данных происходит выбор и выполнение каких-либо процессов.

Поэтому электронная коммуникация требует средств манипуляции с активными данными, т. е. такими, которые могут быть обработаны компьютерными средствами. Итак, современная передача информации подразумевает передачу не только данных, но и технологий обработки этих данных в отличие от простой документальной коммуникации, где технология извлечения информации фактически определяется только человеком, потребляющим эту информацию.

Однако такая информационная технология хотя и позволяет передавать информацию, но не дает возможности передавать знания, так как она в этом случае ориентирована только на извлечение информации из данных.

Наивысшая ступень электронной коммуникации подразумевает глубокую обработку данных — наличие средств извлечения знаний из данных. Таким образом, современная передача информации должна ориентироваться не только на передачу данных и информационных технологий, но и на передачу техноло-

гий знаний, т. е. таких методов, средств и навыков, которые позволяют читателям (приемникам) не только воспринимать извлекаемую информацию, но и генерировать новые данные, информацию и знания, отсутствующие в передаваемых базах данных.

Современная крупная библиотека должна предоставлять три вида коммуникации: личные, документальные, электронные (компьютерные). Особенностью электронной коммуникации является необходимость глубокой обработки документов и доступ к библиотечным базам данных.

ГПНТБ СО АН СССР могла бы не создавать свои собственные библиотечные базы данных, если бы их централизованно делали и оперативно распространяли соответствующие организации. Однако этого пока нет. Отсутствует также глубокая обработка документов, так как поставляемые информационными центрами с большим опозданием библиографические базы данных не обладают этим свойством.

При глубокой обработке данных документ может переносить в пространстве и во времени какие-либо отображения действительности и познания передающего элемента, предназначенные для воздействия на принимающие элементы коммуникационной среды. В качестве документов могут выступать отображения, которые отражают сами себя, либо нечто, либо неизвестно что. Они могут переноситься в пространстве и во времени в виде тел, процессов и полей, образованных естественными, и искусственными частицами.

Глубокая обработка данных документов, поступающих в ГПНТБ СО АН СССР, подразумевает анализ (разложение) документа и фиксацию элементов этого анализа, а также свойств, атрибутов и признаков этих элементов и самого документа. Для анализа документа предлагается использовать фасетный метод Ранганатана, модифицированный в категорийно-фасетный подход и позволяющий разложить предмет документа на тензоры (основные, метатензоры, мегатензоры), вычислить некоторые координаты документа в псевдосемикоординатном пространстве, в частности *рейтинг документа* [5]. *Кроме того, необходимо* фиксировать его связи с другими документами посредством фиксации некоторых *из них из списка литературы. Особенность* этой фиксации — занесение в память компьютера не всех записей из списка литературы, а только в пределах до семи, отобранных по специальному правилу с помощью датчика случайных чисел. В основе лежит гипотеза о нелинейном характере зависимости точности индексирования документа списком литературы от количества фиксируемых документов из этого списка.

При фиксации данных документа в режиме реального времени с возможностью прямого доступа к библиотечным базам данных объем вводимых данных резко сокращается за счет постепенного накопления вводимых ссылок. По оценкам специалистов ISI базы данных могут содержать до 70—80% документов из списков литературы [6]. Кроме фиксации связей между документами необходимо фиксировать связи между документами и людьми, запрашивающими эти документы, в виде фиксации книговыдачи, запросов МБА, заказов на копирование, перевод и т. п. И, наконец, нужно запоминать некоторые атрибуты документа, связанные с видом носителя, т. е с формальными характеристиками документа, не имеющими отношения ни к предмету документа, ни к процессу коммуникации.

Особенностью библиотечных баз данных является наличие в их записях так называемого смыслового ключа, который автоматически формируется при полном вводе библиографической записи из фрагментов библиографического описания документа.

В качестве элементов поискового смыслового ключа выберем четыре поля: автор (первый), заглавие (основное), название сборника, журнала и т. п. и время издания. Эти поля выбраны не случайно. Документ в общем случае является квантом каких-либо данных, отчужденных от источника этих данных, поэтому к основным элементам поискового ключа относятся атрибуты источника и этого кванта. Источник в смысловом ключе отражается полем — автор. Заглавие документа отражает предмет, и название сборника отражается полем: сборник указывает на коллектив, организацию, издающую документ. Наконец, дата издания позволяет идентифицировать документы с одинаковым автором, заглавиями, но вышедшими в разное время. Особенность атрибута: дата для него обязательна, остальное — необязательно, поэтому возможно семь сочетаний по трем основным элементам библиографического описания документа (см. табл.).

Таблица

№	Автор	Заглавие	Сборник	Ключ
1	+	—	—	AAAAA**9
2	—	+	—	*AAAAA*9
3	—	—	+	**AAAAA9
4	+	+	—	AAAAAA*9
5	+	—	+	AAAAA*A9
6	—	+	+	*AAAAAA9
7	+	+	+	AAAAAAA9

В таблице приведен смысловой ключ, составленный из фрагментов библиографического описания. Он назван смысловым потому, что его может составить читатель по смыслу, заложенному в самом библиографическом описании. Смысловой ключ состоит из четырех полей. Первое поле формируется из символов, взятых из фамилии, имени, отчества автора (первого) и содержит 5 знаков. Три первых составлены из первых букв фамилии, а два остальных знака — из первых букв имени и отчества. Следующий (шестой) знак выделяется на первую букву заглавия, седьмой — на первую букву названия сборника, журнала, конференции и т. п. И, наконец, восьмой знак служит для указания даты путем фиксации последней цифры года издания. Таким образом, смысловой ключ формируется из восьми знаков, пять из которых отведены для идентификации источника, один — для идентификации предмета и один — для идентификации косвенного источника.

Общая длина ключа в восемь символов взята из следующих соображений. Она должна быть кратна восьми для удобства обработки на компьютере, не должна превышать 7—8 символов по психофизическим свойствам человека, набирающего на клавиатуре какое-либо число, а также количество знаков и их сочетание должно удовлетворять требованию уникальности ключа.

Смысловой ключ для поискового образа документа формируется автоматически компьютером из фрагментов библиографического описания, которое набирается вручную при первичном вводе, либо из фрагментов описания на магнитных лентах какого-нибудь информационного центра, например ВИНИТИ.

Смысловой ключ для поискового образа запроса формируется вручную читателями с помощью клавиатуры дисплея. Рассмотрим правила, с помощью которых формируется смысловой ключ.

1. Если отсутствует какой-либо основной элемент библиографического описания, то в первый знак соответствующего поля ставится *. (Например, отсутствует фамилия, имя, отчество, в этом случае ключ изображается в виде последовательности, аналогичной № 2 в табл.)

2. Если первое слово из основного элемента содержит менее трех букв, то оставшиеся знаки ключа дополняются символами из второго и т. д. слов основных элементов описания. (Например, имеется только заглавие «О птицеводстве». В этом случае ключ имеет вид: *ОПТИЦ*9. Общее количество букв должно быть равно пяти.)

3. Если общее количество букв в ключе менее пяти, то оставшиеся позиции ключа заполняются знаком вопроса (?). (Например, имеется заглавие «Кот». В этом случае смысловой ключ будет выглядеть следующим образом: *КОТ??*9.)

4. Если в документе нет автора, а есть редактор, то на место фамилии, имени, отчества автора ставят фамилию, имя и отчество редактора.

5. Если в ключе имеется * в поле фамилии, имени, отчества и одном из других полей, то он составляется из букв нескольких слов по следующему правилу: если поле содержит пять и более слов, то выбираются первые буквы из пяти слов; если слов меньше пяти, то из первых слов выбираются первые буквы, а оставшиеся позиции ключа до пяти заполняются вторыми и т. д. буквами из последнего слова, например, журнал «Советское радио» будет представлено «СРАДИ». Это сделано для устранения неинформативности первых слов во многих названиях журналов.

6. Если в ключе имеется * в поле фамилии, имени, отчества и в поле заглавия, то вместо этих * надо ставить номер сборника (журнала).

7. Если в поле сборника имеется *, то вместо * надо ставить код вида документа из следующей шкалы: 0 — автореферат, 1 — алгоритмы и программы, 2 — справочник, 3 — методическое пособие (учебник), 4 — НТД, 5 — патент (авторское свидетельство), 6 — обзор, 7 — материалы конференции, 8 — книга, 9 — каталог.

8. Если какое-либо значение в разделе ключа отсутствует, необходимо ставить знак вопроса (?).

Смысловой ключ можно использовать не только для поиска шифра хранения документа. Рассмотрим более подробно аспекты использования смыслового ключа, их три.

Первый — пользовательский (прагматический). Смысловой ключ используется для поиска шифра хранения документа, он должен быть удобным, понятным и набираться с помощью простого алгоритма.

Второй — технологический (процедурный). В данном случае смысловой ключ используется библиотекарем-технологом, который набирает этот ключ на клавиатуре компьютера с целью проверки на дублетность обрабатываемого документа, т. е. проверки на нахождение данного ключа в базе данных, что соответствует наличию документа в библиотеке и электронного библиографического описания в памяти компьютера. Если набранный ключ имеется в базе данных, то соответствующее ключу описание можно вызвать на экран, и, если оно соответствует докумен-

ту, не вводить описание повторно в память компьютера.

Третий — научный (познавательский). В этом случае смысловой ключ используется для библиометрических исследований. Например, пристатейный (прикнижный) список литературы можно не вводить в память компьютера, а вводить только смысловые ключи и указывать на связь цитируемых смысловых ключей с смысловым ключом документа, где эта пристатейная (прикнижная) литература используется. Смысловые ключи в этом случае могут применяться для создания массива связности данных ключей, по которому можно анализировать связи между документами.

СПИСОК ЛИТЕРАТУРЫ

1. **Wood J. B.** The growth of Scholarship: an online bibliometric comparison of dissertations in the sciences and humanities // Scientometrics.— 1988.— V. 13.— № 1—2.— P. 53—62.

2. **Гиляревский Р. С., Калошин В. В.** Тенденции развития информатики (по отечественным диссертациям от 1965 г. до 1980 г.) // НТИ. Сер. 1.— 1988.— № 4.— С. 2—8.

3. **Нестеров А. В., Иловайский И. В.** Экспертиза и консультации в современной библиотеке // Эффективность использования документальных банков данных в научных исследованиях / ГПНТБ СО АН СССР.— Новосибирск, 1989.— С. 37—44.

4. **Изучение** работы эксперта в библиотеке: Отчет о НИР / ГПНТБ СО АН СССР.— Новосибирск, 1988.— 48 с.— Машинопись.

5. **Компьютерное** рабочее место документально-информационного работника на ЕС-1841: Отчет о НИР / ГПНТБ СО АН СССР.— Новосибирск, 1989.— 37 с.— Машинопись.

6. **Маркусова В. А., Черный А. И.** Информационная продукция и технология ее подготовки в институте научной информации США // НТИ. Сер. 1.— 1985.— № 12.— С. 6—15.