



С.Х. Ляпин, А.В. Куковякин

**Контекстное знание и его изучение с
помощью инструментов
полнотекстовой библиотеки**

Рекомендуемая форма библиографической ссылки

Ляпин С.Х., Куковякин А.В. Контекстное знание и его изучение с помощью инструментов полнотекстовой библиотеки // Научный сервис в сети Интернет: труды XVIII Всероссийской научной конференции (19-24 сентября 2016 г., г. Новороссийск). — М.: ИПМ им. М.В.Келдыша, 2016. — С. 240-248. — URL: <http://keldysh.ru/abraul2016/9.pdf>

Размещена также [презентация к докладу](#)

Контекстное знание и его изучение с помощью инструментов полнотекстовой библиотеки

С.Х. Ляпин¹, А.В. Куковякин²

¹ Университет ИТМО, Архангельский краеведческий музей

² ООО «Константа»

Аннотация. Рассматривается использование сервисов электронной полнотекстовой библиотеки для извлечения и изучения контекстного знания; исследование проводилось на технологической базе информационной системы T-Libra (разработка ООО «Константа», Архангельск). Приведены примеры экспликации микроконтекста («горизонтального» контекста в пределах авторского абзаца); макроконтекста («вертикального» контекста в пределах документа или их совокупности); кластеризуемой совокупности микроконтекстов; создания управляемой тематической коллекции запросов (интерактивное знание); мультимодального контекста (текст + графика); выявления предметно-тематических трендов (по тематике «электронного правительства»);

Работа выполнена при поддержке гранта РГНФ № 14-03-12017.

Ключевые слова: полнотекстовый поиск, контекстный поиск, абзацно-ориентированный запрос, частотно-ранжированный запрос, терминограмма, гибридный поиск, предметно-тематические тренды, мультимодальные ресурсы, коллекция тематических запросов.

Введение

Изучение структурно организованных, тематических и смысловых контекстов различного типа и вида, автоматически извлекаемых из неструктурированных данных, относится к числу проблем, важных для задач использования информационных технологий в сферах образования, науки, культуры, управления, бизнеса [1], [2].

С ней тесно связана задача *экспликации контекстного знания*. Решение последней обычно связывают с разработкой специализированных информационных систем [3]. Но вопрос может решаться – *и это является предметом настоящего доклада* – и на пути развития сервисов полнотекстового поиска в электронных библиотеках (как специализированных по содержанию ресурсной базы, так и универсальных).

При этом контекстное знание (т.е. знание, содержащееся в различных контекстах, полученных в результате полнотекстовых запросов) может

существовать в своих *различных видах и формах*; оно может быть извлечено и изучено в результате полнотекстового поиска и последующей его обработки.

В докладе рассматривается несколько конкретных случаев экспликации разных видов контекстного знания, полученного в рамках существующих сервисов ИС T-Libra, а также обсуждаются пути дальнейшего развития этого направления исследований и разработок.

Поиск для целей доклада проводился по базе данных, предоставленной ООО «Константа», и включающей около 3000 полнотекстовых русскоязычных ресурсов гуманитарной и общенаучной тематики.

2. Архитектура и поисковые сервисы информационной системы

Архитектура. Информационная система T-Libra, предназначенная для создания многофункциональных электронных полнотекстовых библиотек, 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.

Сервисы полнотекстового поиска

В используемой нами версии электронной библиотеки 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. 1).

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.

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"; слои между собой находятся в отношении *логического пересечения*, термины внутри слоя – в отношении *логического объединения*. 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), актуально используемых в запросе; б) указания максимального расстояния между терминами в составе авторского абзаца, принадлежащими разным слоям: от 0, когда слова из двух разных слоев запроса в составе абзаца примыкают друг к другу, до произвольной величины.

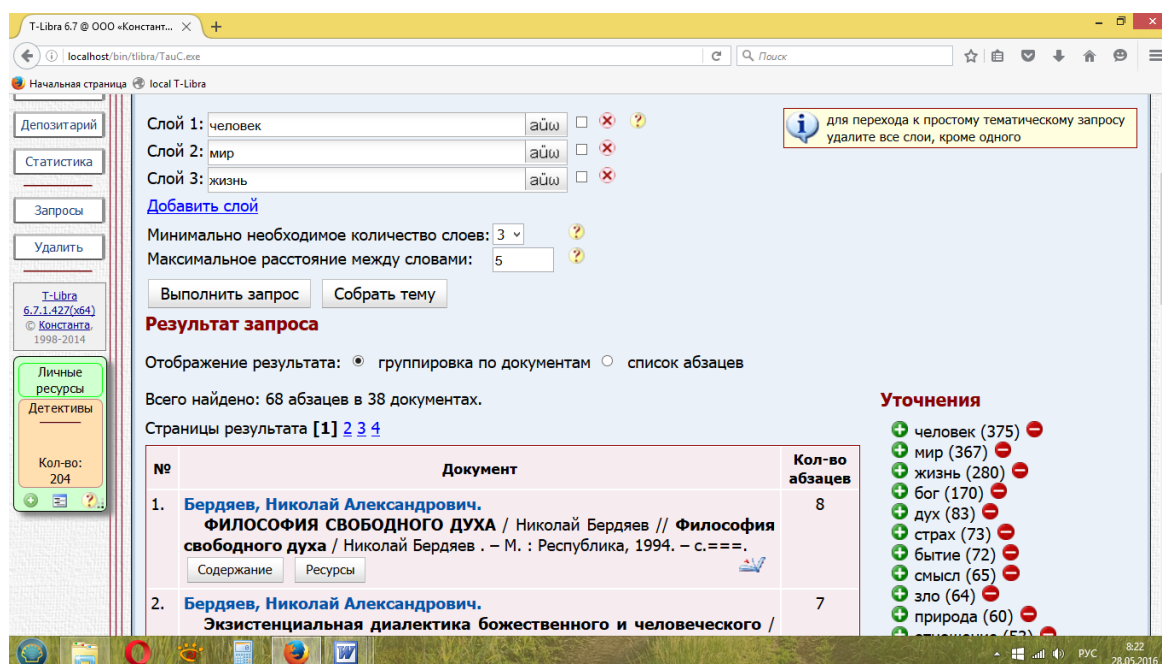


Рис. 1. Результат трехслойного запроса: слой 1 «человек», слой 2 «мир», слой 3 «жизнь». Условия фокусировки: все 3 слоя обязательны, расстояние в абзаце между терминами, находящимися в разных слоях запроса – не более 5 слов.

При этих условиях найдены 68 абзацев в 38 документах (по базе в 2979 документов). В правом столбце на странице – результаты кластеризации результатов этого запроса (получаются автоматически при осуществлении запроса).

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.

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).

Относительный частотный запрос

Ресурсы №1 Бердяев (68 ед.) Ресурсы №2 Мамардашвили (10 ед.) Ресурсы №3 Кант (5 ед.) **Параметры и результаты**

Опорный термин (существительное): ?

Глубина выборки: ?

[Выполнить запрос](#)

Результат запроса

Ресурсы №1 Бердяев (68 ед.)				Ресурсы №2 Мамардашвили (10 ед.)				Ресурсы №3 Кант (5 ед.)			
№	Слово	Кол-во	Частота (‰)	№	Слово	Кол-во	Частота (‰)	№	Слово	Кол-во	Частота (‰)
1	человек	12817	10.797	1	человек	2101	5.398	1	человек	380	6.003
2	мир	8385	7.064	2	мир	1470	3.777	2	разум	350	5.529
3	жизнь	7438	6.266	3	смысл	1225	3.147	3	природа	301	4.755
4	свобода	4823	4.063	4	жизнь	1193	3.065	4	цель	281	4.439
5	дух	4607	3.881	5	время	963	2.474	5	закон	271	4.281
6	бог	4341	3.657	6	слово	821	2.109	6	понятие	225	3.554
7	личность	3354	2.825	7	должен	717	1.842	7	должен	217	3.428
8	сознание	2813	2.370	8	тема	705	1.811	8	отношение	200	3.160
9	должен	2703	2.277	9	предмет	704	1.809	9	суждение	152	2.401
10	природа	2657	2.238	10	сознание	620	1.593	10	основание	148	2.338
11	бытие	2264	1.907					11	образ	145	2.291

Рис. 2. Пример автоматического построения терминограммы (частотный запрос относительно термина «человек» на глубину в 30 терминов) по трем корзинам ресурсов: произведения Бердяева (68 док.), Мамардашвили (10 док.), Канта (5 док.); на скриншоте показаны первые 11. Дата осуществления: 25.05.2016

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 для последующего контент-анализа.

3. Виды эксплицируемого контекстного знания

3.1. Экспликация микроконтекста («горизонтального» контекста)

Экспликацию микроконтекста («горизонтального» контекста в пределах авторского абзаца) удобнее всего производить с помощью многослойного тематического запроса, показанного, напр., на рис.1. Можно раскрыть в любом из найденных документов любой из найденных абзацев – например, из книги Н.А.Бердяева «Философия свободного духа».

Аналогичным образом можно раскрыть другие найденные абзацы (а в результаты этого запроса попали 38 произведений 17 различных авторов), и произвести анализ связанного с ними контекстного знания.

В нашем случае мы получим описание примерно 60 единиц контекстного знания по выбранной теме («Проблема человека в русской философии») для трехслойного запроса (рис.1) и на нашей ресурсной базе.

3.2. Экспликация макроконтекста («вертикального» контекста)

На рис. 2. приведен пример построения частотно-ранжированной *терминограммы*: таблицы терминов (существительных) *относительно* термина «человек» по трем корзинам ресурсов (произведениям Н.А.Бердяева, М.К.Мамардашвили и И.Канта). Соответствующий запрос строит эту терминограмму на указанную пользователем глубину (30 терминов в нашем случае) на подмножестве из тех абзацев, где имеется указанный термин («человек» в нашем случае).

Запрос позволяет определить те термины, с которыми в рамках авторского абзаца используется интересующий нас термин. Тем самым эксплицируется предметная область использования термина, и становится возможным сравнительный анализ по этому параметру различных авторов/документов.

3.3 Кластеризуемая совокупность микроконтекстов

В ИС T-Libra может быть использован инструмент *кластеризуемой совокупности микроконтекстов*, полученных в результате абзацно-ориентированного запроса.

На рис. 1 в правой части экранной страницы видна частотно-ранжированная таблица терминов (существительных), автоматически построенная во время осуществления запроса по найденному множеству релевантных абзацев. Она и представляет собой предварительную кластеризацию этого множества абзацно-ориентированных запросов.

Вместе с тем строки этой таблицы *активны*; можно, нажав на значок (-) справа возле термина, исключить этот термин из результатов запроса. Мы автоматически получим на этой же странице подмножество результатов исходного запроса, в абзацах которого нет этого термина. Соответственно будет изменяться и контекстное знание, эксплицированное запросом.

Таким образом, получаем кластеризуемую совокупность микроконтекстов (= результатов абзацно-ориентированных запросов), своего рода *динамический контекст*, и инструмент для работы с ним.

3.4. Тематические коллекции запросов (интерактивное контекстное знание). Сочетание абзацно-ориентированных и частотно-ранжированных запросов может быть использовано для выявления качественно нового вида контекстного знания – *тематических коллекций полнотекстовых запросов*, которые одновременно могут использоваться и как *готовое* тематизированное знание, расширяющее состав информационных ресурсов электронной библиотеки, и как *пользовательский инструмент* для создания и развития аналогичных коллекций.

В структуре ИС T-Libra имеется подсистема **Личные ресурсы**. Для работы с тематическими коллекциями разрабатывается специальный блок этой подсистемы: Коллекции запросов.

В нашем случае 68 найденных абзацев (см. рис. 1) помещаются пользователем с помощью механизма drag and drop («перетаски и отпусти») в этот специализированный блок. Там размещаются различные функциональные элементы этой коллекции, а также инструменты управления ею.

3.5. Экспликация мультимодального контекста (функциональная интеграция иконографических и полнотекстовых ресурсов). Одной из задач развития современных электронных библиотек является функциональная интеграция размещаемых в них нетекстовых ресурсов (например, иконографических описаний и изображений икон, реализуемых через каталог), и полнотекстовых ресурсов.

Иконографические описания могут быть размещены в электронном каталоге ИС T-Libra. Для этого используются поля библиографического каталога (стандарт RUSMARC), который имеется в информационной системе. Например, в поле Аннотация (поле 330 стандарта RUSMARC: «Резюме или реферат») помещается весь текст «музейной этикетки»; обычно это около 1 страницы текста.

Таким образом, средствами каталога может быть отображено иконографическое описание, а средствами поиска по каталогу найдена электронная карточка с прикрепленными к ней текстовыми и нетекстовыми (например, графическими) представлениями ресурса.

Далее пользователь нажимает на специальный ярлычок (Контекстный поиск), находящийся в электронной карточке, и запускает *гибридный запрос* одновременно по каталогу и полнотекстовой базе данных.

Алгоритм этого запроса включает в себя а) предварительную кластеризацию иконографического описания и, на основе этой кластеризации б) многослойный абзацно-ориентированный запрос по полнотекстовым ресурсам. Этот алгоритм учитывает также весовые коэффициенты: с *большим* весом – термины из *названия* иконы, с *меньшим* – из *аннотации*. Подбор весовых коэффициентов при разработке алгоритма производится экспертами.

Мы называем этот запрос *квазисемантическим*, поскольку семантика запроса учитывается косвенно, через указанный алгоритм. [4]

Результатом такого запроса является некоторое количество релевантных тематических абзацев из документов электронной библиотеки.

Для экспериментальной разработки алгоритма квазисемантического гибридного поиска нами был использован механизм многослойного тематического запроса.

В первый слой запроса были включены термины из названия иконы (в нашем случае – «Спас Нерукотворный»). Во второй – «убрус» и «лик». В третий – «Константинополь» и «Христос». Все эти термины были взяты из соответствующей электронной карточки (иконографического описания).

Варьируя другие параметры запроса, можно найти приемлемое количество релевантных абзацев. В нашем случае было найдено 59 абзацев в 44 документах (по базе в 2572 документа).

Характерно (для широкого контекстного поиска, о котором и идет речь), что получаемые абзацы – не только из иконографических книг или статей. Такого рода релевантных абзацев из общего числа найденных (59) оказалось около 20 (из таких ресурсов, как: словари, художественная литература, детективы, философия, богословие, история, культурология.). Это обстоятельство позволяет изучать более широкий культурно-исторический контекст рассматриваемой иконы.

Аналогичные исследования были проведены на вышеобозначенной ресурсной базе еще с 9 иконографическими описаниями.

Результаты исследований были использованы для разработки алгоритма квазисемантического гибридного поиска, функционально объединяющего иконографическое описание и полнотекстовые ресурсы.

Разумеется, изложенный здесь подход может быть использован для интеграции других (не только иконографических) ресурсов: графических, аудио, видео и т.п. – с полнотекстовыми ресурсами электронной библиотеки и создания мультимодальных тематических коллекций гибридных запросов, эксплицирующих интерактивное контекстное знание (см. п. 3.4.).

3.6. Экспликация мультимодального контекста (интеграция виртуального атласа и полнотекстового поиска). Этот вид контекстного поиска реализован в рамках проекта по созданию распределенной информационной среды «Пространство Ломоносова» для Архангельского краеведческого музея. Функциональную часть этой распределенной среды составляет виртуальный гипермедиа атлас «Земля Ломоносова». Он создается на основе технологии Google Earth [5] и содержит около 100 локаций («метки» на глобусе), связанных с именем и деятельностью М.В. Ломоносова.

В пределах конкретной локации реализовано как взаимодействие с другими элементами виртуального глобуса, так и автоматизированное взаимодействие с сервисами электронной полнотекстовой библиотеки. [6] Поиск «от локации на виртуальном глобусе – к электронной библиотеке»

позволяет в интерактивном режиме связать локации на глобусе с релевантными тематическими абзацами электронной библиотеки, тем самым актуализовать культурно-исторический контекст соответствующих событий и имен, связанных с деятельностью М.В. Ломоносова.

3.7. Анализ понятийно-тематических трендов

В рамках совместных исследований с сотрудниками Центра технологий электронного правительства (ЦТЭП) Университета ИТМО по использованию инструментов полнотекстового поиска ИС T-Libra для контент-анализа и выявления предметно-тематических трендов тематической коллекции текстов получены следующие результаты:

- предложена методика проведения контент-анализа коллекции текстов (5,2 тыс. новостных сообщений за пять лет (2011-2015) и выявления в них предметно-тематических трендов по теме «электронное правительство», сочетающая полнотекстовый поиск с автоматическим построением терминограмм (на основе T-Libra), инфографику (на основе MS EXCEL), статистический анализ (на основе прикладной программы SPSS) и экспертный анализ, осуществленный сотрудниками ЦТЭП Университета ИТМО. [7]

Из проведенного исследования следует, что в процессах электронного взаимодействия власти и гражданского общества современной России доминирующие позиции занимает *власть*. Связаны данные процессы прежде всего с развитием *интернета* в разных *регионах* (областях) и основную роль в этих процессах СМИ отводит *правительству*, оказывающему государственные услуги через специализированные *порталы*.

Факторный анализ с использованием программы SPSS показал также, что употребляемые термины статистически описываются двумя факторами. Главный фактор можно обозначить как взаимодействие *человека* (пользователя) и *государства* (власти, организатора процессов). Другой фактор – взаимодействие *управления* (чиновника, исполнителя процессов) и *технологии* (инфраструктуры). Выделение данных факторов позволяет внести коррективы в разработанные ранее модели электронного взаимодействия власти и общества.

С точки зрения тематики данного доклада, можно говорить о новом виде контекстного знания: *предметно-тематических трендах* за тот или иной период и по той или иной теме.

Заключение

Как показывают изложенные выше результаты, исследования и разработки по использованию инструментов продвинутого полнотекстового поиска для выявления контекстного знания достаточно эффективны и приводят к выявлению новых видов контекстного знания.

Они могут быть продолжены – как в плане расширения ресурсной базы, на которой они проводится, так и в плане развития соответствующего инструментария.

Литература

1. Integrating Unstructured Text into the Structured Environment // Prentice Hall — 0132360292, <http://cdn.ttgtmedia.com/searchDataManagement/downloads/UnstructuredData4.pdf>. (дата обращения 29.04.2015).
2. Ляпин С.Х., Куковякин А.В. Обработка неструктурированных данных (извлечение контекстного знания) с помощью сервисов полнотекстового поиска в электронной библиотеке // XVII Международная конференция DAMDID/RCDL-2015 Обнинск, 13-16 октября 2015. – Труды конференции / под ред. Л.А. Калиниченко, С.О. Старкова – Обнинск, ИАТЭ НИЯУ МИФИ, 2015 – 525 с. – С. 164-167.
3. Николай Ильин, Сергей Киселев, Владислав Рябышкин, Сергей Танков. Технологии извлечения знаний из текста // «Открытые системы», № 06, 2006. URL: <http://www.osp.ru/os/2006/06/2700556/>. (дата обращения 29.04.2015)
4. Ляпин С.Х., Куковякин А.В. Северная икона в мультимодальной библиотеке: к функциональной интеграции иконографических и полнотекстовых ресурсов // Сборник научных статей XVIII Объединенной конференции «Интернет и современное общество» IMS-2015, Санкт-Петербург, 23-25 июня 2015 г. – с. 216-224.
5. Google Earth (обзор) // URL: http://itc.ua/articles/google_earth_22033/ (дата обращения: 15.04.2016).
6. Ляпин С.Х., Куковякин А.В. Пространство Ломоносова: опыт функциональной интеграции виртуального атласа и полнотекстовой библиотеки. // Сборник научных статей межд. конф. EVA-2016. Санкт-Петербург, Университет ИТМО, 22-24 июня 2016 г. [В печати].
7. Ляпин С.Х., Куковякин А.В., Кудрявцева М.В. Использование инструментов электронной библиотеки для выявления понятийно-тематических трендов // Сборник научных статей XIX Объединенной конференции «Интернет и современное общество» IMS-2016, Санкт-Петербург, Университет ИТМО 22-24 июня 2016 г. [В печати].