Тематические коллекции полнотекстовых запросов для изучения контекстного знания (проект Humanitariana)

С.Х. Ляпин[1], А.В. Куковякин[2]
[1] Санкт-Петербургский национальный исследовательский университет информационных технологий, механики и оптики (Университет ИТМО)
lyapins@yandex.ru
[2] ООО «Константа»
magicmagus@yandex.ru

Аннотация

Рассматривается использование сервисов полнотекстового поиска для создания *тематических коллекций полнотекстовых запросов*, представляющих собой одновременно и новую разновидность экспертного знания, и пользовательский инструмент для их (коллекций) дальнейшего пополнения. В этой связи кратко описывается функционал информационной системы T-Libra (используемой в рамках проекта Humanitariana), позволяющий сочетать абзацно-ориентированные и частотно-ранжированные запросы, а также осуществлять предварительную кластеризацию их результатов. В качестве примера описана методика создания коллекции запросов по теме «Проблема человека в русской философии».

Ключевые слова: проект Humanitariana; сервисы полнотекстового поиска; тематические коллекции поисковых запросов; кластеризация результатов запросов; абзацно-ориентированный запрос; частотно-ранжированный запрос.

1. Введение. От поисковых сервисов – к тематическим коллекциям запросов

Развитие поисковых сервисов над полнотекстовыми и комбинированными информационными ресурсами представляет собой одну из важных тенденций развития современной информационной среды [1]. Сказанное относится и к развитию многофункциональных электронных полнотекстовых библиотек и специализированных информационных систем, создаваемых на их основе [2], в частности к нашему проекту Humanitariana, направленному на создание виртуального информационно-ресурсного центра по извлечению и изучению контекстного знания в гуманитарной сфере [3]. Инструментарий

Сборник научных статей XVIII Объединенной конференции «Интернет и современное общество» IMS-2015, Санкт-Петербург, 23-25 июня 2015 г.
© Университет ИТМО

соответствующих информационных систем может быть использован для создания качественно нового вида информационных ресурсов, - ***тематических коллекций полнотекстовых запросов***, которые одновременно могут использоваться и как готовое тематизированное (экспертное, контекстное) знание, расширяющее состав информационных ресурсов электронной библиотеки, и как пользовательский инструмент для создания и развития аналогичных коллекций.

2. Архитектура информационной системы и базовые сервисы полнотекстового поиска для создания тематических коллекций поисковых запросов

2.1. Архитектура

Используемая нами в различных проектах информационная система T-Libra, предназначенная для создания многофункциональных электронных полнотекстовых библиотек (разработка ООО «Константа», Архангельск), функционирует в клиент-серверной Интернет/Интранет архитектуре. На стороне пользователя предполагается лишь наличие Интернет-браузера и стандартных прикладных программ по работе с файловыми ресурсами. На стороне сервера – операционная система Windows, СУБД MySQL (или аналогичная по функционалу), Веб-сервер Apache, сервер приложения (разработчик – ООО “Константа»). Вся бизнес-логика вынесена в сервер приложения.

Примечание. На серверах может функционировать также операционная система Linux с развернутой на ней виртуальной машиной Windows.

2.2. Сервисы полнотекстового поиска

В используемой нами текущей версии электронной библиотеки [4] имеются следующие типы полнотекстового поиска: ***а) абзацно-ориентированный, б) частотно-ориентированный.*** При этом абзацно-ориентированный поиск представлен разновидностями работы как в локальной, так и в распределенной среде. Для целей настоящего доклада используется его версия, включающая кластеризацию результатов запроса в ходе его выполнения (см. далее рис. 2).

Абзацно-ориентированный поиск предназначен для поиска и презентации текста с точностью до отдельных авторских абзацев, содержащих заданную пользователем терминологическую структуру (тем самым эксплицируется «горизонтальный» микроконтекст, в котором в составе абзаца находятся искомые термины). Авторский абзац выбран в качестве естественной единицы смыслового членения текста. Поиск ведется с учетом словоизменительной парадигматики (для русского языка). Обеспечивается поддержка нескольких видов и различных форм презентации результатов этого поиска.

Простой («однослойный») тематический поиск, с одним комплексным полем для ввода терминов и использованием для этих терминов операторов логического *объединения*, обязательного *исключения* или обязательного

включения термина в запрос. Результатом поиска является список абзацев, удовлетворяющих заданным условиям.

Каждый из абзацев, входящих в результаты запроса, может быть одним «кликом» мышки раскрыт до своего полного вида. Используя опцию «Контекст» в левом меню, можно последовательно раскрыть абзацы до и после найденного – вплоть до кластера из семи абзацев (три абзаца «до», три абзаца «после», плюс сам абзац – результат запроса).

Имеется возможность посмотреть, с этой же экранной страницы, соответствующий ресурс (статью, книгу и т.д.) в файловом виде; ресурс при этом может быть представлен в различных форматах: текстового документа, графического образа документа (важно для архивных ресурсов), сопровождающего документ аудио- или видеофайла (важно для организации электронных выставок и коллекций).

Расширенный («многослойный») тематический поиск. Этот вид поиска содержит функционал дополнительной тематической фокусировки запроса. Соответствующий инструментарий включает в себя: а) формирование нескольких поисковых полей («слоев») и б) включение в запрос дополнительных количественных параметров его фокусировки.

Поисковое поле "слой" представляет собой технический инструмент для выделения того или иного содержательного "аспекта" интересующей пользователя "темы"; всего может быть сформировано от 2 до 8 слоев. Например, в первом слое вводим термин «человек», во втором – термин «мир», в третьем – термин «жизнь». Тем самым в структуре запроса тематика «человека» специализирована (аспектуализирована) в связи с «миром» и «жизнью».

Еще более точная тематическая фокусировка запроса достигается за счет выполнения дополнительных условий: а) указания минимально необходимого количества поисковых слоев (от 2 до 8); б) указания максимального расстояния между терминами, принадлежащими разным слоям: от 0, когда слова из двух разных слоев запроса в составе абзаца примыкают друг к другу (например, «жизнь человека» и т.д.), до произвольной величины.

Частотно-ориентированный поиск предназначен для построения частотно-ранжированных списков терминов (существительных), и тем самым экспликации различных «вертикальных» макроконтекстов, неявно присутствующих в отдельном документе или их выбранной совокупности. Получающиеся таблицы списков терминов, с указанием абсолютного (в обычных числах) и относительного (в ‰, промилле) количества их встречаемости в тексте, мы называем «терминограммами» (по аналогии с «рентгенограммами»). Поиск может проводиться одновременно по 1, 2 или 3 корзинам ресурсов.

Примечание. Этот тип поиска используется также для предварительной кластеризации результатов абзацно-ориентированных запросов; в этом случае он встроен в механизм их осуществления.

Обеспечивается поддержка двух видов частотно-ориентированного поиска и различных форм презентации его результатов:

абсолютный частотный, результатом которого является частотно-ранжированный список существительных, входящих в ресурсы области поиска

и приведенных к нормальной форме (именительный падеж, единственное число);

относительный частотный, результатом которого является частотно-ранжированный список существительных, входящих только в те абзацы, которые содержат заданный пользователем термин (тем самым список строится «относительно» этого термина).

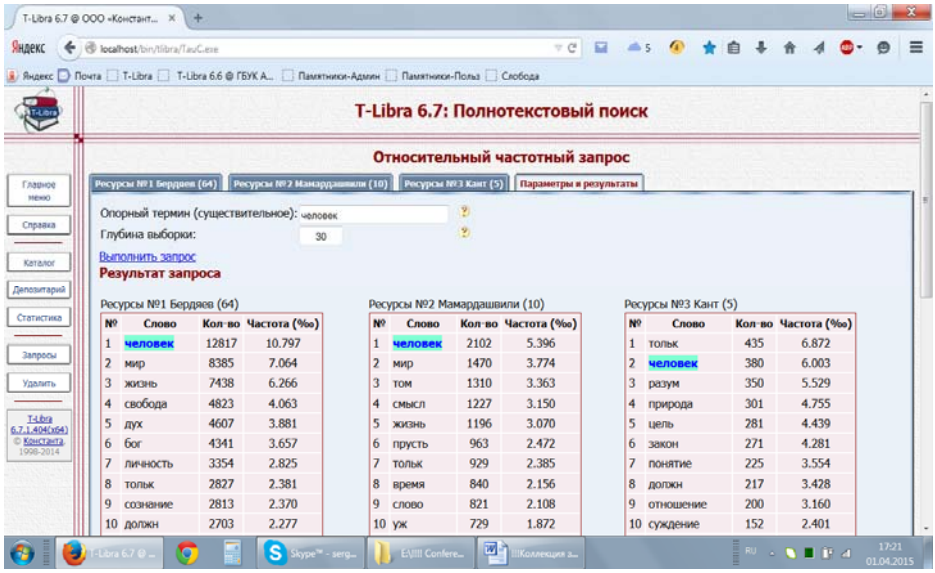

Рис. 1. Пример автоматического построения «терминограмм» (частотный запрос относительно термина «человек») по трем корзинам ресурсов: произведениям Бердяева (64 ед.), Мамардашвили (10 ед.), Канта (5 ед.).

Все термины, входящие в итоговую терминограмму, являются активными; «кликнув» по любому из них, можно выйти на уже сформированный абзацно-ориентированный запрос по данному термину, и эксплицировать его микроконтекст.

Эти виды частотного поиска могут использоваться для целей текстологического анализа документа; для выявления и описания предметной области документа; для составления списка ключевых слов; для сравнительного анализа предметных областей различных авторов или различных документов; для проверки текстов на плагиат и т.д.

3. Методика создания коллекции, фрагмент описания семантических полей коллекции «Проблема человека в русской философии»

Начать создание тематической коллекции запросов можно, осуществив многослойный запрос [Человек мир жизнь] по всей базе (в нашем случае - 2551 документ). Термин в первом слое «человек» определяется темой коллекции.

Термины «мир» и «жизнь» во втором и третьем слоях формы запроса мы заимствовали из терминограммы Бердяева (см. рис. 1, левая корзина ресурсов). Разумеется, можно было бы взять их из терминограммы Мамардашвили («мир» и «смысл») или Канта («разум» и «природа») – см. рис. 1, средняя и правая корзины ресурсов; – но это было бы не вполне корректно (во всяком случае, для Канта) с точки зрения темы «Проблема человека в *русской* философии».

Дополнительные параметры фокусировки запроса (все три слоя обязательны, расстояние между поисковыми терминами в абзаце не более 5 слов) подбираются пользователем (экспертом) с тем расчетом, чтобы в результате запроса было обозримое количество абзацев, содержание которых затем предстоит изучать.

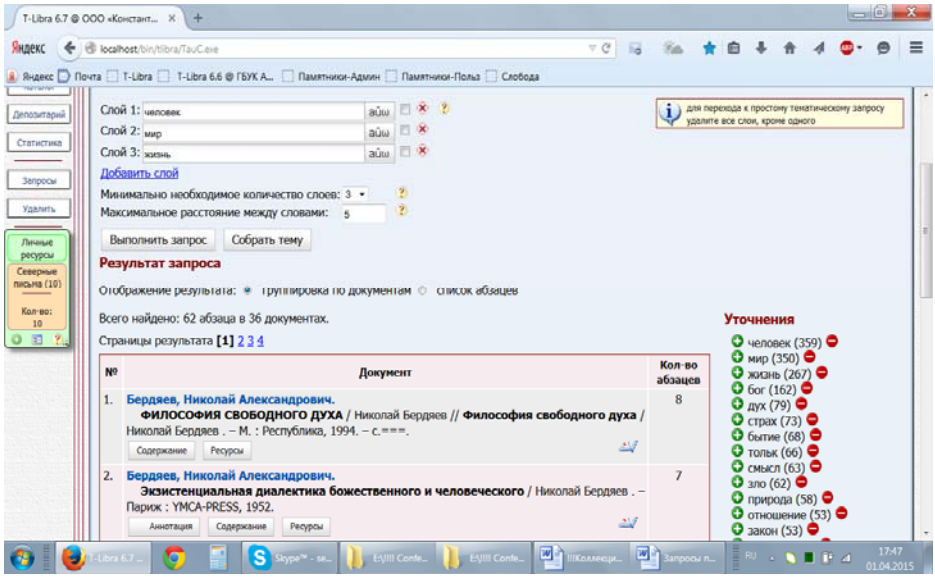

Рис.2. Результат многослойного тематического запроса [Человек мир жизнь] по всей базе в 2551 документ. Найдено 62 релевантных абзаца в 36 документах. В правом столбце экранной страницы - уточняющие результаты кластеризации по подмножеству из этих 62 найденных абзацев.

В нашем случае имеется 62 абзаца в 36 документах (см. рис. 2); произведена ранжировка по релевантным документам. Во время осуществления запроса производится предварительная кластеризация его результатов: на подмножестве из 62 найденных абзацев строится терминограмма (см. правую колонку «Уточнения» на странице результата запроса, рис. 2). В круглых скобках указано количество терминов (существительных) в подмножестве этих абзацев. Прямо с этой страницы можно *исключать* (знак –) и *включать* (знак +) термины из этой таблицы в абзацный поиск, тем самым варьируя поисковые запросы уже с учетом кластеризации.

Затем можно раскрыть один из найденных релевантных абзацев, например, из книги Н.А.Бердяева «Философия свободного духа» (см. рис. 3).

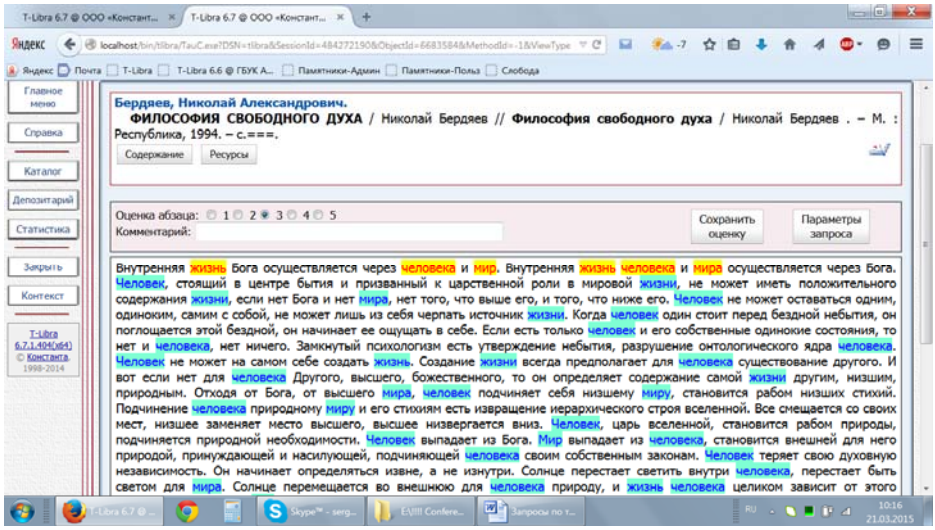

Рис.3. Раскрытый абзац из книги Н.А.Бердяева «Философия свободного духа». Красным цветом на желтом фоне покрашены термины, удовлетворяющие условиям фокусировки запроса (все три слоя обязательны, расстояние в абзаце между терминами запроса не более 5 слов).

Аналогичным образом можно раскрыть другие найденные абзацы (8 абзацев из книги Бердяева «Философия свободного духа», и остальные 54 абзаца из других произведений, найденные в результате запроса, см. рис. 2).

Далее проводим экспертный анализ найденных абзацев, и формируем перечень семантических (смысловых) полей и соответствующих ключевых слов. Для 8 абзацев из книги Н.А.Бердяева «Философия свободного духа» это будут (курсивом нами выделены ключевые слова):

- *Внутренняя жизнь Бога* осуществляется через человека и мир; *Внутренняя жизнь человека и мира* осуществляется через Бога; для утверждения *онтологического ядра человека* необходима *иерархии бытия*: высшего (Бога) и низшего (природы); *одинокий человек* поглощается *бездной небытия*; *Откровение* совершается в духе, а не в объективном мире и не в субъективном, психическом мире. Откровение есть *прорыв из духовного мира* в наш мир.

- В жизни человека как *природного существа* нельзя обнаружить *логоса*; *сознание без смысла*, как *приспособление* к *круговороту природного мира*; сознание, обращенное к Богу, открывает смысл, ему даны *знаки иного мира*.

- *Христианство* есть *религия благодати*, *теософия безблагодатна*; *иррациональное начало зла, вера* в *благую весть* от избавления от зла и наступлении *Царства Божьего*; *эсхатология; карма*; теософическое миросозерцание *магично*, а не *мистично*.

- Религия как *откровение Божества* в человеке и мире; выход из *состояния одинокости;* откровение как *событие духовной жизни* во внутреннем мире человека; *духовный опыт*; *Бог-Сын* как откровение небесной любви, *божественной соборности*.

- *Духовная индивидуальность* и Бог; потеря индивидуальности в *«природном» человеке*.

- *Мистерия Троичности* в жизни человека и мира*; встреча «одного» с «другим*», ее [мистерии] *разрешение в «третьем*».

- Всякое *лицо человеческое*… нужно *любить в Боге* и *через Бога*; *Христова любовь*; *откровение Духа Святого* в жизни человека; *мистика любви.*

- *Отвлеченная метафизика*; *гипостазирование явлений*; спиритуализм, материализм, идеализм; *конкретность бытия*; *бытие как жизнь*; панлогизм; *гипостазирование познающего субъекта*; *гипостазирование познаваемого объекта*; *сущность жизни*; метафизика германского идеализма; метафизика докантовской философии; *бытие есть акт* (а не субстанция); философия *духовной жизни*.

Аналогично могут быть раскрыты и проанализированы другие абзацы из других произведений, найденные в результате запроса.

В нашем случае мы получим описание примерно 60 смысловых полей по выбранной теме «Проблема человека в русской философии» для трехслойного запроса (рис. 2) и на нашей ресурсной базе (около 2500 полнотекстовых ресурсов гуманитарной направленности, из которых в результаты данного запроса попали 36 документов).

Можно сформировать многослойный запрос с другими терминами и тем самым другими аспектами темы. Например, [человек смысл Бог] и т.д. Это существенно (в разы или даже десятки раз) увеличит количество абзацев для дальнейшего анализа и формирования тематической коллекции. На практике можно ограничиться вышеприведенным запросом и кластером из 60 смысловых полей, поскольку сформированная на его основе «экспертная» коллекция сама становится инструментом для дальнейшего ее развития в соответствии с текущими интересами пользователя.

4. Функциональное место коллекции в структуре информационной системы T-Libra

В структуре ИС T-Libra имеется подсистема **Личные ресурсы.** Она предназначена для создания и управления совокупностью личных (персональных) хранилищ информации, полученной с помощью сервисов информационной системы T-Libra и на ее ресурсной базе.

Это могут быть: подборки библиографических описаний с прикрепленными к ним полнотекстовыми (полносодержательными) ресурсами; произвольные абзацы или их совокупности, полученные в результате полнотекстового поиска; готовые тематические подборки материала (опция Собрать тему в полнотекстовом поиске); терминограммы (частотно-ранжированные списки терминов); готовые структуры пользовательских запросов и т.д.

Для работы с тематическими коллекциями в составе ИС T-Libra разработан специальный блок этой подсистемы: Коллекции запросов.

В нашем случае 62 найденных абзаца (см. рис. 2) помещаются пользователем с помощью механизма drag and drop («перетащи и отпусти») в этот специализированный блок. Там размещаются:

1. Статическая форма / скриншот с формой запроса и его результатами (см. рис.2);

2. Список абзацев результатов запроса (в нашем случае – 62 ед.), с возможностью их раскрытия;

3. Предварительная кластеризация этих результатов (терминограмма и/или семантическое картирование подмножества найденных абзацев);

4. Фиксация и краткое описание семантических (смысловых) полей в подмножестве этих абзацев, сделанное экспертами/пользователями;

5. Выход в соответствующую поисковую форму электронной библиотеки (см. рис. 2) с возможностью дальнейшего тематического поиска по ее ресурсам. Таким образом, коллекция запросов является *интерактивной и динамической*, то есть является инструментом для настраиваемого полнотекстового поиска и тем самым для пополнения и развития коллекций.

6. Инструменты управления разделом «Коллекции запросов» личного хранилища.

5. Заключение

Методика и технология создания ***тематических коллекций запросов*** на основе сервисов продвинутого полнотекстового поиска электронной библиотеки позволяет существенно расширить и качественно развить состав информационных ресурсов библиотеки, добавив к ним готовое тематизированное (экспертное, контекстное) знание.

Поскольку коллекция запросов функционально интегрирована в электронную библиотеку, она является также ***пользовательским инструментом*** для расширения данной конкретной коллекции, а также создания аналогичных коллекций самой различной тематики.

Доклад подготовлен при поддержке гранта РГНФ № 14-03-12017

Литература

[1] С.Х. Ляпин, А.В.Куковякин, А.В.Чугунов. Распределенный полнотекстовый поиск в электронных библиотеках: технология и проекты // Труды 16-й Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» – RCDL-2014, Дубна, Россия, 13–16 октября 2014 г. С. 231–236. URL: http://rcdl.ru/doc/2014/paper/RCDL2014_231-236.pdf, (дата обращения: 01.04.2015).

[2] М.Е. Шварцман. Электронная библиотека как основа виртуальной исследовательской инфраструктуры. URL: http://www.aselibrary.ru/blogs/archives/1209/ (дата обращения: 01.04.2015).

[3] Ляпин С.Х., Куковякин А.В., Мбого И.А., Толстикова И.И, Чугунов А.В. Сервисы полнотекстового поиска в распределенной информационной среде (проект Humanitariana) // Научно-технический вестник информационных технологий, механики и оптики, Санкт-Петербургский национальный исследовательский университет информационных технологий, механики и

оптики (Санкт-Петербург), Т.15, №1, 2015 г. С. 155–162. URL: http://elibrary.ru/item.asp?id=23028758. (дата обращения: 26 марта 2015).
[4] См. Информационная система T-Libra для создания многофункциональных электронных библиотек с возможностями гибкого тематизируемого полнотекстового поиска. URL: http://demo.tlibra.ru (дата обращения: 01.04.2015).

Thematic collections of full-text queries for the study of contextual knowledge in the humanitarian field (project Humanitariana)

S. Kh. Lyapin[1], A.V. Kukovyakin[2]
[1]ITMO University, [2]Constanta, Ltd

Discusses the use of services full-text search to create thematic collections of full-text queries, which is both a new kind of expert knowledge, and a custom tool for them (collections) further funding and development. In this regard, briefly describes the functionality of the information system T-Libra (used in the framework of the project Humanitariana), combining a paragraph-oriented and frequency-ranked queries, and perform preliminary clustering results. As an example, the technique of creating a collection of queries on the topic "the Problem of man in Russian philosophy".

Keywords: the project Humanitariana; services full-text search; thematic collection of search queries; clustering of query results; paragraph-oriented request; a frequency-ranked query.