УДК 02-029:004
ББК 78.347.8

Возможности пользователя при поиске информации в электронных библиотеках, или «Витязь на распутье»

Рассматриваются вопросы организации тематического поиска в электронных библиотеках. В любой информационной системе человек оказывается в роли путника, которому предлагаются возможные варианты выбора пути с целью получения результата. В роли пресловутых надписей на сказочных камнях выступают экраны компьютеров с текстами меню. Далеко ли информационные системы ушли от известных с детства текстов типа «направо пойдешь — счастье найдешь»?

Ключевые слова: *тематический поиск, электронная библиотека, электронный каталог, семантические сети, графы, тезаурусы, онтологии, библиотечно-библиографическая классификация, ББК, система представления знаний, словесная формулировка индексов, семантическая паутина, Semantic Web.*

Ольга Александровна Лаврёнова,
заведующая отделом развития компьютерных технологий и лингвистического обеспечения Российской государственной библиотеки, кандидат филологических наук

Варианты формирования путей поиска

В каком бы виде люди ни читали или ни прослушивали книгу, журнал, карту, диссертацию, их сначала нужно найти. Рискнем утверждать, что средства поиска часто играют более существенную роль в информационном обеспечении, чем объемы хранимой информации, а требования к качеству поисковых средств растут вместе с увеличением объемов данных.

Глядя на экран поисковой системы, пользователь нередко чувствует себя «витязем на распутье»: казалось бы определил для себя некую информационную потребность, но не знает, куда идти. Вся надежда на надписи на экране. Что же может быть представлено на нем в форме меню? Как правило, это фраза наподобие следующей: «Введите слово или сочетание слов». В электронной библиотеке с базой данных объемом в сотни тысяч полных текстов это означает примерно следующее: скачи напролом в дебри этих текстов, где и найдешь написанные тобою слова, возможно, в нужных тебе текстах. Короче говоря, получишь «то, не знаю что», а потом разберешься. Хорошо еще, если в таком поиске предусмотрен учет грамматических парадигм. В этом случае хотя бы найдутся твои, витязь, слова сразу во всех грамматических формах. Еще более заботливые системы учтут, что найденное слово не обязательно характеризует содержание текста, а встретилось между прочим. Такие системы,

используя хитрые методы, найдут слова, встречающиеся чаще других или совместно с другими, заданными как поисковые, в одном абзаце или в одном и том же предложении, стоящие рядом или разделенные конкретным количеством других слов. Другие системы, приближаясь к смысловому подходу, проверят, а принадлежат ли найденные слова к некому заранее составленному словарю терминов, специфичных для определенной области знаний. Дебри, конечно, будут прорежены, но загадочность результата все же остается, и многих витязей, т. е. пользователей, это обстоятельство не смущает. Однако повсеместно встречаются такие люди, которым необходимы заранее проложенные пути-дороги да еще и с четкими дорожными знаками. В этом направлении серьезное значение имеют, в частности, разработки лингвистических процессоров и программных продуктов, обеспечивающих формальное обнаружение идентичных фрагментов текстов, а также попытки создания систем автоматизированного индексирования, включая систематизацию.

Следует отметить, что разработчики систем, в которых поиск основан преимущественно на вычислениях сходства последовательностей знаков в тексте запроса и тексте документа или построенных для него метаданных, проявляют чудеса изобретательности в алгоритмах программных продуктов. При этом создаются, в частности, интересные системы, обладающие исключительно полезными функциями. Рассмотрим две из них: «Антиплагиат. РГБ» [1] и T-Libra [3]. Первая система позволяет выявить настолько похожие фрагменты текстов, что проверяемый по полнотекстовой базе данных документ может быть признан системой как возможный плагиат, если совпадения достигают определенной, заранее заданной степени сходства.

Система T-Libra дает, в частности, возможность отыскивать в полнотекстовой базе данных по тематическому запросу не документы, а их фрагменты, релевантные запросу, например, абзацы, т. е. служит подобием сказочного клубка ниток, который приводит путника прямо к цели. Искал жар-птицу — получай 200 штук на выбор. Много лишних — следовало точнее формулировать ее характеристики. Пользователь в этой системе практически получает факты, которые с большой вероятностью касаются запрашиваемой темы или предмета. Варианты реализации системы T-Libra достаточно многообразны, но можно предложить и новые, например, поддержку технологии автоматизированной систематизации на основе ББК или УДК — по аналогичной схеме сравнения текстов документов и текстов словесных формулировок индексов.

Можно утверждать, что в настоящее время наблюдаются *две основные тенденции повышения качества тематического поиска в электронных ресурсах:*

- использование формальных характеристик встречаемости и взаимного расположения поисковых признаков в текстах (по этому поводу имеется интересная информация, например, в материалах Международной конференции по компьютерной лингвистике «Диалог» [7], Европейской конференции по теории и практике электронных библиотек — TPDL [8,15], Российской конференции по электронным библиотекам — RCDL [9]); при этом используется достаточно сложный математический аппарат обработки текстов и результаты тонких исследований в области компьютерной лингвистики;
- создание и использование различных способов (систем) представления (организации) знаний, которые управляют семантическими связями между поисковыми признаками [15].

Системы представления знаний для электронных библиотек

Прошли времена эйфории первых десятилетий создания электронных библиотек (ЭБ), когда многим разработчикам и пользователям ка-

залось, что невероятные объемы полнотекстовых ресурсов сами по себе обеспечат полноту поиска информации, поскольку в них чрезвычайно много слов. К настоящему времени большинству специалистов в области создания ЭБ стало ясно, что высокое качество тематического поиска в полнотекстовых ресурсах возможно только при условии, если система может учитывать при поиске семантические связи между поисковыми признаками (синонимию, иерархические и ассоциативные отношения). На обеспечение этих условий направлены современные проекты создания систем представления знаний, к которым, в частности, относятся онтологии, тезаурусы, классификации. Разумеется, это не новые явления в лингвистике информационных систем, но они едва не стали жертвами иллюзии безграничных масштабов оцифровки данных, хотя успешно использовались с середины XX в. для увеличения полноты поиска в автоматизированных информационных системах (АИС) и, в частности, в электронных каталогах (ЭК) библиотек [5].

Наиболее сложной структурой и большой трудоемкостью разработки характеризуются *онтологии*. Формально онтология как модель представления знаний состоит из понятий, терминов, организованных в таксономию, их описаний и правил вывода. Онтологии в настоящее время строятся только для достаточно ограниченных областей знаний (например, для отдельных разделов астрономии и биологии).

Большое место занимают разработки *информационно-поисковых тезаурусов*, история создания которых в информатике насчитывает уже примерно пять десятилетий. Кстати, наблюдается тенденция преобразования систем предметных рубрик в тезаурусы, так как предметизация была необходима для обеспечения расстановки карточек с предметными рубриками по различным аспектам в предметных каталогах библиотек. При этом парадигматические связи типа иерархических в них развивались слабо. Особый интерес представляет *современный стандарт на тезаурусы ISO 25964* [16], в котором учтены тенденции использования таких словарей для поиска электронных информационных ресурсов в сетях (см., в частности, работы Johan De Smedt и Jutta Lindenthal в материалах семинара [15]). Однако тезаурусы при всем их значении для тематического поиска требуют значительных затрат труда квалифицированных специалистов в каждой области знания и лингвистов, способных эту работу организовать, что делает возможным создание серьезных тезаурусов только по отраслям знаний.

Если мы посмотрим «с холодным вниманьем вокруг», то становится очевидным, что средств и кадров в любых универсальных библиотеках недостаточно для формирования нигде не созданного универсального тезауруса, но при этом существуют глубоко разработанные *библиотечно-библиографические классификации*, в которые вложены знания многих ученых и специалистов. Следует отметить, что в сфере создания систем представления знаний для обработки электронных ресурсов классификациям уделяется большое внимание как перспективным средствам моделирования знаний в электронной среде. Классификации как способы (модели) иерархического представления (организации) знаний управляют семантическими связями между поисковыми признаками.

Классификация как путеводитель

Российская государственная библиотека в соответствии с указанной тенденцией использует *национальную Библиотечно-библиографическую классификацию (ББК)* в качестве основы тематического поиска в интегрированной электронной библиотеке РГБ (как принято в едином ЭК Библиотеки) [6].

Многие библиотеки размещают на веб-сайтах полные классификации для своих ЭК в их исходной форме, но для пользователей не так-то просто конструировать поисковые признаки, если классификация имеет сложную структуру. Современный пользователь хочет получать все немедленно и по умолчанию.

В связи с этим РГБ предусмотрела *два варианта технологии использования семантических связей при поиске информации в ЭК и ЭБ:*

- включение иерархических цепочек формулировок индексов (hierarchical strings of captions) ББК в каждую библиографическую запись;
- размещение в открытом доступе рабочего варианта классификации, связанной с ЭК, в качестве модели представления знаний.

Библиографическая запись в РГБ включает:

- индексы ББК,
- иерархические цепочки словесных формулировок индексов,
- свободные (неконтролируемые) ключевые слова.

Рабочие таблицы ББК будут включать:

- иерархические деревья всех индексов, сформированных при каталогизации различных видов документов в РГБ,
- цепочки словесных формулировок индексов,
- данные о количестве библиографических записей в ЭК для каждого индекса, найденного при поиске, и для индекса вместе с его нижестоящими делениями.

Рассмотрим первый вариант на примере 1.

Пример 1. Запрос для поиска в электронной библиотеке авторефератов и диссертаций: *«Работы по волжской группе финно-угорских языков»*. Одна из списка найденных диссертаций:

Заглавие: *Деривация отрицания в марийском языке : автореферат дис. ... кандидата филологических наук : 10.02.22*
Индекс ББК: *Ш166.32-211*
Словесная формулировка индекса ББК:
Филологические науки. Художественная литература -- Языкознание -- Финно-угорские языки -- Волжская группа языков -- Марийские (мари, черемисский) языки -- Грамматика -- Морфология -- Словообразование
Дополнительные ключевые слова: *деривация, отрицание.*

Понятно, что без расшифровки индекса ББК в библиографических данных эта диссертация не могла быть найдена на данный запрос. Даже в заглавии нет поисковых признаков. Однако пользователю не придется входить в образ витязя на распутье: в иерархической цепочке, представляющей собой фрагмент иерархического дерева, для поисковой системы проложен путь (маршрут) от требуемой темы *«Волжская группа финно-угорских языков»* до любой более узкой темы (например, «*Словообразование в марийских языках*»), что обеспечивает автоматическое обнаружение документов и по данной достаточно узкой теме. При этом пользователь может даже не догадываться о том, что система привела его к нужному документу, имея в библиографических записях описания этого маршрута, и о том, что о «путнике» позаботились заранее те, кто разрабатывал таблицы классификации, составлял классификационные индексы, расшифровывал их цепочками словесных формулировок.

Пример 2. Книга по зоологии, найденная в ЭК РГБ наряду с другими по запросу: «*Зоогеография перепончатокрылых*».

Заглавие: *Структура населения муравьев тайги*
Индекс ББК: *Е691.894.73For-81,0* Словесная формулировка индекса ББК: *Биологические науки -- Зоология -- Систематика животных -- Беспозвоночные -- Членистоногие -- Насекомые -- Перепончатокрылые -- Жалящие -- Муравьи -- Экология*
Индекс ББК: *Е685.9(2Р36),0*
Словесная формулировка индекса ББК: *Биологические науки -- Зоология -- Зоогеография -- Российская Федерация -- Урал*

Данная книга найдена на различных уровнях иерархии не в одной цепочке, но одновременно по двум «маршрутам».

Для книг с 1998 г. обработки и с 2003 г. для диссертаций и авторефератов словесные формулировки индексов в записях ЭК РГБ существуют, но нецелесообразно оставлять вне полноценного тематического поиска документы, обработанные ранее, например те, для которых миллионы записей будут включены в ЭК в результате ретроконверсии карточного каталога. Для ЭБ это тоже важно, так как и оцифрованные старые книги тоже требуют эффективных средств поиска. Соответственно, следующая задача состоит в *«декодировании» индексов в максимально возможном количестве существующих записей на книги, авторефераты и диссертации*. Они должны получить иерархические цепочки словесных формулировок. Работы в этом направлении основаны *на идее восстановления словесных формулировок «старых» индексов по аналогии* с теми индексами, которые уже имеют словесные формулировки. Понятно, что в исходном машиночитаемом эталоне таблиц ББК могут быть только исходные индексы, но никак не составленные систематизаторами при обработке поступающих в библиотеку документов. Многолетний опыт показывает, что программы, помогающие систематизатору составить индекс классификации (называемые «рабочее место систематизатора»), как правило, сложны как для разработки, так и для использования специалистами в этом деле и совершенно бесполезны для конечного пользователя. Достаточно присмотреться к работе опытного и, тем более, неопытного систематизатора — и становится ясным, что он, стараясь, как и всякий нормальный человек, экономить время и силы, но при этом не терять качества результатов, использует готовые индексы, составленные ранее для других документов. Тот, кто хоть раз попробовал построить индекс ББК по истории, биологии или философии, поймет сложность этого труда. Систематизатор

заимствует их из библиографических записей ЭК или из составленных им же самим перечней и дорабатывает в соответствии с тонкостями содержания документа. Мало того, много лет назад в РГБ были созданы машиночитаемые таблицы ББК, в которых программно сформированы готовые цепочки словесных формулировок для соответствующих индексов эталона таблиц, и загружены в качестве справочника в систему АЛЕФ, поддерживающую АИБС РГБ [11]. Из этих цепочек можно было составлять более сложные для индексов, создаваемых в библиографических записях ЭК. Оказалось, и этого мало. Считалось, что в предыдущем (собственном) программном обеспечении ЭК — АИС МЕКА — технология была более удобной. В ней при вводе в новую запись индекса, уже имеющегося хотя бы в одной записи ЭК, в нужное поле составляемой записи программно включалась полная цепочка словесных формулировок этого индекса, ранее составленная каким-либо систематизатором в другой библиографической записи. Затем данные можно было дорабатывать.

Таким образом, в Библиотеке появился проект *создания* упомянутых выше *машиночитаемых таблиц ББК, названных «рабочими»*, так как они, по сути, формируются в процессе систематизации. Навигация по таким таблицам и составляет основу второго варианта использования семантических связей при поиске информации.

Необходимо отметить, что развитие данного проекта вряд ли было бы успешным без идей и разработок таких специалистов РГБ, как А.И. Вислый, А.А. Винберг, Т.В. Аветисова, Т.И. Жебрунова.

Построение рабочих таблиц ББК

Вопрос: что же может стать источником исходных данных для машиночитаемых таблиц? В результате дискуссий выбрано *три основных источника данных:*

- разделители Генерального систематического каталога (ГСК) РГБ как основной источник;
- библиографические записи из ЭК (индексы и цепочки словесных формулировок) — как дополнительные источники готовых цепочек на более низких уровнях иерархии, чем на разделителях ГСК;
- машиночитаемый эталон (полные таблицы) ББК — для уточнения структуры данных и терминологии.

Почему ГСК определен как главный источник данных? Разделители систематического каталога расставлены по иерархии индексов ББК. За разделителем расположено ограниченное количество карточек с библиографическими записями, на которых имеются индексы ББК, начинающиеся с индекса на разделителе или полностью совпадающие с ним. Задача заключается, собственно, в том, чтобы восстановить для индексов в старых записях формулировки хотя бы до уровней разделителей.

Как известно, на каждом разделителе указан индекс или его конечная часть и словесная формулировка непосредственно «конца» индекса. С целью создания рабочих таблиц *оцифрованы разделители ГСК* РГБ и получены деревья «индекс — его конечная формулировка».

Пример 3. Фрагменты иерархического дерева разделителей ГСК.

Щ Искусство. Искусствознание
Щ31 Музыка
Щ315 Инструментальная музыка
Щ315.3/9 Музыкальные инструменты. Инструментоведение
Щ315.31 Старинные инструменты
Щ315.32 Народные инструменты
Щ315.4 Клавишные инструменты
Щ315.41 Клавикорд. Клавесин (Чембало). Спинет
Щ315.42 Фортепиано

По данному примеру можно судить о том, что обычной сортировкой индексов по знакам для формирования иерархических деревьев в ББК не обойтись. Были разработаны специальные *программные продукты* для сортировки данных, обнаружения в каталоге пропусков разделителей, т. е. уровней иерархии. С помощью экспертов в области систематизации организовано восстановление пропусков индексов со словесными формулировками, а в настоящее время проводится профессиональное редактирование файлов по всем тематическим разделам. После редактирования из деревьев разделителей программно формируются *иерархические деревья, состоящие из полных цепочек для индексов.* Автором этого комплекса программ является Т.В. Аветисова.

Пример 4. Получение цепочки словесных формулировок для *Щ315.42* «*Фортепиано*» из примера 3.

Щ315.3/9 Искусство. Искусствознание -- Музыка -- Отдельные виды музыки и музыкального исполнения -- Инструментальная музыка -- Музыкальные инструменты. Инструментоведение
Щ315.31 Искусство. Искусствознание -- Музыка -- Отдельные виды музыки и музыкального исполнения -- Инструментальная музыка -- Музыкальные инструменты. Инструментоведение -- Старинные инструменты
Щ315.32 Искусство. Искусствознание -- Музыка -- Отдельные виды музыки и музыкального исполнения -- Инструментальная музыка -- Музыкальные инструменты. Инструментоведение -- Народные инструменты

Щ315.4 Искусство. Искусствознание -- Музыка -- Отдельные виды музыки и музыкального исполнения -- Инструментальная музыка -- Музыкальные инструменты. Инструментоведение -- Клавишные инструменты
Щ315.41 Искусство. Искусствознание -- Музыка -- Отдельные виды музыки и музыкального исполнения -- Инструментальная музыка -- Музыкальные инструменты. Инструментоведение -- Клавишные инструменты -- Клавикорд. Клавесин (Чембало). Спинет
Щ315.42 Искусство. Искусствознание -- Музыка -- Отдельные виды музыки и музыкального исполнения -- Инструментальная музыка -- Музыкальные инструменты. Инструментоведение -- Клавишные инструменты -- Фортепиано

Из индексов с цепочками будут программно строиться деревья для рабочих таблиц ББК.

Обеспечение навигации по классификационным маршрутам

Условия для поиска по рабочим таблицам

Схема технологии изображена на рис. 1.

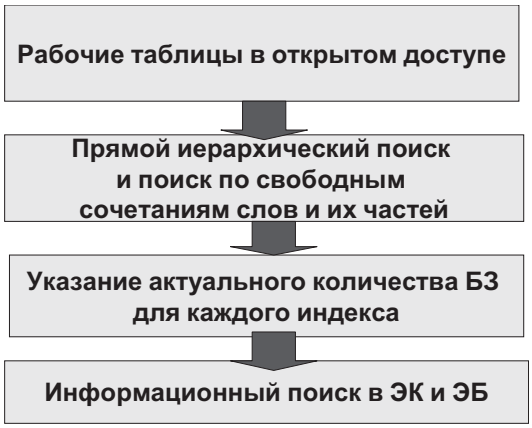

Рис. 1. Схема функционирования рабочих таблиц ББК при поиске

Таблицы планируется разместить в открытом доступе и функционально связать с ЭК. Пользователь сможет проводить поиск путем навигации с верхнего уровня нужного раздела ББК или задать поисковые признаки для прямого выхода на деления более низких уровней [14]. Для каждого деления должно быть указано количество библиографических записей в ЭК именно с найденным индексом, а также количество записей со всеми индексами более «длинными» (в нижестоящих разделах). Пользователь перемещается вверх или вниз по иерархическому дереву, выбирая пути поиска в соответствии с собственными решениями, и получает на экране библиографические записи для выбранного деления.

Формирование маршрутов поиска в библиографических записях

Для создания возможно большего числа готовых маршрутов для пользователей, у которых нет времени или желания прокладывать собственные пути на каждом распутье, была, как упоминалось выше, поставлена задача дополнения индексов в «старых» библиографических записях цепочками их словесных формулировок. Для того чтобы убедиться в работоспособности идеи заимствования цепочек для индексов из базы данных рабочих таблиц по принципу аналогии произведена *экспериментальная проверка технологии программного «декодирования» (расшифровки) индексов* в старых библиографических записях по аналогии с данными в машиночитаемых рабочих таблицах [4, 14]. Для эксперимента отбирались случайные индексы ББК из карточного алфавитного каталога. Специальное программное обеспечение сотрудника РГБ В.А. Калачихина сравнивало каждый индекс с имеющимися в базе данных полными индексами ББК. При обнаружении индекса, максимальное количество начальных знаков которого совпадало с заданным, словесная формулировка найденного индекса приписывалась заданному. Ниже приведены примеры 5—9 некоторых результатов эксперимента.

Пример 5

На входе — *Е693.32-739.1,0* из старой библиографической записи (подчеркнута часть индекса, формулировки которой восстановлены). Система находит в рабочих таблицах путем сравнения индекс *Е693.32-73* и заимствует для заданного индекса цепочку словесных формулировок:
Цепочка: *Биологические науки -- Зоология -- Таксономия животных -- Chordata. Хордовые -- Vertebrata. Позвоночные -- Зоология позвоночных -- Pisces. Рыбы. Ихтиология -- Физиология, биофизика и биохимия -- Физиология*

Пример 6

На входе — *Р410.150.11* . Найден и «расшифрован» полностью.
Цепочка: *Здравоохранение. Медицинские науки -- Внутренние болезни -- Болезни систем кровообращения и лимфообращения -- Болезни сердца -- Болезни перикарда -- Перикардиты*

Пример 7

На входе — *Д443.426.3у(2)8.*
Найден: *Д443.426.3*
Цепочка: *Науки о Земле (геодезические, геофизические, геологические и географические науки) -- Геологические науки -- Геология -- Геологическая разведка -- Методика и техника поисков и разведки -- Геофизические методы поисков и разведки -- Буровая (промысловая) геофизика. Каротаж -- Радиоактивный каротаж -- Нейтронный гамма-каротаж (НГК)*

Пример 8

На входе — *В661.41я26.*
Найден: *В661.41*
Цепочка: *Физико-математические науки -- Астрономия -- Звезды и диффузная материя -- Звездные характеристики -- Спектры звезд -- Спектральная классификация звезд*

Пример 9

На входе — *Е693.363.99.* Найден и «расшифрован» полностью.
Цепочка: *Биологические науки -- Зоология -- Систематика животных. Специальные зоологические науки Chordata. Хордовые -- Vertebrata. Позвоночные. Зоология позвоночных -- Mammalia. Млекопитающие. Териология (Маммалиология) -- Cetacea. Китообразные. Киты, дельфины, кашалоты -- Дельфины*

Из приведенных примеров видно, до какой степени продуктивным для поиска в ЭК может оказаться восстановление словесных формулировок для индексов в библиографических записях, полученных в результате ретроконверсии алфавитного карточного каталога. Разумеется, не обойдется без сложностей, связанных, например, с изменением структуры систематического каталога и правил систематизации за прошедшие десятилетия, но при наличии профессиональных экспертов найдутся приемлемые решения.

Дополнительно приведем примеры (10 и 11) использования средств тематического поиска в электронной коллекции работ уникальных произведений великих русских ученых XVIII—XIX вв. (грант РФФИ 11-07-00750). При включении в их библиографические записи цепочек словесных формулировок индексов ББК *издания XVII—XIX вв. включаются в общий поиск по темам* в Национальной электронной библиотеке и, кроме того, в едином электронном каталоге (ЭК) РГБ вместе с публикациями на традиционных носителях.

Пример 10

Мережковский, Константин Сергеевич (1855—1921).
Исследования о губках Белого моря / К.С. Мережковский. — Санкт-Петербург : Тип. В.Ф. Демакова, 1879.
Приписано 2 индекса ББК.
Цепочки их словесных формулировок:
1) Биологические науки -- Зоология -- Систематика животных -- Беспозвоночные -- Губки
2) Биологические науки -- Гидробиология -- Региональная гидробиология морей и океанов -- Белое море

Пример 11

Анучин, Дмитрий Николаевич (1843—1923). К истории искусства и верований у приуральской чуди : чудския изображения летящих птиц и мифических крылатых существ : с 3-мя таблицами (фототипиями) и 129-ю рисунками в тексте / Д.Н. Анучин. — Москва : Тип. М.Г. Волчанинова, 1899. — [2], 87—160 с., [3] л. ил. : ил.; 33 см. Из Материалов по археологии восточных губерний, изд. Императорским Моск. археологическим обществом, т. 3
Документу приписано 2 индекса ББК.
Цепочки их словесных формулировок (расшифровка индексов):
1) Искусство. Искусствознание -- Скульптура -- Виды скульптуры -- Станковая скульптура -- Мелкая пластика -- История мелкой пластики -- Российская Федерация -- Мелкая пластика Урала и Приуралья -- Темы и образы -- Иллюстративные издания
2) Этнография -- Историческая этнография -- Россия -- Историческая этнография Европейской части -- Финно-угорские народы и племена -- Духовная культура -- Религия. Верования
Дополнительные ключевые слова: *птицы, мифические крылатые существа, приуральская чудь*

Классификация как метод обогащения

Итак, цепочка формулировок индексов, включенная непосредственно в состав электронного ресурса, автоматически прокладывает по соответствующей ветви иерархического дерева классификации путь к документу от верхнего уровня до наиболее узкого понятия. В результате работа может быть найдена при поиске с любого уровня иерархии, заданного в запросе, по умолчанию, без участия пользователя. Цепочки могут работать как иерархический путь к документу и в библиографической записи ЭК, и во встроенных метаданных полнотекстовой базы данных, и в иных формах представления документов.

Добавление поисковых признаков путем дополнительной интеллектуальной обработки полнотекстовых ресурсов принято называть в современных публикациях *обогащением* (enrichment) полнотекстовых ресурсов. Оно формирует дополнительные возможности их обнаружения по наибольшему количеству запросов. Предоставление более богатых, привычных человеческой логике и соответствующих принятой структуре наук средств поиска информационных ресурсов в электронных библиотеках обеспечивает, с одной стороны, полноту поиска и, с другой стороны, существенно повышает востребованность предоставляемых электронных ресурсов, создание которых так дорого обходится библиотекам.

Классификация в семантической паутине

В последние годы многие библиотекари выражают опасения, что библиотеки со своими способами описания тематики текстов могут не вписаться в зарождающуюся *технологию «Semantic Web»* [13] и, возможно, потеряются в сетевом пространстве. Далее покажем, что предлагаемые технологии тематического поиска на основе классификации и словесных формулировок индексов удачно вписываются в современную тенденцию представления библиотечных ресурсов на основе концепции «Semantic Web» (семантическая паутина, в переводе с английского, или «семантический веб» — в качестве некой комбинированной формы термина, употребляемой достаточно часто).

Термин «семантическая паутина» был впервые введен сэром Тимом Бернерсом-Ли (изобретателем Всемирной паутины WWW) в 2001 году. Она названа им «следующим шагом в развитии Всемирной паутины». Позже он предложил в качестве синонима термин Гигантский Глобальный Граф (“Giant Global Graph”), но согласился, что первый термин уже завоевал свои позиции [2]. В обычной структуре «всемирной паутины» WWW, основанной на HTML-страницах, информация заложена в тексте страниц и извлекается человеком с помощью какого-либо браузера. Семантическая паутина как способ представления знаний предполагает запись информации в виде семантической сети с помощью онтологий [10].

«Семанти́ческая сеть — информационная модель предметной области, *имеющая вид ориентированного графа,* вершины которого соответствуют объектам предметной области, а дуги (ребра) задают отношения между ними. Объектами могут быть понятия, события, свойства, процессы» [10]. Поскольку в математическом смысле и семантическая сеть, и семантическая паутина являются графами, без использования терминов теории графов [12] не обойтись, и имеет смысл дать некоторые пояснения для читателей, далеких от «царицы наук». В термине «семантическая сеть» соединены термины из двух наук: «семантика в языкознании изучает смысл единиц языка, а сеть в математике представляет собой разновидность графа — набора вершин, соединенных дугами (ребрами). В семантической сети роль вершин выполняют понятия базы знаний, а дуги (причем направленные) задают отношения между ними. Таким образом, семантическая сеть отражает семантику предметной области в виде понятий и отношений» [10].

В результате основной формой представления семантической сети является граф. Понятия семантической сети записываются в овалах или прямоугольниках и соединяются стрелками с подписями — дугами. Такой способ изображения

различных технологий, путей и других структур знаком практически каждому человеку. Графическому изображению сети, как любому графу, соответствует строгая математическая запись, и компьютерные программы по ним и реализуют свои вычисления в процессе выполнения своих функций.

Иерархические деревья — это разновидность графов. Каждая цепочка словесных формулировок индексов ББК — фрагмент направленного графа (дерева), где узлы (вершины) — это словесные формулировки, связанные дугами (ребрами, стрелками) — отношениями. «Иерархия типов и подтипов является стандартной характеристикой семантических сетей. Иерархия может включать длинную цепочку сущностей» [10].

В качестве иллюстрации построения семантических сетей на основе цепочек словесных формулировок индексов ББК рассмотрим представление в виде графа цепочек для пяти диссертаций из электронной библиотеки РГБ.

Пример 12. Заглавия диссертаций с цепочками, на основании которых построена семантическая сеть на рис. 2.

1) Горномарийские полисемантичные глаголы и их русские эквиваленты в горномарийско-русских словарях (3 индекса)
Филологические науки. Художественная литература -- Языкознание -- Финно-угорские языки -- Волжская группа языков -- Марийские (мари, черемисский) языки -- Лексикология -- Семантика (семасиология)
Филологические науки. Художественная литература -- Языкознание -- Финно-угорские языки --Волжская группа языков -- Марийские (мари, черемисский) языки -- Лексикология -- Семантика (семасиология)
Филологические науки. Художественная литература -- Языкознание -- Финно-угорские языки -- Волжская группа языков -- Марийские (мари, черемисский) языки -- Лингвистическая стилистика. Перевод -- Перевод
Филологические науки. Художественная литература -- Языкознание -- Индоевропейские языки -- Славянские языки -- Восточнославянские языки -- Русский язык -- Лингвистическая стилистика. Перевод -- Перевод

2) Прагматические аспекты функционирования слова в эрзянском языке
Филологические науки. Художественная литература -- Языкознание -- Финно-угорские языки -- Волжская группа языков -- Мордовские языки -- Мордовско-эрзянский (эрзя-мордовский) язык -- Лексикология

3) Категория градуальности в современных мордовских языках: на материале имени прилагательного
Филологические науки. Художественная литература -- Языкознание -- Финно-угорские языки -- Волжская группа языков -- Мордовские языки -- Грамматика -- Морфология -- Части речи -- Имя прилагательное

4) Ихтионимы и лексика рыболовства в марийском языке
Филологические науки. Художественная литература -- Языкознание -- Финно-угорские языки -- Волжская группа языков -- Марийские (мари, черемисский) языки -- Лексикология -- Термин и терминология

5) Заглавие и цепочка из примера 1.

Таким образом, наши цепочки словесных формулировок индексов ББК, описывающие темы документов, можно интерпретировать как естественные составляющие семантических сетей. Классификационные семантические сети состоят из маршрутов, проложенных для поиска пользователем данных в электронных библиотеках.

Список источников

1. *Авдеева Н.В.* Система «Антиплагиат.РГБ»: задачи, проблемы, результаты, перспективы / Н.В. Авдеева [и др.] // Материалы Междунар. конф. «Интеллектуализация обработки информации» (ИОИ-9) (сент. 2012). — М. : Торус Пресс, 2012. — С. 593—596.
2. *Бернерс-Ли Тим.* Гигантский Глобальный Граф «Giant Global Graph» [Электронный ресурс] / Тим Бернерс-Ли. — Режим доступа: http://goodarticles.narod. ru/ggg.html
3. Информационная система T-Libra [Электронный ресурс]. — Режим доступа: http://www.tlibra.ru/tlibra
4. *Лаврёнова О.А.* Каким способом можно «расшифровать» классификационные индексы в библиографических записях для обеспечения тематического поиска в электронных ресурсах библиотеки [Электронный ресурс] / О.А. Лаврёнова. — Режим доступа: http://www.aselibrary.ru/blogs/archives/1128/
5. *Она же.* Тематический поиск в электронных каталогах и электронных библиотеках // Библиотековедение. — 2004. — № 5. — С. 42—50.
6. *Она же.* Традиционные средства библиотечного поиска в электронной среде. — Доступность электронных ресурсов библиотек, музеев, архивов как актуальная проблема развития информационного общества : материалы VII Рос. науч.-практ. конф. «Электронные ресурсы библиотек, музеев, архивов» (31 окт. — 2 нояб. 2011 г., Санкт-Петербург). — СПб. : Политехника-сервис, 2011. — С. 179—191.
7. Международная конференция по компьютерной лингвистике «Диалог» [Электронный ресурс]. — Режим доступа: http://www.dialog-21.ru/digest/2012/

Рис. 2. Семантическая сеть на основе словесных формулировок индексов ББК для нескольких документов из ЭБ диссертаций

8. Международная конференция по теории и практике электронных библиотек TPDL [Электронный ресурс]. — Режим доступа: http://tpdl2012.org/
9. Российская конференция по электронным библиотекам RCDL. Труды [Электронный ресурс]. — Режим доступа: http://rcdl2012.pereslavl.ru/section.php?id=79
10. Семантическая_сеть [Электронный ресурс]. — Режим доступа: http://ru.cybernetics.wikia.com/wiki/Семантическая_сеть
11. Таблицы ББК по технике и естественным наукам в системе АЛЕФ [Электронный ресурс]. — Режим доступа: http://aleph.rsl.ru/F/?func=file&file_name=base-list-tst01
12. *Харари Ф.* Теория графов / Ф. Харари. — М. : УРСС, 2003. — 300 с.
13. *Fensel Dieter.* Spinning the Semantic Web: Bringing the World Wide Web to Its Full Potential / Dieter Fensel [et al.]. — The MIT Press, 2002.
14. *Lavrenova Olga.* The National Library Bibliographic Classification (BBK) as a Base for Subject Search in the Integrated RSL Digital Library. The Project presentation [Электронный ресурс] / Olga Lavrenova. — Режим доступа: https://www.comp.glam.ac.uk/pages/research/hypermedia/nkos/nkos2012/abstracts/Lavrenova.pdf
15. Networked Knowledge Organization Systems and Services. The 11th European Networked Knowledge Organization Systems (NKOS) Workshop [Электронный ресурс]. — Режим доступа: https://www.comp.glam.ac.uk/pages/research/hypermedia/nkos/nkos2012/programme2012.html
16. Project ISO 25964-1 [Электронный ресурс]. — Режим доступа: http://www.niso.org/schemas/iso25964