ЯЗЫКИ ТЕМАТИЧЕСКОГО ПОИСКА В ЭЛЕКТРОННОМ КАТАЛОГЕ

© Г. А. Скарук, 2009

Использование в отечественных электронных каталогах (ЭК) широкого набора разнообразных по структуре, лексическому наполнению, свойствам лингвистических средств представляет собой устойчивую тенденцию. Между тем дискуссия об оптимальном наборе информационно-поисковых языков (ИПЯ) далеко не завершена.

Это вынуждает библиотечное сообщество на новом уровне обратиться к задаче оптимизации комплекса лингвистических средств, т. е. определения состава лингвистического обеспечения (ЛО) и параметров каждого из поисковых языков, необходимых и достаточных для эффективного функционирования системы в целом.

Предметные рубрики (ПР) наряду с ключевыми словами имеют самую длительную историю использования в ЭК. В автоматизированном режиме стали очевидными многие ограничения языка предметных рубрик (ЯПР), которые в карточных каталогах были не так заметны. Особенно острая критика в адрес этого языка звучала на начальных этапах создания ЭК. Тогда он воспринимался как «наследие системы карточных каталогов». Вопрос о бесперспективности системы предметных рубрик Библиотеки Конгресса США ставился даже в 1996 г. [1], по истечении более чем четвертьвекового периода развития Online Public Access Catalogue (OPAC). Поэтому подверглись изменениям практически все элементы этого поискового языка.

'. Е. Б^леП [2] так обозначила основные этапы развития ЯПР в США: с конца XIX в. до 70-х гг. XX в. - период следования традиционным правилам каталогизации, разработанным Ч. Кетте-ром, в 1970-1980-х гг. (с началом эпохи электронных каталогов) основное внимание уделялось созданию полного универсального свода правил по формулированию ПР. В 1990-х доминирует стремление к максимальному учету интересов пользователей. К настоящему моменту основная задача -адаптация языка предметных рубрик к различным видам информационной среды, в том числе электронной. По всей видимости, в создании универсального общероссийского списка предметных рубрик нам предстоит пройти все эти этапы, и сейчас мы находимся на втором из них. Хотя имеются рекомендации использовать в качестве общесетевого лингвистического средства список предметных рубрик Библиотеки Конгресса США [3], не подтвержденные, впрочем, серьезным анализом рекомендуемой системы.

В современных ЭК возможен поиск либо по каждому подзаголовку, либо по каждому отдельному слову предметной рубрики. Так, документ с предметной рубрикой «Питание спортсменов -гигиена» будет выдан в ответ как на отдельные запросы «Питание», «Спортсмен$», «Гигиена», так и на объединенный запрос «Питание» \* «Спорт-смен$» \* «Гигиена». Таким образом, читателям становится доступен поиск по произвольному сочетанию признаков.

Снижение эвристической функции ведущего слова в ПР в авторитетном файле и непосредственно в поисковом образе документа изменило и подходы специалистов к оптимизации структуры сложных рубрик. В методической литературе предлагаются следующие шаги: отказ от описательных формулировок в пользу сложных рубрик, ужесточение правил принятия решений о применении инверсии [4], использование в структуре ПР реляторов [5].

Это приводит к сближению свойств ЯПР и де-скрипторного поискового языка. Менее значимыми становятся и отличия ЯПР от иерархических классификаций. Это касается прежде всего локализации отдельных поисковых признаков в линейных структурах лексических единиц (ЛЕ), в том числе отмеченная Е. М. Ручимской [6] множественность локализации отраслей по отношению к одному предмету в ЯПР в противовес множественности локализации предметов в иерархических классификациях.

Но в результате теряется характерное отличие языка предметных рубрик - жесткая связанность терминов (заголовков и подзаголовков) в текстах ПР. Принятая в карточных предметных каталогах форма ПР, во-первых, исключает ложную координацию при поиске и, во-вторых, позволяет пользователям «увидеть сочетаемость терминов» [7] и уточнить контекст темы уже при составлении поискового предписания.

Возможность сохранить все перечисленные преимущества, увеличив в то же время гибкость поиска, дает изменение способа организации машиночитаемого нормативного словаря - авторитетного файла предметных рубрик. Появляются разнообразные дополнительные способы доступа к нормативным записям. Возможен доступ пользователя:

• к ПР, где названное пользователем слово является ведущим (способ поиска в карточных предметных каталогах);

• ко всем ПР словаря, содержащим любое из слов запроса;

• к ПР, максимально близким к полному тексту запроса.

Среди проблем, унаследованных от карточных каталогов, F. Calderon называла затруднения при использовании ЯПР для поиска информации об общих понятиях и по широким темам [8].

Если вы в процессе поиска по словам из предметных рубрик приводите какое-то широкое понятие, то более узкие, подчиненные ему понятия, не войдут в выдачу, если для их выражения в нормативном словаре существуют самостоятельные, более точные предметные рубрики. Например, задав поисковое предписание «цветны$» \* «металл$», в результате поиска в ЭК вы получите только опи-

сания книг, посвященных цветным металлам в целом. Книги только о медных рудах, никелевых рудах, серебряных рудах и т. д. в выдачу не войдут, так как для описания их содержания существуют самостоятельные рубрики, которые и были приведены в поисковых образах документов. Для такого поиска более целесообразно использовать иерархические классификации.

Иерархические классификации имеют обширную и длительную практику применения в карточных каталогах. В электронной среде классификации приобретают действительно уникальные свойства. Их возможности принципиально отличаются от свойств различных типов вербальных ИПЯ.

На ранних стадиях автоматизации библиотек высказывались мнения о том, что классификации устарели. Но эти доводы сегодня признаны несостоятельными [9]. Более того, очевидно, что возможности классификационных ИПЯ в электронной среде намного расширились.

В отличие от языка предметных рубрик, иерархические классификации позволяют создавать и просматривать обширные тематические массивы библиографических записей по несколько иному (отраслевому) принципу. При этом, меняя длину индекса путем усечения в ту или иную сторону, мы либо сужаем, уточняем, либо расширяем тематические рамки обзора. Подобными возможностями в таком объеме не располагает больше ни один поисковый язык ЭК.

Поиск по усеченным классификационным индексам обеспечивает дополнительные (по сравнению с карточными каталогами) и действительно уникальные возможности. Эта процедура предполагает автоматическое включение в выдачу всех лексических единиц, иерархически подчиненных приведенным в запросе. Таким образом, с помощью классификационного ИПЯ мы имеем возможность в кратком поисковом предписании объединить большое число понятий по отраслевой теме.

Например, если пользователя интересуют книги о болезнях органов дыхания (а это бронхиальная астма, пневмонии, бронхиты, плевриты и т. д.), нет необходимости перечислять их все. Достаточно привести в ПП индекс ББК с усечением «Р412$»1 и в поле поиска будут включены индексы: Р 412.2 Болезни бронхов Р412.201.1 Бронхиты Р 412.202 Бронхиальная астма Р 412.3 Болезни легких Р 412.301.1 Пневмонии и т. д.

Такой поиск дает возможность, не изменяя объема отраженного в индексе понятия, существенно расширить поле обзора и объем выдачи.

1 Примеры здесь и далее приводятся по полным таблицам ББК.

Рассмотрим основные проблемы поиска с использованием иерархических классификаций.

Иерархическое структурирование выдачи. Одной из проблем поиска в ЭК остается сложность тематического структурирования больших массивов библиографических записей, выданных в ответ на широкотематические запросы. Такой подход обеспечивался структурой карточных систематических каталогов. А значит, снижается обозримость представленного пользователю информационного массива. Структурирование выдачи по иерархическому принципу гипотетически возможно, но пока такая возможность нигде не реализована.

Понимание смысла индексов. Практики указывают на то, что без специальных знаний или вспомогательных средств пользователи ЭК часто затрудняются оценить содержание буквенно-цифровых кодов.

Недостаточная гибкость. Иерархические классификации имеют наиболее жестко структурированный по ограниченному числу оснований и связей нормативный словарь. Поэтому могут возникнуть затруднения при поиске по произвольному сочетанию признаков, по узкопредметным и межотраслевым запросам.

Но использование этого языка в ЭК в принципе позволяет преодолеть ограничения, связанные с линейностью сложных индексов. Предлагаются различные методы реализации классификационных ИПЯ в ЭК, а именно:

1. Использование полных индексов классификации как единой совокупности знаков, максимально отражающих содержание документа, в том числе поиск по индексам с различной степенью усечения в зависимости от поисковой задачи.

2. Применение иерархической классификации как посткоординатной системы, подразумевающее использование всех фрагментов полного индекса как независимых поисковых элементов, с возможностью усечения каждого элемента.

3. Использование словесных эквивалентов к цифровым индексам.

3.1. Замена каждого индекса в таблицах классификации соответствующей вербальной формулировкой (тезаурус на базе классификации).

3.2. Связывание классификационных индексов с соответствующими терминами «предметного входа» (тезаурус для классификации).

3.3. Связывание классификационных индексов с ЛЕ других вербальных ИПЯ.

Возможности поэлементного поиска по индексам комбинационных классификаций были описаны еще в 1976 г. В. А. Мишиным [10], однако в отечественных ЭК данная стратегия до сих пор на практике не реализована.

Продемонстрировать гипотетические возможности такого поиска можно на примере. Читателю

необходимы книги по истории российского казачества в XVII-XVIII вв. Тему адекватно отражает индекс ББК «Т3(2)46-294», но документы с этим индексом в систематическом каталоге не найдены. Используя вышестоящий индекс, читатель получил две книги, содержащие в числе прочей и информацию об указанном в запросе историческом периоде. В ЭК можно еще более эффективно построить поисковое предписание (если бы в современных ЭК была возможность поиска по элементам индексов), используя усечение: «Т3(2$» \* «$-294$». При такой формулировке в выдачу войдут документы не только о всех более широких исторических периодах, но и об истории казачества различных регионов и населенных пунктов России.

Пока же электронные каталоги в основном допускают поиск только по полным индексам с правосторонним усечением.

Аргументацию критиков о применении иерархических классификаций в ЭК по причине их громоздкости и сложности для восприятия опровергает N. Williamson [11]. Она указывает на то, что разработчики часто путают простоту самой классификационной системы с простотой ее использования. В электронном каталоге, в отличие от карточных, появляется возможность создания различных путей поиска для пользователей с разным уровнем библиографической и профессиональной подготовки. Каждый пользователь карточного систематического каталога в любом случае должен был выполнять один и тот же набор шагов. В ЭК же для неопытных пользователей часть этих процедур может осуществляться «по умолчанию». В то же время высококвалифицированные специалисты получают возможность просматривать классификационные таблицы в полном объеме и строить сложные поисковые предписания. Возможности раскрытия структуры и содержания ИПЯ в авторитетных файлах гораздо шире, чем в печатных таблицах классификации.

Судя по публикациям, под поиском по ключевым словам подразумеваются достаточно разнородные процессы: поиск по ключевым словам, отобранным из текста документа на основании различных методик, при этом поисковый образ документа может быть дополнительно обогащен либо индексатором, либо с помощью автоматических процедур; поиск по полным текстам аннотаций и рефератов; по полным текстам заглавий и по наиболее информативным словам заглавий; а также по словам из традиционных предметных рубрик в режиме посткоординации. Каждый из перечисленных методов поиска имеет свои особенности. В современных ЭК чаще всего функция поиска по ключевым словам предполагает поиск по отдельным словам из всех текстовых полей биб-

лиографических записей: заглавия, предметных рубрик, ненормируемых ключевых слов, аннотации или реферата (если они есть).

Мнения по поводу роли языка ключевых слов (ЯКС) в составе ЛО ЭК очень разнообразны: от высказываний о целесообразности придания ЯКС статуса основного языка каталога крупной библиотеки [12, 13] до предложений об использовании его в качестве дополнительного поискового средства [14]. Подводя итоги эксперимента по поиску в электронном каталоге на ЯКС, Н. П. Никольцева,

О. А. Седышева делают вывод, что этот язык будет интересен прежде всего для узкоспециальных, профильных библиотек, библиотек, не имеющих в составе ЛО языков с развитым нормативным словарем, а также специальных БД в составе ЭК (например, краеведческих) [15]. Столь же разнообразны и высказывания различных авторов по поводу функций ключевых слов в автоматизированном поиске по отношению ко всем типам автоматизированных информационно-поисковых систем. Некоторые специалисты говорят о том, что дескрипторные ИПЯ уступают лидирующие позиции в области автоматизированного поиска языку ключевых слов, хотя отмечают в этой связи снижение показателей качества поиска [16]. В то же время Н. И. Гендина пишет об ошибочности тезиса об «отмирании» в условиях развития новых информационных технологий всех иных ИПЯ, кроме языка ключевых слов [17]. J. Rowley называет в качестве преобладающей тенденцию сочетания при поиске естественных и нормализованных языков индексирования [18].

Общим признаком всех перечисленных методов является значительно меньшая степень лексического контроля при формировании поисковых словарей или же его отсутствие. Поэтому при составлении поискового образа запроса рекомендуется перечислять все возможные синонимы каждого слова в наименовании темы. Так происходит, например, когда пользователю ЭК нужна информация о психологии детей в возрасте от 14 до 18 лет. Педагоги, как правило, называют их старшеклассниками, врачи и психологи могут выбрать другие варианты. Поэтому в запросе указывается несколько наименований:

(«старшеклассни$» + «дет$» + «подростк$» + «юнош$») \* «психологи$».

Как и для дескрипторного поискового языка, основным достоинством считается возможность поиска по любому сочетанию признаков, по новым проблемам, узкотематическим запросам, терминология которых не отражена в контролируемых словарях.

Проблемы поиска. В плане помощи пользователю возможности ЯКС очень ограниченны. Поиск по ключевым словам предполагает большую

интеллектуальную нагрузку на пользователя и применение более сложной поисковой логики. Эти свойства очень затрудняют использование ключевых слов при самостоятельном индексировании запросов неподготовленными пользователями, на которых ориентированы современные ЭК. Это означает, что обеспечить удовлетворительную полноту поиска способен только пользователь, хорошо владеющий терминологией отрасли.

Однако практики говорят о предпочтении пользователями этого языка. В то же время следует учитывать, что пользователь выбирает язык общения с системой (результаты использования которой он не всегда может предвидеть), но никак не стратегию поиска в каталоге. Имея в виду, что в ЭК язык индексирования запроса может отличаться от языка индексирования документа [13], а также существует возможность автоматического выбора поисковой стратегии, приведенные данные нельзя рассматривать как однозначный аргумент в пользу вербальных ИПЯ. Решающую роль при выборе ИПЯ должны играть требуемые результаты поиска. Но на текущей стадии развития лингвистического и программного обеспечения ЭК совпадение плана выражения естественного языка и ИПЯ, безусловно, имеет большое значение. Поэтому необходимо выявлять для пользователей специфические свойства ЯКС или в процессе обучения, или в процессе диалога с поисковой системой.

Подведем итоги. Изменение принципов реализации поисковых процедур и возможностей доступа к элементам сложных индексов в условиях автоматизации закономерно приводит к сокращению различий между предкоординатными и посткоор-динатными ИПЯ, обусловленных структурой и порядком следования отдельных элементов индексов. В то же время большую значимость приобретают принципы образования классов понятий и специфика плана выражения, а также принципы структурной организации нормативных словарей, характерные для цифровых иерархических классификаций и вербальных поисковых языков. Обусловленные этими свойствами функциональные различия на данный момент являются определяющими при использовании в ЭК различных лингвистических средств.

Каждый из рассматриваемых ИПЯ имеет ряд устойчивых свойств, которые в одних вариантах поиска могут проявиться как несомненные преимущества, а в других - привести к неудаче. Соответственно, и сторонники, и критики применения того или иного поискового языка в ЭК находят в каждом случае веские аргументы. Закономерно предположить, что для реализации различных типов поисковых запросов будут эффективны различные ИПЯ. В этой ситуации следует говорить о

преимуществах того или иного ИПЯ по отношению к каждому конкретному запросу и требованиям пользователя по его реализации. А значит, возможен дифференцированный подход к выбору ИПЯ в зависимости, с одной стороны, от специфических характеристик запроса и предъявившего его пользователя, а с другой - от специфических свойств ИПЯ определенного типа.

На основе анализа требований к ЛО, предъявляемых в рамках различных концепций тематического поиска, получены выводы о противоречивости одних требований и взаимодополняющем характере других. В частности, в противоречие вступают требования: с одной стороны - упрощения из-за расширения самостоятельного поиска непрофессиональными пользователями, а с другой -большей гибкости и семантической мощности ввиду усложнения решаемых задач; с одной стороны - унификации в связи с созданием сводных каталогов и корпоративных сетей, а с другой -большей дифференциации в целях «персонализации» процесса поиска. Следовательно, для решения конкретных поисковых или технологических задач необходимо построение иерархии критериев и требований, на основе которых осуществляется их реализация и контроль качества выполнения, согласованной с системой требований и предпочтений пользователей ЭК.

Вместе с тем в различных концепциях информационного поиска присутствуют совпадающие требования: множественность используемых лингвистических средств, разнотипность ИПЯ, дифференциация методов и процедур их использования в зависимости от характеристик пользователя ЭК, наличие механизмов помощи в формулировании и корректировке поискового предписания.

Сопоставление мнений о свойствах отдельных типов информационно-поисковых языков позволило сделать вывод о взаимодополняемости их поисковых возможностей, что подтверждает целесообразность применения в ЭК комплекса взаимодополняющих языков на принципах адаптивного информационного поиска.