Тематический поиск в электронных библиотеках: иллюзии и реальность

Subject Search in Electronic Libraries: Illusions and Reality

Тематичний пошук в електронних бібліотеках: ілюзії та реальність

О. А. Лавренова
Российская государственная библиотека, Россия, Москва

Olga Lavrenova
Russian State Library, Moscow, Russia

О. О. Лаврьонова
Російська державна бібліотека, Москва, Росія

Предлагается рассмотреть распространенные иллюзии относительно использования тематического (предметного) поиска в электронных библиотеках. Дается анализ средств тематического поиска в электронных каталогах и полнотекстовых электронных ресурсах с точки зрения лингвистической теории языкового общения, с одной стороны, и поисковых намерений пользователей электронных библиотек, с другой стороны.

The author examines the wide-spread illusions related to the topical (subject) search in e-libraries. She analyzes the tools of information retrieval in electronic catalogs, as well as in full-text electronic resources from the viewpoint of the linguistic theory of language communication, on one hand, and search intentions of e-library users, on the other hand.

Автор пропонує розглянути поширені ілюзії щодо використання тематичного (предметного) пошуку в електронних бібліотеках. Подано аналіз засобів тематичного пошуку в електронних каталогах та повнотекстових електронних ресурсах із погляду лінгвістичної теорії мовного спілкування, з одного боку, та пошукових намірів користувачів електронних бібліотек, з іншого боку.

1. Разработчиков электронных библиотек (ЭБ) вдохновляют огромные вычислительные мощности современных компьютерных систем и богатые **полнотекстовые массивы полнотекстовых электронных документов,** вызывая массовые иллюзии по поводу теоретически неограниченных возможностей поиска информации, в частности, по требуемым темам и предметам. Иллюзия вырастает до таких приятных сердцу разработчиков идей, как отказ от каталогов и громоздких MARC-форматов, неприлично трудоемких сооружений типа тезаурусов и библиотечных классификаций.

Действительно, такие поисковые системы в сети Интернет, как Yandex, Rambler, Alta Vista и даже Google работают в таком относительно вольном стиле и позволяют себе из всех лингвистических средств АИС обходиться грамматическими парадигмами, характеристиками частотности и сочетаемости поисковых элементов и т. п. На полноценные семантические средства потребовались бы непомерные финансовые затраты. Соответственно, пользователь этих поисковых систем и не ждет от них особых тонкостей, так как не питает иллюзий относительно полноты или точности поиска информации в них, и готов получить хотя бы один нужный сайт и 1000 ненужных.

О пользователях электронных библиотек этого не скажешь. Обращаясь в ЭБ, человек имеет совершенно другие поисковые намерения: он пребывает в иллюзии, что ему выдадут все то, что есть полезного по его запросу в электронном «фонде» данной библиотеки. Он интуитивно предполагает некую аккуратную организацию массива данных ЭБ и поиска в нем. Может, сказывается влияние традиционных библиотек? Задача тематического поиска в ЭБ, в самом деле, отличается от таковой в сфере создания электронных каталогов (ЭК) традиционных библиотечных фондов не столь принципиально, как представляется многим, тем более что поиск в ЭБ преимущественно основан на тех же библиографических записях (БЗ) в ЭК или соответствующих им по содержанию метаданных, структурированных иными способами (например, с помощью языка разметки текстов XML) [4, 5].

Добиться высокого качества поисковых результатов в ЭБ можно лишь с учетом двух важнейших факторов: законов речевого (языкового) общения и типов поисковых намерений пользователей электронных библиотек.

2. **Естественное объяснение** того, что свободный поиск по КС или по словам в полных электронных текстах документов не может обеспечить высокие показатели полноты, заключается **в следующих непреложных законах языкового общения**:

- языковые способы передачи знаний (информации) обеспечивают экономию используемых средств; ограничение передаваемой каждым сообщением (текстом) информации позволяет людям сообщать друг другу сведения в приемлемые отрезки времени; передача информации невозможна без восстановления получателем (адресатом) информации, не отраженной непосредственно в тексте, на основе фонда общих знаний;
- отсутствие необходимых фрагментов фонда общих знаний у адресата обусловливает непонимание им сообщения (текста).

Конкретный текст, идентифицированный в качестве единицы в речевом материале, **представляет собой языковую модель некоторого фрагмента системы знаний,** отраженного в семантике данного текста. Смысловое содержание текста не может быть понято (интерпретировано адекватно заложенному в нем смыслу) без ориентации соответствующего фрагмента системы знаний в системе знаний соответствующей области науки, техники и т. д. Обращение к знаниям экстралингвистического характера считается одной из важнейших задач лингвистической семантики (см., например, работы [1, 2]). В той ситуации общения, когда между людьми как коммуникантами функционирует некая автоматизированная информационная система (АИС), например, электронная библиотека, условия понимания только усложняются. Между системой знаний автора и системой знаний адресата (читателя, пользователя ЭБ) существенную интерпретирующую роль играют структуры «знаний о мире», заложенные в систему.

Спрашивается, почему разработчики систем нередко полагают, что некая совокупность средств вычислительной техники и программного продукта без обеспечения системы фрагментами «знаний о мире» сможет адекватно обнаруживать необходимые человеку данные? Особенно распространена **иллюзия относительно того, что поиск в полнотекстовых базах данных электронных библиотек сам по себе достаточен**, так как все слова в нем могут быть поисковыми.

Допустим, программный продукт отыскивает в электронных текстах указанные в запросе последовательности знаков и, соответственно, выдает те документы, в которых они найдены.

Рассмотрим возникающие при этом **проблемы:**

1) Текст составлен из словоформ, которые используются в основном не в исходной форме. Поэтому нужен машинный грамматический словарь, в котором содержатся парадигмы склонения и спряжения слов и т. п., т. е. требуется лингвистический процессор. Словарь должен включать как общеязыковую лексику, так и терминологию соответствующей области знания. Что это означает для универсальной библиотеки? Словарь должен содержать хотя бы основную терминологию практически всех областей знаний одновременно. Насколько это реально выполнить? Остается использовать поисковые признаки с отсечением окончаний и частично суффиксов, как это делается в электронных каталогах (например, *физи* – физика, физический, в физике*).

2) Лексические единицы обладают свойствами синонимии, омонимии и полисемии. Следовательно, в электронной библиотеке нужны специальные словари (переходим к семантике). Например, если в системе не зафиксирована синонимия слов и словосочетаний *«погода», «метеообстановка», «метеоусловия», «метеоявления», «погодные условия», «синоптические условия»*, то, естественно, при использовании в тексте только одного из них, электронный документ не будет найден по запросам, в которых используются другие слова с эквивалентным значением.

3) Автор экономит силы и использует только слова и словосочетания, необходимые для понимания адресатами с соответствующим фондом общих с ним знаний. Никто не может поручиться, что в своем тексте автор использует все слова и словосочетания, обозначающие более широкие понятия по отношению к тем, которые непосредственно требуются для изложения смыслового содержания работы. Автор далеко не всегда расписывает в тексте иерархическую структуру области знания, полагая, что она знакома будущим читателям (они должны знать, что *кислородная система и система регулирования давления воздуха* относятся к *системам жизнеобеспечения*,

карельский или марийский язык – к *финно-угорским языкам*, а *экология микроорганизмов* – к *микробиологии*). Таким образом, автор предполагает наличие некоторого общего с читателем фонда знаний, необходимого и достаточного для понимания, т. е. адекватной интерпретации смысла текста. Можем ли мы как специалисты, создающие электронные каталоги и электронные библиотеки, полагаться на случай (авось, автор непосредственно в своем тексте обеспечит иерархический поиск в нашей системе)? Логичнее создать в ней «общий фонд знаний о мире», т. е. сформировать иерархическую классификацию, информационно-поисковый тезаурус или, как теперь стало популярным в сфере электронных библиотек, построить онтологию.

4) **Ассоциативные связи между понятиями – важное свойство мышления человека.** Ассоциативные связи, отображаемые в тексте, целиком зависят от способа мышления автора. Требуется сформировать в системе ассоциативные связи, общие для некоторого круга специалистов *(*например, *безопасность полетов* – ассоциативные понятия*: герметичность, обледенение, гроза, полоса безопасности, противообледенительное оборудование, опасное сближение, прочность, столкновение и т. д.).*

Следовательно, полнота полнотекстового поиска при всей его первозданной красоте также зависит от использования сильных лингвистических средств: грамматических словарей, тезаурусов, классификаций.

3. Поиск в электронных библиотеках (ЭБ) реализуется:

- по обычным электронным каталогам ЭБ с использованием принятых в них средств;
- по метаданным, представленным не в форме записей ЭК (XML, HTML – разметка);
- по полным текстам документов с лингвистическими процессорами и без них;
- по частям полных текстов (например, оглавлениям, рефератам, аннотациям, наиболее информативным разделам).

Разумеется, неплохие результаты дают статистические методы обработки текстов, использование оценки значимости терминов в тексте путем ранжирования их в соответствии с некоторыми показателями «веса». Однако в больших массивах полнотекстовых данных проверить качество этих средств чрезвычайно трудно, приходится полагаться на некоторые вероятностные оценки результатов поиска.

Вопрос о том, **какой информационно-поисковый язык (ИПЯ) лучше для организации тематического поиска, не имеет смысла**. Напротив, для обеспечения всего разнообразия поисковых намерений пользователей необходимо сочетание этих средств: КС, предметные рубрики обеспечивают высокую точность поиска, классификации и тезаурусы – его полноту [3-6].

Наиболее распространенные **ИПЯ предоставляют**, как известно, **следующие средства тематического поиска в электронных каталогах**: свободные ключевые слова (КС), перечни КС, тезаурусы, предметные рубрики (предметные классификации), иерархические классификации (индексы, наименования делений), рубрикаторы (коды, наименования рубрик – для приблизительного распределения документов по широким темам). Для некой имитации тематического поиска используется поиск по всем элементам библиографических записей или по полным текстам и их частям.

4. **Выбор средств тематического поиска должен определяться типами информационных потребностей пользователей,** для которых создается информационная система, точнее – типами их поисковых намерений.

Можно выделить следующие **типы поисковых намерений пользователей** ЭК и ЭБ при поиске по теме [5]:«Найти хоть что-то, но в точном соответствии с предметом поиска» (достаточно свободных КС и предметных рубрик);

- «Подобрать максимально полную информацию по теме, но желательно без случайных документов» (полезны классификации и тезаурусы);
- «Найти все, что касается темы поиска, допустима выдача лишних документов» (нужны тезаурусы и классификации в сочетании со свободными КС);
- «Хорошо бы сначала определиться, какие бывают темы в рамках интересующей области знания, разобраться со связями между темами» (необходима иерархическая классификация).

Сравнение возможностей разных средств тематического поиска в различных сетевых электронных каталогах, включая каталоги электронных библиотек, представляет собой достаточно увлекательное и доступное занятие. Поскольку фонды библиотек по своему составу

существенно различаются, простым поиском не обойдешься. Рекомендуем найти в одном из ЭК записи по интересному запросу, а затем отыскать по авторам и заглавиям те же документы в каталогах других библиотек, проанализировать поля, по которым реализуется в них тематический поиск – и сделать вывод, можно ли их, в принципе, найти по заданной теме, используя предусмотренные в БЗ поля.

Эксперименты, в частности, показали [5], что наиболее продуктивным средством поиска на полноту можно считать, кроме тезаурусов, **иерархические цепочки словесных формулировок индексов ББК,** которые строятся для документов систематизаторами РГБ на основе таблиц-эталонов. Цепочки, в сущности, расшифровывают все уровни иерархии соответствующей ветви иерархического дерева ББК, используемые при построении индекса. Наличие в БЗ таких цепочек обеспечивает тематический поиск по содержащимся в них КС, что означает автоматический учет иерархических связей между темами при поиске. Дополнительно используются свободные КС.

Хотелось бы подчеркнуть, что использование в БЗ электронных каталогов индексов классификаций без машиночитаемых таблиц и предметного доступа, без расшифровки индексов выглядит как чистая дань библиотечным традициям и карточным каталогам. Кто-нибудь видел пользователя, который знает индексы наизусть? Разве что систематизаторы. Индексы, построенные с большими затратами труда, не играют в сетях никакой роли. Интересно, какие иллюзии можно питать относительно пользы включения индексов в БЗ в таких условиях, разве что желание печатать правильные каталожные карточки?

Исследование показало, что **без ввода в БЗ словесных формулировок индексов и без тезаурусов** при поиске в электронных каталогах получаются **огромные потери** информации, намного большие, чем можно себе представить априори, и при этом скрытые от пользователя.

В качестве примера рассмотрим результаты тематического поиска в базе данных авторефератов ЭК РГБ (АЛЕФ) [5], имея в виду, что массив авторефератов включается в электронную библиотеку РГБ.

Запрос 1: *«*Авторефераты по финно-угорским языкам». Поисковые признаки: финно угорск* язык*. Найдено 109 БЗ. 33 найденных автореферата обработаны в те годы, когда для индексов ББК формировались цепочки. Из них без наличия цепочек могло быть найдено только 2 (по сочетанию наименования специальности и заглавия). Все остальные 31 БЗ выданы только по словам, входящим в словесные формулировки индексов, так как в них зафиксированы иерархические связи названий конкретных языков со словосочетанием «финно-угорские языки». 76 авторефератов прежних лет найдены почти исключительно по наименованию специальности «10. 02. 07 Финно-угорские и самодийские языки», которую теперь, похоже, отменили. При этом в 33-х БЗ наиболее новых авторефератов встретились, в частности, следующие специальности: 05. 13. 01 Системный анализ, управление и обработка информации (по отраслям), 10. 02. 01 Русский язык, 10. 02. 19 Теория языка, 10. 02. 20 Сравнительно-историческое, типологическое и сопоставительное языкознание, 13. 00. 08 Теория и методика профессионального образования, 24. 00. 01 Теория и история культуры. Это, в частности, демонстрирует абсолютную непродуктивность тематического поиска в ЭБ по специальностям ВАК.

Ниже приведены примеры цепочек словесных формулировок индексов в библиографических записях из ЭК, благодаря структуре которых были найдены авторефераты по указанному запросу.

(а) Заглавие: *Иноязычная лексика в мордовских **языках** (индоевропейские* заимствования)

Тема: *Филологические науки. Художественная литература -- **Языкознание** -- Финно-угорские **языки** -- Волжская группа **языков** -- Мордовские **языки** -- Лексикология -- Словарный состав **языка** -- Заимствования. Интернационализм*

(б) Заглавие: *Вокалическая система кильдинского диалекта саамского **языка** в свете русско-саамской интерференции:: автореферат дис.... кандидата филологических наук: 10. 02. 19 / С. - Петерб. гос. ун-т*

Тема: *Филологические науки. Художественная литература -- **Языкознание** -- **Финно-угорские языки** -- Саамская (лопарская) группа **языков** -- Саамский (лопарский) **язык** -- Диалектология и диалектография -- Местные (территориальные) диалекты*

(в) Заглавие: *Зоонимическая лексика карельского **языка**:: автореферат дис.... кандидата филологических наук: 10. 02. 22 / Петрозавод. гос. ун-т*

Тема: *Филологические науки. Художественная литература -- **Языкознание** -- **Финно-угорские языки** -- Прибалтийско-финская группа **языков** -- Карельский **язык** -- Лексикология -- Термин и терминология*

(г) Заглавие: *Реконструкция праобско-угорского вокализма:: автореферат дис.... кандидата филологических наук: 10. 02. 20 / Рос. гос. гуманитар. ун-т (РГГУ)*

Тема: *Филологические науки. Художественная литература -- **Языкознание** -- **Финно-угорские языки** -- **Угорская** группа **языков** -- Обско-угорские **языки** -- Фонетика*

(д) Заглавие: *Топонимия бассейна реки Казым:: автореферат дис.... доктора филологических наук: 10. 02. 02 / Удмурт. гос. ун-т*

Тема: *Филологические науки. Художественная литература -- **Языкознание** -- **Финно-угорские языки** -- **Угорская** группа **языков** -- Обско-угорские **языки** -- Хантыйский (остяцкий) **язык** -- Лексикология -- Словарный состав **языка***

(е) Заглавие: *Модальные слова и словосочетания в современном марийском **языке**: автореферат дис.... кандидата филологических наук: 10. 02. 22 / Марийс. гос. ун-т*

Темы: *Филологические науки. Художественная литература -- **Языкознание** -- **Финно-угорские языки** -- Волжская группа **языков** -- Марийские (мари, черемисский) **язык** -- Грамматика -- Синтаксис -- Словосочетание Филологические науки. Художественная литература -- **Языкознание** -- **Финно-угорские языки** -- Волжская группа **языков** -- Марийские (мари, черемисский) **язык** -- Грамматика -- Морфология -- Части речи -- Модальные слова*

По поводу величины потерь при поиске авторефератов с 1986 по 2002 г. (без цепочек словесных формулировок индексов ББК) по упомянутому запросу пользователь останется в неведении. Дело в том, что и поиск по заглавиям не даст положительного результата: они содержат, как правило, названия конкретных языков или подгрупп: финский, карельский, коми-пермяцкий, удмуртский, марийский, хантыйский, венгерский, мордовские (мокшанский и эрзянский) языки и т. д. Парадоксально требовать от пользователя проведения поисков сразу по всем конкретным финно-угорским языкам в заглавиях. Приведенный пример показывает, что при поиске по полному тексту без таблиц классификации пользователь должен уповать на автора, который должен догадаться, описывая в автореферате работу по конкретному языку упомянуть в тексте всю иерархию терминов из соответствующей ветви дерева классификации языков.

Реальность такова, что, игнорируя законы речевого общения и особенности поисковых намерений пользователей, электронная библиотека (и ЭК обычной библиотеки тоже) формирует у людей **опасную иллюзию обеспечения полной информацией** о наличии в базе данных ЭБ или в фонде только тех документов, которые выданы при поиске, хотя для этого в системе не обеспечены никакие условия.

Понятно, что при создании электронных библиотек недопустимо подобное положение вещей. На сайте каждой ЭБ необходимо четко указывать, что может пользователь ожидать при использовании тех или иных поисковых средств. Кроме того, **сообщество организаций, создающих электронные библиотеки, должно сосредоточиться на проблемах семантического представления содержащихся в них текстов, развития способов оптимального использования классификаций, разработки тезаурусов и найти для этого источники финансирования.**

Литература

1. Гак В. Г. К проблеме синтаксической семантики (семантическая интерпретация «глубинных» и «поверхностных» структур) //Инвариантные синтаксические значения и структура предложения. –М., 1969. – С. 77–85
2. Звегинцев В. А. Язык и лингвистическая теория. – М.: 1973. – 248 с. Лавренова О. А. Методика разработки информационно-поискового тезауруса. – М.: Пашков дом, 2001. – 54 с.
4. Лавренова О. А. Моделирование семантики научно-технических текстов для АИС и его теоретические основы // Электронные библиотеки: перспективные методы и технологии, электронные коллекции. Труды Восьмой Всероссийской научной конференции (RCDL'2006). Суздаль, 17-19 октября 2006 г. – Ярославль: Ярославский государственный университет им. П. Г. Демидова, 2006. – С. 345–352
5. Лавренова О. А. Средства тематического поиска в информационных ресурсах библиотек // Корпоративные библиотечные системы: технологии и инновации. Труды V научно-практической конференции АРБИКОН. Санкт-Петербург (Россия), Лаппеенранта (Финляндия), 1-7 июля 2007 г. – Санкт-Петербург, 2007. – С. 40–45

6. Лавренова О. А. Тематический поиск в электронных каталогах и электронных библиотеках // Библиотековедение. – 2004. – №5, – с. 42–50