СПОСОБЫ ОПИСАНИЯ ИНФОРМАЦИОННЫХ ПОТРЕБНОСТЕЙ ПОЛЬЗОВАТЕЛЕЙ ПРИ ВЗАИМОДЕЙСТВИИ С ЭЛЕКТРОННЫМИ БИБЛИОТЕКАМИ ЧЕРЕЗ ПОСРЕДНИК С ТЕЗАУРУСОМ И РУБРИКАТОРОМ

Казаков Евгений Николаевич

Всероссийский научно – технический информационный центр (ВНТИЦ), 125801, Москва, ГСП-47, ул. Смольная, 14, postmaster@vntic.org.ru

Аннотация

Рассматривается взаимодействие пользователей с цифровыми библиотеками, содержащими многочисленные неоднородные коллекции. Предполагается, что взаимодействие реализуется с помощью специальной инфраструктуры: посредника, имеющего три уровня представления метаинформации: персонализованный, интероперабельный (федеративный) и локальный. Предполагается, что посредник содержит тезаурус и предметный классификатор (рубрикатор), которые обеспечивают интеллектуальную помощь пользователям.

Предлагаются способы фиксации исходных (начальных) информационных требований пользователей. Посредник учитывает эти требования при обработке запросов. Пользователи могут корректировать эти требования в диалоговом режиме.

Пользователь заполняет анкетную форму, состоящую из следующих блоков: общие указания, текст запроса, пояснение к запросу, использование тезауруса и предметного классификатора, критерий релевантности.

Описывается каждый блок и его влияние на результат поиска. Предлагаются команды навигации по тезаурусу.

Работа выполнена при поддержке грантов РФФИ №01-07-90084, №00-07-90086.

1. Введение

Электронные библиотеки (ЭБ) представляют собой электронную информационную среду, которая должна обеспечивать комфортный доступ ко всем видам накопленной человечеством информации (текстовой, аудио, видео, мультимедиа и другой) в любое время в любом месте каждому пользователю, пожелавшему получить нужную ему порцию информации.

Комфортный доступ предполагает учет самых разнообразных пожеланий и возможностей пользователей, обусловленных как различием их опыта взаимодействия с ЭБ, так и большим количеством и разнообразием коллекций.

Одна из наиболее важных задач – решение проблемы масштабирования, то есть ограничение сложности сценария обращения при сколь угодно большом количестве ЭБ и электронных коллекций с учетом разнообразия их структур и принципов функционирования. Эффективный подход к решению данной задачи разрабатывается группой сотрудников Института проблем информатики Российской Академии наук (ИПИ РАН) под руководством Л. А. Калиниченко.

Основная идея этого подхода по созданию интегрированных ЭБ с неоднородными ресурсами заключается в создании инфраструктуры доступа, представляющей промежуточный слой между электронными коллекциями и потребителями информации. Основными компонентами промежуточного слоя являются информационные посредники, существующие независимо от электронных коллекций. В посреднике различаются три уровня представления информации и метаинформации локальный уровень,

представляющий метаинформацию о разнородных коллекциях, федеративный (интероперабельный) уровень, содержащий интегрированную метаинформацию, и персонализованный уровень с метаинформацией для конкретных пользователей. Подробную информацию о структуре посредника можно найти в [1].

Еще одной важной проблемой комфортного доступа к ЭБ является обеспечение интеллектуальной помощи пользователю при взаимодействии с ЭБ. Решение этой проблемы связано с введением в состав посредника политематического тезауруса и предметного классификатора (рубрикатора) [2, 3], а также онтологических определений. Интеллектуальная помощь пользователю в зависимости от его пожеланий может быть оказана неявно (в режиме умолчания), либо в режиме подсказки в процессе диалога.

В статье рассматривается начальная фаза взаимодействия с ЭБ, в которой пользователь должен достаточно точно и, по возможности полно, сформулировать свои исходные информационные потребности. От того, насколько точно и полно пользователь опишет свои требования к поиску, зависит качество ответа.

В статье предполагается, что электронные коллекции могут содержать различные виды информации (текст, аудио, видео, мультимедиа), но должны иметь для каждого документа, произведения или самостоятельного фрагмента явное текстовое описание наиболее значимых для пользователя характеристик (вид информации, тип носителя, стандарт записи, авторы, названия, год и место выпуска, исполнители, действующие лица, объем документа или длительность записи и т.д.).

2. Общая схема взаимодействия с ЭБ

Взаимодействие пользователя с ЭБ через посредник с тезаурусом и рубрикатором рассматривается как интерактивный процесс диалогового общения. Предполагается, что на основе исходных информационных потребностей, заданных пользователем, посредник генерирует с применением хорошо формализованного и структурированного языка (например, SQL [4]), а также с использованием тезауруса и рубрикатора серию запросов к тем коллекциям, в которых потенциально имеется нужная пользователю информация. После получения ответов от всех коллекций через локальный уровень и интеграции ответов посредник передает их пользователю на персонализованный уровень.

Пользователь, получив ответ и оценив его качество, может либо закончить диалог, либо скорректировать исходные информационные потребности, либо инициировать диалог с тезаурусом, рубрикатором и онтологиями. Пользователь может прервать этот интерактивный процесс либо, получив удовлетворительный ответ, адекватный своим потребностям, либо, поняв, что все усилия не дают улучшения качества результата.

Процесс взаимодействия пользователя с ЭБ существенно зависит от его желания и возможности влиять на диалог с посредником. Пользователь может привести самые общие сведения о своей потребности вплоть до предъявления текста или ссылки на текст, известного ему релевантного документа. Заинтересованный и искушенный пользователь может потребовать непосредственного доступа к тезаурусу и рубрикатору посредника и самостоятельно управлять процедурой расширения и трансформации запроса.

3. Структура исходных информационных потребностей пользователя.

Представление о структуре информационных потребностей пользователя можно получить, анализируя опыт работы больших широко тематических документальных и фактографических информационных систем (например: АСИНИТ (ВНТИЦ), АССИСТЕНТ (ВИНИТИ) и др.).

Естественно, что пользователь прежде всего формулирует проблему, для решения которой он осуществляет поиск. Кроме того, он определяет к каким тематическим областям может относиться

интересующая его информация. Ему также важно, какого типа поиск он хочет осуществить. В одном случае ему достаточна небольшая порция информации (например, несколько документов), очень точно соответствующая проблеме, т.е. поиск “на точность”. В других обстоятельствах (например, при подготовке обзора, патентном исследовании, выборе направления исследования) ему необходим исчерпывающий поиск, т.е. поиск “на полноту”, который приводит к выдаче очень больших порций информации (например, много документов разного объема). С этой проблемой связана глубина ретроспективы, которая для точного поиска может быть небольшой, а для исчерпывающего поиска может оказаться значительной.

Пользователь должен также решить, какие виды информации, и виды документов его интересуют, каков характер требующихся ему материалов (методические, теоретические, расчетные, обзорные документы и т.д.). Только учитывая все выше сказанное, пользователь может сформулировать текст запроса, дать описание наиболее нужных ему аспектов и тематической области поиска и оценить способы использования тезауруса и рубрикатора, по крайней мере, на первом этапе поиска. Очень важной характеристикой поиска является способ оценки найденной информации на соответствие запросу, то есть выбор критерия релевантности.

Подводя итог, можно предложить в качестве основы для фиксации исходных потребностей пользователя ЭБ некоторую анкетную электронную форму, состоящую из блоков. Каждый блок может включать несколько стандартных формулировок с перечнем допустимых значений. Например, стандартная формулировка “Виды информации” может содержать допустимые значения: текст, аудио, видео и т.д. В этой форме пользователь может отмечать нужные ему стандартные формулировки, допустимые значения, а также вписывать свои значения, которые должны рассматриваться в контексте выбранной стандартной формулировки.

Можно считать целесообразным, чтобы форма для фиксации потребностей пользователя ЭБ содержала следующие блоки:

- общие указания,
- формулировка запроса,
- пояснения к формулировке,
- указания по использованию рубрикатора,
- указания по использованию тезауруса,
- указания по оценки релевантности.

Не приходится сомневаться, что приемлемая для многих пользователей анкетная форма появится в будущем в результате длительной работы многих специалистов, а также после постепенного накопления опыта взаимодействия с ЭБ широкого круга разных пользователей при общении с большим количеством разнообразных коллекций. Не пытаясь предвосхитить такую форму, приведем в данной статье конспективное, иллюстративное описание блоков, ограничившись примерами, поясняющими основную идею подхода.

3.1. Блок общих указаний.

В этом блоке пользователь может задать основные ограничения, которым должны удовлетворять данные, получаемые через посредник от ЭБ. Ниже приводится перечень стандартных формулировок с примерным перечнем их значений. Для каждой формулировки пользователь может выбрать одно или несколько значений, а также после знака “+” дописать те значения, которые ему необходимы, причем эти значения будут рассматриваться в контексте заданных формулировок.

Таким образом блок общих указаний может иметь следующий вид:

1. Виды информации: текст, аудио, видео, мультимедиа, +_____
2. Тип документов: книги, журналы, статьи, патенты, стандарты, инструкции, отчеты, диссертации, картины, ноты, фильмы, музыкальные записи, телепередачи, радиопередачи, +_____
3. Характеристики выдаваемой информации: библиографическое описание документов (авторы, название, год и место издания, издательство, объем в страницах, +_____) описание фильмов и теле-радиопередач (название, авторы, студия, год, действующие лица, исполнители, сюжеты, длительность, + _____) описание музыкальных записей (название, авторы, исполнители, год записи, длительность звучания, + _____)
4. Объем порции: фрагменты, рефераты, аннотации, оглавления, полные документы, полные записи, + _____

5. Условия поиска: 1. поиск на точность ,
   2. поиск на полноту
6. Глубина ретроспективы: годы (1, 2, 3, 4, 5, 6, + _____)
7. Характер искомых материалов: описания произведений, произведения, обзоры, теории, методики, + _____
8. Тематическая область поиска (задается перечнем кодов рубрик предметного классификатора посредника)

3.2. Блок формулировки запроса

В этом блоке пользователь может записать любой текст, отражающий семантику его информационной потребности с той степенью детализации содержания, которую он считает достаточной. Он может описать свою потребность перечнем лексических единиц, кодов рубрик, их комбинациями, текстами документов, а также ссылками на документы и т.д.

Таким образом, блок формулировок запроса может иметь вид:

1. Связный текст: _____
2. Перечень лексических единиц (ЛЕ) слова и словосочетания в виде ключевых слов, понятий и терминов, взятых из текстов, словарей, тезаурусов и рубрикаторов: _____
3. Перечень кодов рубрик предметного классификатора посредника: _____
4. Комбинация перечня кодов рубрик и ЛЕ: _____
5. Тексты, известных пользователю документов, соответствующих его потребности: _
6. Ссылки на документы, фрагменты, адекватные его потребности: _____
7. Структурированные перечни ЛЕ и кодов рубрик (например, объединение ЛЕ или рубрик в смысловые группы с указанием цифрового идентификатора группы): ____
8. Список известных пользователю авторов или организаций, работающих по проблеме, адекватной информационной потребности пользователя: _____

Блок формулировки запроса является обязательным для заполнения, однако выбор формы заполнения определяется пользователем.

3.3. Блок пояснений к формулировке запроса

Этот блок не является обязательным. В нем целесообразно указывать наиболее важные для пользователя аспекты запроса. Это можно сделать либо в виде текста произвольного объема, либо в виде перечней ЛЕ, объединенных в смысловые группы. Например, обязательные ЛЕ, отсутствие которых в тексте документа делают его нерелевантным, желательные ЛЕ, которые могут отсутствовать даже в релевантных документах, запрещенные ЛЕ, присутствие которых в документе делает его нерелевантным.

Таким образом блок может иметь вид

1. Текст:_____
2. Обязательные ЛЕ (перечень): _____
3. Желательные ЛЕ (перечень): _____
4. Запрещенные ЛЕ (перечень): _____

Пользователь должен отдавать себе отчет, что от точности, подробности и тщательности составления формулировки запроса и пояснений к ней существенно зависит качество ответа, поскольку именно перечень ЛЕ в этих двух блоках будет определять исходную точность и полноту.

3.4. Блок указаний по использованию рубрикатора

Этот блок имеет вид:

Диалоговый доступ к рубрикатору посредника через:

1. семантический перечень рубрик
2. алфавитно-предметный указатель

Пользователь может отметить и первую и вторую позиции. Если выбрана первая позиция, то будут последовательно выводиться коды и названия рубрик первого уровня. Для выбранной пользователем рубрики первого уровня выводятся все подчиненные ей рубрики второго уровня, а для выбранной рубрики второго уровня выводятся все подчиненные ей рубрики третьего уровня и т.д.

Если отмечена вторая позиция, то пользователь может, задав усечение слова, слово или словосочетание, получить список кодов и названий рубрик, которые включают заданный пользователем лексический фрагмент. Диалоговый доступ к рубрикатору посредника позволяет пользователю выбрать нужный перечень рубрик с учетом формулировки запроса и промежуточных результатов поиска, а также использовать этот перечень для уточнения тематической области поиска или формулировки запроса.

Если пользователь не заполняет этот блок, то посредник при обработке запроса по умолчанию учитывает тот перечень рубрик, который указан в блоке общих указаний в позиции 8. Тематическая область.

3.5. Блок указаний по использованию тезауруса

Этот блок имеет вид:

Расширение запроса:

1. По рекомендации пользователя:
    - 1.1. синонимические ЛЕ
    - 1.2. узкие ЛЕ
    - 1.3. широкие ЛЕ
    - 1.4. ассоциативные ЛЕ
2. В процессе диалога пользователя с тезаурусом и лексикой посредника

Если пользователь не отметит ни одной позиции, то посредник при обработке запроса по умолчанию использует данные из блока общих указаний позиция 5. Условия поиска. Если задан поиск на точность, то посредник включит в запрос ЛЕ синонимичные ЛЕ из блока формулировки запроса и пояснений к формулировке. Если задан поиск на полноту, то посредник дополнит запрос не только синонимичными, но и более узкими и ассоциативными ЛЕ. А если объем выдачи окажется небольшим, то будут использованы и широкие ЛЕ.

Если пользователь отметит первую позицию блока, то он должен будет задать те виды связей, которые надо использовать при расширении запроса. Например, отметив позицию 1.1. синонимические ЛЕ, пользователь заставит посредник включить в запрос все ЛЕ, синонимичные лексическим единицам в блоках формулировки запроса и пояснений к формулировке.

Если пользователь отметит вторую позицию блока, то посредник включит режим диалога пользователя с тезаурусом и дополнительной лексикой посредника. В этом случае пользователь командами навигации по тезаурусу из раздела 4 сможет включить в запрос все ЛЕ тезауруса и дополнительной лексики, которые повышают качество поиска. Следует заметить, что дополнительная лексика появляется в процессе регистрации коллекций в посреднике [1] и представляет собой лексику коллекции, не совпавшую с лексикой тезауруса посредника. Дополнительная лексика доступна пользователю в процессе диалога с посредником, что позволяет использовать для поиска ЛЕ, не вошедшие в тезаурус.

3.6. Блок оценки релевантности

Указания по оценке релевантности связаны с выбором из множества критериев релевантности, предоставляемых посредником, критерия, который наилучшим образом отвечает интересам пользователя. В качестве примера можно упомянуть логические, приоритетные, весовые и прочие виды критериев. С каждым критерием связан способ упорядочивания документов в выдаче. Для логического критерия пользователь задает логическую формулу над ЛЕ запроса, используя, как правило, операции И, ИЛИ, НЕ. В этом случае в выдачу входят те документы, для которых логическая формула истинна. Для приоритетных критериев каждый смысловой группе приписывается некоторое число, определяющее ее приоритет. Документы в выдаче упорядочиваются по убыванию сумм приоритетов ЛЕ, входящих в их тексты. В случае весовых критериев каждой ЛЕ или смысловой группе ЛЕ приписывается определенный вес. Формула вычисления весового критерия обычно суммирует веса ЛЕ в документе, что позволяет обеспечить эшелонирование выдачи по убыванию весов документов.

Пользователь может выбрать критерий оценки релевантности и задать способ его вычисления для данного запроса. В процессе диалога с посредником пользователь может поменять критерий оценки релевантности и формулу его вычисления в зависимости от промежуточных результатов поиска.

4. Команды навигации по тезаурусу

Эти команды используются в процессе диалога пользователя с тезаурусом и дополнительной лексикой посредника. В статье предлагается следующий перечень команд навигации: СИМВОЛЫ, УСЕЧЕНИЯ, КОМБИНАЦИЯ, СВЯЗЬ, ГНЕЗДО, ИЕРАРХИЯ.

Предполагается, что в тезаурусе фиксируется: связь между ЛЕ и составляющими их словоформами; семантические связи, определяемые стандартами ISO [5, 6], а также дополняющие их виды отношений; частота употребления ЛЕ и словоформ; количество ЛЕ, включающих данную словоформу; количество рубрик, в которых встречается ЛЕ; количество ЛЕ, связанных с данной ЛЕ; номер уровня ЛЕ в иерархии связей. Перечисленные характеристики позволяют пользователю обоснованно принимать решения по преобразованию запросов.

Команда СИМВОЛЫ выделяет словоформы и ЛЕ, содержащие заданную последовательность символов. Команда УСЕЧЕНИЕ выбирает из тезауруса словоформы и ЛЕ, включающие заданное усечение словоформы. Команда КОМБИНАЦИЯ выбирает ЛЕ, в которые входит заданная комбинация усечений словоформ. Команда СВЯЗЬ выделяет из тезауруса ЛЕ, отношение которых с заданной ЛЕ описывается видом связи, отмеченным пользователем. Команда ГНЕЗДО выдает для заданной ЛЕ, которая служит заглавной ЛЕ для гнезда лексических единиц, все синонимические ЛЕ и ЛЕ, отстоящие на один уровень иерархии вверх и вниз по отношению к заглавной ЛЕ. Команда ИЕРАРХИЯ выявляет все ЛЕ, входящие в одну иерархию с заданной ЛЕ.

Все выдаваемые командами фрагменты тезауруса оформляются в листаемые списки ЛЕ. Против каждой ЛЕ в списке пользователь может поставить знак “+” или знак “-”. Пользователь размечает каждый список ЛЕ таким способом, что в случае нескольких последовательных ЛЕ с одинаковым знаком соответствующий знак может ставиться только у первой ЛЕ. В запрос включаются все ЛЕ, начиная со знака “+” до ближайшего знака “-” (в том числе и не помеченные знаком “+”). ЛЕ от знака “-” до ближайшего знака “+” (в том числе и не помеченные знаком “-”) не включаются в запрос.

5. Заключение

В статье предложены способы фиксации исходных потребностей пользователей ЭБ в виде структурированной анкетной формы. Предложенный подход позволяет всем категориям пользователей при работе с большим количеством коллекций гибко менять тактику взаимодействия в зависимости от промежуточных результатов общения с ЭБ, а также по мере приобретения опыта и навыков и помогает добиваться эффективных результатов на каждой стадии общения с посредником благодаря использованию тезауруса и рубрикатора.

Литература

[1] Kalinichenko L. A., Briukhov D. O., Skvortsov N. A., Zakharov V. N. Infrastructure of the subject mediating environment aiming at semantic interoperability of heterogeneous digital library collections // Труды Второй Всероссийской конференции по электронным библиотекам // Протвино, 2000 г., с. 78-90.

[2] Kazakov E. N., Vovchenko E. L. Use of polythematic thesaurus for mediators of digital library multicollections supporting their interactions with the users and collections. The RFBR DL Workshop, Moscow, December 1998.

[3] Казаков Е. Н. Формирование и ведение тезауруса в составе посредника между пользователями и сетью электронных библиотек // Труды Первой Всероссийской конференции по электронным библиотекам // Санкт-Петербург, 1999 г., с. 85-89.

[4] Кузнецов С. Д. SQL: Язык реляционных баз данных – М.: Майор 2001, 192 стр.

[5] ISO2788: Guidelines for establishment and development of monolingual thesauri, 2 nd ed., Geneva: ISO1986.

[6] ISO5964: Guidelines for establishment and development of multilingual thesauri, 1 st ed., Geneva: ISO1985.

THE WAYS OF FIXING OF THE INITIAL INFORMATION REQUIREMENTS OF THE USERS INTERACTING WITH DIGITAL LIBRARIES THROUGH THE MEDIATOR CONTAINING THE THESAURUS AND SUBJECT CLASSIFIER

Kazakov E.N.
The Scientific and Technical Information centre of Russia (VNTIC)
Smolnaja 14, Moscow, 125801
e-mail: postmaster@vntic.org.ru

The interaction of the users with digital libraries keeping numerous heterogeneous collections is esteemed. It is supposed, that the interaction will be realised with the help of a special infrastructure: the mediator having got personalization, interoperation (federated) and local levels. It is supposed, that the mediator contains the thesaurus and subject classifier, that provides the intellectual help to the users. The ways of fixing of the initial information requirements of the users are tendered. The mediator allows for these requirements at queries processing. The users can correct these requirements in an interactive mode.

The user fills in the form consisting of following blocks: the general, text of the query, explanation to the query, usage of the thesaurus and subject classifier, criterion of relevance.

Each block and its influencing on the result of the search is described. The commands of navigating under the thesaurus are tendered.