Интеллектуальный информационный поиск

Автор: Э. Р. Сукиасян

УДК 025.4.03

Рассмотрены возможности обеспечения интеллектуального, развивающего поиска в режиме диалога с электронным каталогом, с одновременным использованием двух классификационных систем - УДК и ББК.

Ключевые слова: интеллектуальный информационный поиск, диалоговый поиск, информационно-поисковые языки, классификационные системы, Универсальная десятичная классификация, Библиотечно-библиографическая классификация.

Термин *информационный поиск* ГОСТ 7.73 - 96 (п. 3.1.3) определяет как действия, методы и процедуры, позволяющие осуществлять отбор определенной информации из массива данных. Там же можно найти определения для документального поиска (в этом случае речь идет о поиске документов, содержащих информацию), в том числе - автоматизированного (с использованием ЭВМ), библиографического (если поиск ведется в библиографической базе данных) и др. Меня больше других интересует, в частности, диалоговый поиск, когда пользователь может формулировать информационные запросы в диалоговом режиме, корректировать их в процессе поиска и получать промежуточные результаты.

Что такое диалог? Это коммуникативный процесс с участием двух сторон, одна из которых - пользователь, а другая - информационно-поисковая система. Никого в этом случае не интересует, ручная она, механизированная или автоматизированная. Понятно, что пользователь в любом случае взаимодействует с ИПС, используя язык. Сразу же вводить понятие информационно-поискового языка (ИПЯ) не хочется. Ведь пользователь может работать с ИПС (например, с карточным каталогом) на естественном языке. В начале работы выберет в перечне классов нужное подразделение, затем каталог подскажет (с помощью внешнего оформления, ссылочного аппарата) путь поиска.

Классификационный ИПЯ в систематическом каталоге (СК) - это не только классификационные индексы (сочетания знаков), но и естественный язык, на котором легко и просто общаться: ведь на ящиках и разделителях мы видим не символы, а слова естественного языка, рядом с которыми сто-

---

ят индексы. Эти же индексы и наименования делений читатель видит на полках, если фонды библиотеки открыты. Мы должны помнить: открытый доступ к фондам - столь же важный критерий перехода к открытому информационному обществу, как и открытый доступ к информации. Рано или поздно мы поймем, что

речь идет обо всех фондах, а не только той их части, которая нами отобрана в так называемый подсобный фонд. О "полочной" функции классификационных систем часто забывают.

Считается, что предметизационный ИПЯ - язык предметных рубрик (ПР) карточного предметного каталога (ПК) - тоже естественный. На самом же деле оба языка предварительно элиминированы, очищены, используют множество заранее отобранных терминов и понятий (и, что естественно, не используют массу других терминов и понятий, не употребляемых в этой системе). Модель реализации задач поиска очевидна: пользователь с помощью терминов индексирования (классификационных индексов или ПР) формулирует поисковый образ запроса (ПОЗ) и проводит поиск в каталогах таких документов, у которых окажется аналогичный (или более или менее близкий) поисковый образ документа (ПОД). При совпадении ПОЗ и ПОД мы говорим о релевантности (или пертинентности) полученных результатов.

Это в теории. А на практике мы всего этого не замечаем, пользуемся языком, на котором привыкли общаться. Понятно, что специалист, знающий более или менее свободно хотя бы два языка (русский и иностранный) имеет преимущества в сравнении с владеющими одним. А еще большие возможности у знающих два-три иностранных языка. Меня поймут многие, кто пытался хоть что-нибудь найти в ПК немецких библиотек, не владея свободно немецким. Успешный поиск в ПК во многом определяется знанием двух лингвистических особенностей, присущих каждому языку: нормами словообразования и порядком слов в предложении. Мне, например, было очень интересно работать с "грузинским" ПК, в котором "всё не так", как в "русском".

В этой связи надо особо подчеркнуть интернациональность классификационного ИПЯ. Если знать "свой" индекс, например, УДК (можете спросить специалистов: они его знают всегда, он указан в записной книжке рядом с номером паспорта), то язык каталога (т.е. тот естественный язык, на котором общаются библиотекари с пользователями) никакого значения не имеет. Находясь в Германии, Венгрии или Нидерландах, даже в Японии, можно открыть каталог с этим индексом, не пользуясь услугами переводчика и не обращаясь к хозяевам.

Мне уже приходилось писать о том, что на этапе "входа" в открытое информационное общество каждому из нас придется решить для себя личную образовательную задачу: сделать один иностранный язык рабочим, чтобы эффективно использовать Интернет. Здесь со мной согласятся многие, может быть - все. Тогда давайте признаем: ИПЯ - это тоже язык. Разные ИПЯ - различные языки, относящиеся, как и естественные языки, к разным языковым группам или семьям. Вот, например, Универсальная десятичная классификация и Библиотечно-библиографическая классификация - языки одной группы комбинационных систем (их неправильно называли полуфасетными, никакой "фасетизации" в них нет). В то время как Классификация двоеточием (КД) Ш. Р. Ранганатана - система из другой группы ИПЯ, правильно называемой аналитико-синтетической (или фасетной). На изучение такого ИПЯ придется затратить гораздо больше времени. Но если вы возьметесь изучать Классификацию Библиотеки Конгресса США (КБК),

неожиданно выясните, что она элементарно проста. Для специалиста это понятно: КБК относится к группе перечислительных систем.

Что можно сказать, сравнивая УДК и Десятичную классификацию М. Дьюи (ДКД)? Обе системы - комбинационные (до 12-го издания и ДКД была перечислительной), они одной группы и даже одного семейства (у них общий "родитель"), но их развитие с начала прошлого века пошло разными дорогами, что привело к поразительному результату: зная одну систему, почти невозможно разобраться с индексами другой.

Мне, наверное, проще судить об этом. Мой учитель Захарий Николаевич Амбарцумян (1903 - 1970) неоднократно повторял: специалист по классификациям не может и не должен ограничивать круг своих знаний одной системой. Тогда он превратится в систематизатора. Изучать их надо, как иностранные языки: последовательно и достаточно глубоко. В школьные годы мне стала родной Десятичная классификация (таблицы Е. Н. Добржинского, других не было); в годы учебы изучал биографию М. Дьюи и ДКД; вступительный реферат в аспирантуру писал о типовых делениях ББК; затем судьба связала меня с РГБ. Два года вместе с коллегами из ГПНТБ России занимался КД и очень глубоко вошел в эту систему (мне доверили отредактировать статью о ДКД в третьем томе Трудов ГПНТБ). Находясь в 1996 г. в Библиотеке Конгресса США, увидел КБК, затем написал статью об этой системе. "Поздняя любовь" - Государственный Рубрикатор НТИ, который с каждым изданием становится всё лучше и лучше. И я не считаю своё образование завершенным: в Великобритании пока не вышло полностью Второе издание Библиографической классификации Г. И. Блисса, еще одной аналитико-синтетической классификации, системы уникальной.

Не надо, наверное, всем знать всё. Но в нашей стране, как и в сосед-

них с нами странах СНГ, сегодня применяются две системы: УДК и ББК. Те, кто занимаются информационным поиском, тем более обслуживанием пользователей, сбором и обработкой информации, должны знать обе классификационные системы. Надо понять: это разные языки, не совпадающие по своей грамматике (ни по лексике, ни по морфологии и синтаксису). Работая в границах одного из этих двух мощных ИПЯ, вы теряете не только информацию в виде документов, но и чрезвычайно важный для каждого пользователя показатель, связанный с аспектностью его запроса.

Пора, наконец, объяснить название статьи. Когда речь идёт об *интеллектуальном информационном поиске,* мы опираемся, конечно, на понятие *интеллект.* Любой словарь пояснит: интеллект - это мыслительные способности человека. Следовательно, интеллектуал - человек с высокоразвитым интеллектом, человек интеллектуального труда. И далее: интеллектуальный - относящийся к разуму, интеллекту, духовный, умственный; отличающийся высоким уровнем развития интеллекта. Я бы добавил: диалоговый интеллектуальный информационный поиск должен быть развивающим для пользователя и одновременно учитывать уровень

его развития. На то и диалог, чтобы подстраиваться, чувствовать пользователя ИПС.

Много лет назад, для 2-й Международной конференции ИСКО (26 - 28 авг. 1992 г. Мадрас, Индия) я подготовил доклад *"Homo Quaerens (The seeking man): On the problem of development of the reader's cognitive capacities in the searching process" (Человек ищущий. К проблеме развития познавательных способностей читателя в процессе информационного поиска).* Он был опубликован в Индии в трудах конференции. Только через десять лет я решился воспроизвести доклад, выступив на конференции "LIBCOM-2001" в Ершово; затем доклад был напечатан в нашем журнале (Науч. и техн. б-ки. - 2002. - N 4. - С. 73 - 83).

Наблюдая многие годы реакцию читателей, работающих с карточным генеральным СК Российской государственной библиотеки, я убедился: взаимодействие с огромным (представьте: 6400 ящиков) универсальным, хорошо организованным каталогом, когда сама обстановка (читатель сидит в удобном помещении за столом, просматривая карточки в ящике) располагает к неторопливому диалогу, в итоге развивает пользователя.

О познавательных механизмах "общения" с СК я писал еще раньше, в практическом пособии "Систематический каталог" (Москва: Кн. Палата, 1990. - 180 с). Студенты, аспиранты, да и многие специалисты впервые видели в каталоге то, что называлось еще в XIX в. "картой знаний". Еще Н. К. Крупская часто применяла этот термин в значении структурированной информации. Сегодня в Интернете можно прочесть, что термин *карта знаний (mind map)* введен, якобы, Т. Бьюзеном *(Tony Buzan)* в его достаточ-

---

но известной книге о мнемотехнике *"Use your memory"* (1984), переводы которой на русский язык издавались с 1989 г. Ну что делать, память у людей не такая уж "глубокая", а многим, особенно читающим в Интернете, кажется, что в истории человечества в "доинтернетовскую эру" ничего и не публиковалось.

Удивительную картину можно было наблюдать в РГБ, когда читатель открывал таблицы ББК (полное издание 1960 - 1968 гг. в 30 книгах: больше 600 учетно-издательских листов, примерно 4 тома БСЭ, плюс примерно 80 выпусков дополнений и исправлений), затем брал книгу, садился за стол и начинал её внимательно изучать, а еще чаще - сразу переписывать в тетрадь. Наибольшим спросом пользовались, как я заметил, таблицы специальных типовых делений, которых в ББК сотни. Ведь это, по сути, глубоко детализированные и хорошо структурированные "карты знаний" тех или иных общих категорий в пределах той или иной дисциплины. Где такое можно было увидеть?

Сегодня читатели библиотек лишены возможности увидеть таблицы классификации, стало быть, познакомиться со структурированной информацией они не могут. Даже если в библиотеке нашли возможность организовать поиск по классификационным индексам УДК или ББК (такие библиотеки есть, их, как мне кажется, не более 3 - 5% от общего числа), то таблиц классификации (тем более

таблиц определителей или типовых делений) читатель на экране не увидит. По странной для меня причине большинство библиотечных администраторов считают, что таблицы классификации - служебный материал, в них могут разобраться только систематизаторы, но уж во всяком случае не читатели. Напомню: когда только мечтали об ЭК, представляя в популярных книжках их внутреннюю структуру, то обязательным элементом ЭК всегда была "База знаний", под которой понимались классификационные таблицы.

Мне очень хочется задать вопрос: Вы уверены, что поиск, который обеспечивают нам находящиеся в эксплуатации библиотек страны ЭК, поиск, при котором читателю тупо предлагается прямоугольная рамочка с возможностью что-то впечатать, хоть отдаленно можно назвать интеллектуальным? Ведь каталог с читателями не разговаривает, их действия не направляет, не показывает, как и что "вписывать", но зато отличается быстродействием, мгновенно отвечая стандартной фразой: "На ваш запрос ничего не найдено"? Что развивает такой ЭК? И кто сказал, что пользователи всегда так торопятся?

Между тем мы располагаем удивительными возможностями сделать наш поиск эффективным. Так получилось, что у нас параллельно применяются две универсальные классификационные системы, обе глубоко струк-

---

турированы, обе имеют мощную лексику. В технологическом и комбинационном отношении они примерно равны, так как обеспечивают поиск по многим признакам. Часть этих признаков признана ведущими (по ним построены основные таблицы). Другие признаки отражены в таблицах стандартных, повторяющихся понятий (определители в УДК, типовые деления в ББК), которые, в свою очередь, делятся на две категории в зависимости от границ распространения. Те, которые относятся к общим категориям, так и называются, образуя систему таблиц общих определителей и типовых делений общего применения. В границах отдельных наук, дисциплин, комплексов есть собственные таблицы повторяющихся понятий (специальных определителей, специальных типовых делений). В совокупности те и другие способны глубоко и многоаспектно раскрыть содержание документа, обеспечить поиск. Только пользователи этого не знают.

Придется здесь остановить ход размышлений и сказать, что на пути от карточного каталога к ЭК мы забыли об очень важном обстоятельстве. Запрос в ЭК реализуется, если ПОД адекватно совпадает с ПОЗ. Если совпадения не происходит, мы получаем стандартный отказ. От машины не узнаешь, что можно обобщить запрос, что тема, быть может, отражена в сборнике, в монографии, которые в системе стоят выше на одну или две ступеньки. Машина - тупая "железка", обычно говорят нам, она не соображает. Не могу с этим согласиться. Карточный СК - тоже "деревяшка с бумажками", но все знают, что здесь можно найти то, чего и не предполагал. В отличие от ЭК, здесь всё открыто, а разделители содержат массу информации на естественном языке. Хороший библиограф может разработать алгоритм грамотного диалога с пользователем во время работы с каталогом. Я встречал такие каталоги за рубежом. У нас таких нет (если обнаружите, напишите мне: sukias@rsl.ru). Почему? Да ясно ведь: у нас

библиографы делают своё дело, а программисты - своё. И сидят они на разных этажах в разных комнатах.

Вернемся, однако, к пониманию того факта, что у нас в одних библиотеках УДК, в других - ББК. На первом этапе горячие головы решили, что две системы в одной стране - дело совершенно нетерпимое. В одном из НИИ (сегодня - ВНИИКИ) организовали отдел, поручили ему научную тему "Проблема единой классификационной системы в стране". В 1960-х гг. издавались и таблицы УДК, и таблицы ББК. Между руководителями министерств и ведомств тогда и началась "гражданская война"; в сознании закрепились некоторые "пункты", пережитки которых иногда всплывают и сегодня. Больше всех страдали специалисты, так как их "объективизм" не устраивал ни тех, ни других. Меня, работавшего сначала во ВНИИКИ, а потом навсегда оказавшегося в "Ленинке", считали "перебежчиком", так

как я полагал возможным ходить и в ВИНИТИ, а в ГПНТБ России даже напечатали мою книжку о централизованной классификации... по УДК.

С годами руководители сменились, мы стали сотрудничать. Все, как мне кажется, осознали тот факт, что нельзя волевым решением "закрыть" какой-нибудь язык только потому, что на нем не все разговаривают. Наконец все узнали, что наиболее распространенной в мире системой является не УДК, а ДКД (благодаря OCLC), а ББК получила статус Национальной классификационной системы России, признанной зарубежными экспертами. Одним словом, установилось динамическое равновесие.

Но это же неудобно: приходится в двух местах индексировать! Хорошо бы сделать переводчик, еще лучше - автоматический. Поставил индекс УДК, нажал кнопку - получил индекс ББК.

Поступило предложение: "Давайте сделаем переводной словарь". Соглашаюсь: "Давайте, как вы себе это представляете?" Мне отвечают: "Вот Вам таблицы УДК, проставьте рядом индекс ББК". Так и опубликуем: слева УДК - справа ББК?

Самое интересное в этих предложениях: они всякий раз высказываются людьми, которые практически ни с УДК, ни с ББК не работали, систематизацией не занимались, таблиц в руках не держали. Не получится ничего, это ясно. Это не "переводной словарь", а таблицы приведения ("прикладывания") индексов ББК к индексам УДК. Заранее отвечу: в 30 - 70% строчек ничего рядом с индексом УДК написать не получится. И если наоборот составлять такой, простите, словарь (от ББК к УДК), то получится та же самая картина в процентах. Пишу об этом, так как цифры получены в эксперименте, который я проводил в Краснодаре со студентами в конце 1960-х гг. Сравнивались "по наукам" УДК (таблицы под редакцией Е. И. Шамурина, огромный том 1963 г.) и новые таблицы ББК. Наивысший процент совпадения, примерно 80%, оказался в химии. Во всех остальных случаях не только структура - выбранные разработчиками классификационные признаки оказывались разными. Обсудив результаты, мы тогда выдвинули новую рабочую

гипотезу: язык таблиц, т.е. термины, понятия, должен совпасть. Будем сравнивать указатель к таблицам по отраслям. По УДК его пришлось сначала сделать, "выбирая" для каждой отрасли рубрики из единого указателя. Сделали. Не получилось: у 40 - 60% карточек в ящике оказалось только по одному индексу - или УДК, или ББК.

Стало понятно, что не надо строить "переводные таблицы". Появилась другая рабочая гипотеза - использовать язык-посредник. Иначе говоря, "приводить" индексы УДК и ББК в соответствие не друг с другом, а приписывать их к делениям третьей классификационной системы. Сегодня мы приблизились к решению этой задачи, а часть её (в отношении УДК)

выполнена. Есть уже несколько изданий Государственного Рубрикатора НТИ (ГРНТИ, применение системы регламентировано ГОСТ 7.59), в котором индексы УДК проставлены. Мы договорились, что при первой возможности решим вопрос и в отношении индексов ББК. Дать оценку полезности этой работы должны специалисты-систематизаторы. Одно можно сказать уже сейчас, без эксперимента: машина "по-умному" работать всё равно не сможет. Когда рядом с одним индексом оказывается несколько индексов другой системы, принять интеллектуальное решение может только систематизатор.

Если для каждого индекса УДК нашелся бы адекватный ему индекс ББК, никакой проблемы не было. Такие "общности" можно провести разве только на уровне самых первых делений. Чем глубже, тем дальше расхождения между системами. Несколько лет назад проведен серьезный анализ возможностей построения таблиц соответствия классификационных систем. После доклада на конференции "LIBCOM-2007" решили: надо проводить исследования в экспериментах, на небольших отраслевых разделах. ВИНИТИ и РГБ готовы участвовать, помогать.

Моя статья была опубликована в нашем журнале (Науч. и техн. б-ки. - 2008. - N 8. - С. 36 - 40). Она заканчивалась словами: "Если мы хотим иметь "Словарь соответствия", надо, прежде всего, начать с поисков финансирования проекта экспериментальной работы на определенном, четко ограниченном, отраслевом разделе. Сначала на малом массиве отработать технику и технологию, учесть и подсчитать все экономические затраты, определить нормативы, необходимые для дальнейшего планирования. А пока надо понять: "таблицы соответствия" - это фикция, которую не стоит даже обсуждать".

Прошло более трех лет. Серьезных подвижек нет. Однако то и дело кто-нибудь вспоминает о "переводных таблицах". А время идет. На словах все готовы "к борьбе за эффективность ЭК". На деле поиск все больше превращается в техническую операцию перебора выданных записей с последующим глубоким разочарованием. Об интеллектуальном информационном поиске с глубоким диалогом остается только мечтать.