СПРАВОЧНЫЙ АППАРАТ БИБЛИОТЕКИ

ШАРОНОВА Н.В. КАНИЩЕВА О.В.

ХНПИ, г. Харьков

ПРОБЛЕМА ИНДЕКСИРОВАНИЯ ПОЛНОТЕКСТОВЫХ ДОКУМЕНТОВ ПО КЛЮЧЕВЫМ СЛОВАМ

Обучение навыкам поиска и сам процесс поиска литературы (текстовой информации) являлись и тем более являются сейчас необходимой частью образовательного процесса. В доэлектронный период поиск информации был организован главным образом посредством каталогов библиотек, архивов и издательств, а также оглавлений самих печатных изданий с использованием системы ссылок (назовем эту технологию для краткости поиском по каталогу).

Такая технология поиска больших трудностей в использовании и обучении не вызывала, поскольку процедура поиска по каталогу однозначна: зная автора или название документа или издательство и год издания, можно в принципе найти искомый документ. А затем, ознакомившись с оглавлением и аннотацией, получить более - менее точное представление о том, насколько данная информация необходима.

Следует отметить, что и тогда технология поиска по каталогу не удовлетворяла полностью потребности пользователя. В известной мере это компенсировалось наличием в ряде научных изданий детально разработанных предметных указателей, содержащих наряду с рубрикой сведения о соответствующих страницах. Однако процесс составления таких ссылок (полнотекстовая индексация) был очень долог и трудоемок и поэтому технология поиска по тексту не могла быть широко распространена.

В настоящее время объем информации многократно увеличился, а структура информации усложнилась. Количество информации возросло не только за счет появления большого числа независимых источников информации, но и в силу того, что электронные технологии сделали информацию более доступной для каждого. Под усложнением структуры информации понимается главным образом то, что эта информация в значительной своей части не систематизирована. Кроме того, при работе с большими объемами информации резко уменьшается время на анализ уже найденных документов.

Накопленные сегодня колоссальные объемы информации, в совокупности с непрерывно увеличивающимися темпами ее роста, определяют актуальность и значимость исследований в области информационного поиска. В этих условиях система традиционного ручного поиска по каталогу становится уже совершенно недостаточной. На смену ей должна прийти, или, по крайней мере дополнить её, технология автоматизированного, а может быть, и автоматического поиска по тексту.

Подобная технология неразрывно связана с проблемой формирования ключевых слов для документа. Разработчики современных автоматизированных информационных библиотечных систем (АИБС) предложили заполнять дополнительные поля, в которые можно вводить аннотации, названия рубрик (тематических, хронологических и др.) и ключевые слова (КС).

В настоящей статье хотелось бы уделить особое внимание поиску по ключевым словам. Нужно сказать, что этот вид поиска положительно оценили и индексаторы, и пользователи. Последним он хорошо знаком по применению в сетях Интернет.

Индексирование по ключевым словам стало обязательным элементом библиотечно - информационных систем. В то же время этот вид индексирования является ахиллесовой пятой для библиотек, так как методические материалы в этой области практически отсутствуют. Естественно, что библиотеки пытаются найти выход из создавшегося положения, разрабатывая методики по формированию тезаурусов на основании собственного опыта, не подозревая о подстерегающих из трудностях.

Ситуация усугубляется и тем, что число библиографических записей в создаваемых электронных каталогах растет с каждым днем. И стоит задуматься, насколько востребованными или конкурентоспособными окажутся они в связи с предстоящей интеграцией в мировое информационное пространство. Нужно заметить, что чем больше проходит времени, тем масштабнее будет редактирование, без которого не обойтись. О его необходимости свидетельствует хотя бы то, что в данный момент при индексировании документа по ключевым словам большую роль играет субъективный фактор, т.е. полнота и точность раскрытия содержания документа прямо зависит от уровня квалификации библиотечного работника, его знакомства с проблематикой отрасли.

В основном выбор ключевых слов и формы их представления определяется по интуиции и не связан ни с какими правилами. Анализ результатов индексирования документов по ключевым словам в наиболее успешно занимающихся автоматизацией массовых библиотеках Москвы, Перми и других городов России показал, что одна и та же книга, проиндексированная разными сотрудниками даже одного и того же отдела, получает совершенно несопоставимые поисковые образы документов.

Один из ведущих библиотековедов России Э. Р. Сукиасян так охарактеризовал положение дел в области индексирования:

"В силу определенных обстоятельств наши каталогизаторы были поставлены перед фактом, получив автоматизированные библиотечные системы, в которых единственным поисковым инструментом, обеспечивающим тематический, содержательный поиск, был поиск вербальный. "Пишите слова," - призывали разработчики этих систем, иногда называя эти "слова" предметными рубриками, а чаще ключевыми словами, но никогда не указывали на то, откуда их взять. Брали, как показывает анализ, кто откуда, по - русски говоря, "как Бог на душу положит". Чаще всего прямо из заглавия, с титульного листа… С началом "накопления массива" становится ясно, что поиск в нем осуще-ствлять будет трудно."

Таким образом, необходимость общего методического решения по формированию ключевых слров очевидна и безотлагательна. Какие же методическими материалами располагают библиотеки ?

Основными документами, которые могут быть использованы при индексировании по колючевым словам,можно считать:

1. ГОСТ 7.25 - 80. Тезаурус информационно-поисковый одноязычный. Правила разработки, структура, состав и форма представления.

2. ГОСТ 7.66 - 92. Индексирование документов. Общие требования к координатному индексированию.

3. ГОСТ 7.59-2003. Индексирование документов. Общие требования к систематизации и предметизации.

4. СТБ 7.74-2002. Информационно - поисковые языки. Термины и определения. В этом документе дано определение понятия "язык ключевых слов".

Имеются также, правда, весьма немногочисленные, и ведомственные разработки, в частности, "Методика составления ключевых слов для электронного каталога библиотеки ВГУ им. П.М. Машерова" (Витебск, 2003.- 13 с., рукопись), "Инструкция по индексированию входного потока документов для формирования массива электронного каталога" (М-во культуры и печати РБ, НББ. - Мн., 1995).

Индексирование по ключевым словам признается специалистами наименее эффективным, потому что непосредственно в процессе индексирования создается неуправляемый словарь. Нормализовать такой словарь можно за счет упорядочения синонимов, введения стандартных подзаголовков, унификации форм наименований.

В одной из публикаций, посвященных рассматриваемой проблеме, Ф. С. Воройский пишет:

> "Качественное индексирование достигается, если систематизатор стремится к полноте и точности отображения содержания документов, отражая все аспекты рассматриваемого в документе предмета или объекта. При этом в первую очередь должны учитываться информационные интересы пользователей системы, представлять логику и принципы поиска информации потребителями. Изучение результатов поиска при обслуживании читателей, корректировка и дополнение поисковых образов документов позволяют сделать электронный каталог или базы данных информативными, с высокими поисковыми возможностями.
>
> Другой составляющей успеха высокого качества индексирования является составление словарей ключевых слов по единой методике, при этом учитываются особенности индексируемого массива документов, т.к. он предназначается в первую очередь для потенциальных потребителей конкретной библиотеки или определенной отрасли."

Технологические процессы и основные правила индексирования документов изложены в п.6 ГОСТа 7 - 59 - 2003.

Общие правила индексирования состоят из следующих взаимосвязанных процессов:

1. анализ содержания документа как объекта индексирования;

2. выявление и отбор смысловых компонентов в содержании документа;

3. принятие решения о составе поиского образа индексируемого документа;

4. оформление отобранных смысловых компонентов как понятий в терминах индексирования в соответствии с системой грамматических средств информационно - поиского языка;

5. редактирование терминов индексирования.

Непосредственно с этими процессами связаны процессы составления правил и словаря КС.

Анализ содержания документа - одна из важнейших операций, способствующих повышению раскрытия информационного содержания документа в соответствии с наибольшим читательским спросом, всех важнейших для потенциального пользователя аспектов содержания.

В зависимости от вида индексируемого документа анализу подлежат:

- заглавие,
- сведения, относящиеся к заглавию,
- оглавление,
- аннотация или реферат,
- иногда и отдельные фрагменты текста.

Анализ содержания документа заканчивается составлением мысленной аннотации, в которой отражаются основные и второстепенные темы документа, представляющие интерес для пользователей ИПС.

Составление мысленных аннотаций - процесс творческий, в котором большую, если не определяющую роль играет личностный фактор, т.е профессиональный уровень лица, осуществляющего индексирование. Тогда и субъективность принимаемых решений сводится к минимуму.

Из чего складывается этот уровень? На наш взгляд, прежде всего из двух составляющих:

- знания предметной области индексируемого документа;
- знания правил составления ключевых слов.

Последнее включают в себя как решения, заложенные в типовых методиках, так и решения, принятые индексаторами конкретного потока документов. Для предотвращения расхождения в индексировании необходимо стандартизировать построение поисковых образов.

Составление словаря ключевых слов происходит путем:

- отбора из заглавий, аннотаций, рефератов и текстов документов слов естественного языка, которые могут использоваться в поисковых образах документов (ПОД) и поисковых предписаниях (ПП);
- разработки классификационных схем, отражающих иерархию и соподчинение используемых терминов в данной предметной области.

Важнейшее требование к словарю ключевых слов – полнота охвата терминологии, так как словарь включает термины, которые фигурируют в документах, вводимых в ИПС.

В зависимости от того, какой подход будет использован при формулировке ключевых слов, будут получены различные словари. Применение унитермов позволяет свободно использовать элементы поисковых образов, обеспечивает глубокое и детальное индексирование, увеличивает количество “точек доступа” к разыскиваемым документам. Вместе с тем в ряде случаях оно оказывается недостаточным.

Разделение устойчивых словосочетаний, соответствующих определенным научно - техническим понятиям, грозит потерей информации при поиске, т.к. определенные понятия не всегда могут быть выражены единичным термином. Иными словами, ключевые слова принимают в словарь с учетом точки зрения поиска информации для каждого ключевого слова отдельно и их лексикографической обработки.

Распространенным считается правило: в первую очередь следует ориентироваться на единичные ключевые слова, сохраняя устойчивые словосочетания, удовлетворяющие приведенным выше лингвистическим и прагматическим критериям.

Для более полного отражения содержания индексируемого документа, как правило, достаточно 10 -12 ключевых слов, не являющихся синонимами.

Кроме этих правил, индексаторы могут, в зависимости от особенностей индексируемого потока документов и запросов потребителей, выработать собственные. Эти правила следует зафиксировать и придерживаться их при создании электронных каталогов и баз данных.

Любой словарь ключевых слов со временем нуждается в редактировании, потому, что обрастает большим количеством синонимов, омонимов, нарушений правил при составлении ключевых слов. Чтобы избежать глобального редактирования, необходимо уже при формировании словников продумать способы устранения явлений, мешающих поиску информации.

В заключение хотелось бы предложить не только библиотекам, но в первую очередь методическим центрам обратить пристальное внимание на данную проблему: ведь от её грамотного решения по - существу зависят возможности информационного поиска, а тем самым и судьба создаваемого библиотекой справочного аппарата - прежде всего на электронных носителях.