Синтетический метод извлечения контекстного знания в русскоязычной социально-гуманитарной сфере: комплексный подход

О.В. Кононова[1], С.Х. Ляпин [1], Д.Е. Прокудин [2, 1]

[1] Университет ИТМО, [2] Санкт-Петербургский государственный университет, Университет ИТМО

kononolg@yandex.ru, lyapins@yandex.ru, hogben.young@gmail.com

Аннотация

Авторы предлагают синтетический метод извлечения контекстного знания как комплексный подход, ориентированный на решение проблемы экспликации контекстного знания в русскоязычных текстах гуманитарной и социальной направленности с последующей типологией контекстного знания как по отдельным дисциплинам, так и по междисциплинарным связям, а также по общенаучным и общекультурным термин-концептам, объединяющем различные дисциплины.

Проблема имеет общенаучное и общекультурное значение, поскольку контекстное знание составляет важную часть явного и неявного знания, которым оперирует человек в научной, образовательной и культурной деятельности.

Предлагаемый подход к решению проблемы экспликации контекстного знания состоит в расширении возможностей технологий гибкого тематизируемого полнотекстового и мультимодального поиска, реализованного и апробированного в рамках деятельности виртуального ресурсно-сервисного центра, созданного в 2014-2016 гг. на базе Университета ИТМО в рамках проекта РГНФ № 14-03-12017 (проект «Гуманитариана»).

Ключевые слова: синтетический метод, полнотекстовый поиск, мультимодальный поиск, контекстное знание, микро-контекст, макро-контекст, терминограмма, семантическое картирование, поисковый тезаурус

1. Введение

Важной тенденцией, характеризующей современное информационное общество, является изменение требований его членов к потребительским

качествам информации, способам и инструментам ее получения и анализа. Под удовлетворением информационных потребностей понимается, в первую очередь, обеспечение открытости и достоверности данных, а также обеспечение возможности самостоятельного управления информационными ресурсами — поиска, сбора, анализа, оценки информации, и таким образом, новых знаний [12, 17, 36]. При этом именно качество и полнота удовлетворения информационных потребностей социума выступает факторами повышения эффективности управления предприятиями, организациями, государственным сектором экономики и государственными институтами, дальнейшего развития информационного общества. Удовлетворение информационных потребностей достигается применением технологий и методов извлечения контекстных знаний из больших массивов данных. Данные методы включены в большинство современных BI систем, предназначены для поддержки проведения бизнес-аналитики, обеспечения свободного поиска и интеллектуальной выборки данных в систематизированных хранилищах информации, а также для поддержки научно- исследовательской деятельности в организациях и компаниях [24].

В настоящее время проводятся различные исследования в области разработки и применения различных методов поиска и извлечения контекстного знания. Исследования можно условно разделить на несколько групп в зависимости от преследуемых целей и / или предметной области. К первой группе следует отнести исследования общетеоретического характера, которые поднимают вопросы разработки эффективных методов и технологий поиска и извлечения контекстного знания как таковых [5, 7, 8]. Как правило, под предметной областью в этом случае понимаются поисковые интернет-системы или поисковые системы частных и публичных библиотек. Отличия — стандартность решений, и, в еще большей степени, стандартность используемых инструментов. Акцент делается на интенсификацию поисковой деятельности и полноту охвата области поиска, а также на реализацию возможности поиска объектов разной природы и формата (текстовые, графические данные и пр.).

Вторая группа исследований и разработок ИС ориентирована на потребности бизнеса, и соответственно на использование контекстного поиска и экспликации знаний как инструментов в аналитических модулях BI систем. Проблемы применения методов контекстного поиска в бизнес-компаниях описаны в статье А. Шуклина [36].

К третьей группе исследований следует отнести проекты, ориентированные на изучение мнений и поведения интернет-сообществ, СМИ публикаций, выделение нового знания, обнаружения тенденций в развитии интересов социума, применением методов извлечения, обобщения и агрегирования разнородных данных. Отличительной особенностью данной группы исследований является применение научных методов обработки данных с целью извлечения знаний к ненаучному текстовому массиву [3, 15, 16, 21, 22, 23, 34].

Общими недостатками этих исследований является то, что в них недостаточно комплексно рассматриваются вопросы разработки алгоритмов поиска и создания соответствующих информационных систем количественного

накопления информации, а также применение этих систем и алгоритмов при качественном анализе рассредоточенных массивов текстовых данных, размещённых в разнородных информационных системах. Это, в свою очередь, актуализирует необходимость выявления типовых структур контекстного знания на основе информационного поиска по мультимодальным базам данных, и, в этом смысле, на автоматизацию и формализацию методов изучения контекстного знания.

Актуальность проблемы выделения контекстного знания для социогуманитарного знания заключается в том, что, хотя контекстное знание активно используется при осмыслении и интерпретации гуманитарных и социальных текстов, а также при организации проектно-контекстной образовательной деятельности, но делается это в основном на интуитивном социально-психологическом уровне, без структурированного и специфицированного понимания «контекста» и «контекстного знания» для различных дисциплин и междисциплинарных связей.

Обработка текстовой и графической информации научной природы, в частности научной информации в области социально-гуманитарной сферы требует особых подходов и методов поиска, извлечения в связи с многочисленностью связей и зависимостей между отдельными термин-концептами научного текста, а также чрезвычайным многообразием контекстов использования одних и тех же понятий, значительных различий в трактовке микро и макро-контекстов использования понятий. Также следует принимать во внимание, что при работе с научными текстовыми массивами фактически речь идет о применении методов обработки данных непосредственно к «знаниям» и, следовательно, нахождения нового знания путем поиска и анализа контекстов уже известного, зафиксированного знания.

Важным аспектом изучения гуманитарного и социального знания является выявление и анализ его содержательных и смысловых контекстов («контекстного знания») разного типа, вида и уровня. Авторами разработан комплексный подход для предварительной экспликации и последующего анализа контекстного знания с использованием технологии полнотекстового поиска, реализованной и развиваемой в информационной системе «Гуманитариана», созданной в 2014-2016 гг. на базе Университета ИТМО в сотрудничестве с партнерами в рамках гранта РГНФ №14-03-12017-в, URL: http://77.234.221.107/tlibra/. В его рамках при координирующей роли Университета ИТМО (г. Санкт-Петербург) поэтапно создавалась междисциплинарная информационная распределенная среда с открытым доступом, разрабатывалась типология задач автоматизированного извлечения контекстного знания из научных текстов, создавались запросы разного типа и вида для типовых задач извлечения знаний.

2. Изучение контекстного знания как исследовательская проблема

Как в русскоязычном, так и в мировом научном дискурсе изучение контекстного знания ведется в основном по двум направлениям: 1) теории и практике контент-анализа и 2) теории и практике контекстного обучения.

1. *Контент-анализ.* Исследование контекстов различного типа, вида и уровня проводятся в рамках традиционного контент-анализа: метода и технологии качественно-количественного анализа документов в целях выявления или измерения социальных фактов и тенденций, отраженных этими документами. Контент-анализ изучает документы в их социальном контексте.

Все более широко распространяется контент-анализ сообщений средств массовой информации, основанный на парадигматическом подходе, в соответствии с которым изучаемые признаки текстов (содержание проблемы, причины ее возникновения, проблемообразующий субъект, степень напряженности проблемы, пути ее решения и др.) рассматриваются как определенным образом организованная структура [31].

Вместе с тем, традиционный контент-анализ и его результаты существенно определяется заранее заданными и внешними по отношению к тексту категориями (целевыми задачами и ключевыми понятиями, задаваемыми исследователем).

2. *Контекстное обучение.* Для изучения контекста и контекстного знания характерным является также социально-психологический подход, наиболее распространенный в теории и практике контекстного обучения. Речь идет о проектировании и использовании обучающих социальных ситуаций и ролевых игр как форм контекстного обучения. Концептуальной основой такого подхода является теория контекстного обучения и воспитания, разработанная в научно-педагогической школе А.А. Вербицкого. Суть контекстного обучения — последовательное моделирование в формах учебной деятельности студента предметного и социального содержания его будущей профессиональной деятельности [14, 20]. Несмотря на практическую эффективность контекстного обучения, в рамках этой методологии и технологии само контекстное знание используется на интуитивном уровне, не формализуется и не исследуется его содержательная структура.

3. *Комплексное изучение контекстного знания*. Для систематического изучения контекстного знания в 2014 году был создан виртуальный информационно-сервисный центр Humanitariana («Гуманитариана»). Этот центр создан на базе Университета ИТМО (с партнерами) в рамках проекта РГНФ № 14-03-12017. Его основой является многофункциональная электронная библиотека с возможностями гибкого тематизируемого полнотекстового поиска как в локальной, так и в распределенной среде (http://77.234.221.107/tlibra). На его ресурсно-сервисной основе были реализованы пилотные проекты по изучению некоторых конкретных аспектов контекстного знания: разработка коллекции тематических полнотекстовых запросов для экспликации контекстного знания; экспликации понятийно-тематических трендов по тематике «Электронное правительство» на основе публикаций в электронных СМИ. Результаты этих и других исследований, апробированные на крупных

российских и международных конференциях в 2014-2016 годах и опубликованные в рецензируемых журналах, подтвердили предположение об актуальности изучения проблем контекстного знания, правомерности и эффективности использования для этих целей автоматизированных методов полнотекстового поиска.

Помимо этих направлений можно выделить подходы, затрагивающие данную тематику:

1) технологический, направленный на разработку информационных систем и реализации в поисковых системах алгоритмов контекстуального [9, 25, 33];

2) семантический, связанный с разработкой и применением лингвистических подходов к анализу тестов и выявлению в них определённых смыслов [1, 10, 18, 32, 35];

3) содержательный, состоящий в прикладном применении алгоритмов поиска, анализа информации в функционирующих информационных системах для количественной обработки текстов и качественном анализе в них содержащихся смыслов из определённых предметных областей, например, [6]. Кроме этого, достаточно важным направлением исследований является анализ возможности обработки разнородных данных (текстов), распределённых в различных разнородных информационных системах с доступом как из глобальных, так и из локальных сетей. Это особенно актуально в свете развития Grid-технологий распределения информации для возможности её систематизации и обобщения при постоянном её количественном приращении [11, 13, 19].

Однако, проводимые в настоящее время исследования не являются в полной мере комплексными, т.е. объединяющими в себе все указанные подходы: от разработки алгоритмов поиска и создания соответствующих информационных систем количественного накопления информации до применения этих систем и алгоритмов при качественном анализе рассредоточенных массивов текстовых данных, размещённых в разнородных информационных системах.

Отличительной особенностью предлагаемой к использованию для изучения контекстного знания информационной системы «Контекстное знание» является ее многоплановая работа с мультимодальными полнотекстовыми библиотеками (или интегрированной системой библиотека + мультимодальная коллекция), в том числе в распределенной информационной среде.

Исследованиями в данной области занимается ряд специалистов как за рубежом, так и в России. В настоящее время ведущими центрами изучения контекстного знания, методов и технологий его извлечения и анализа являются университетские лаборатории и центры, специализированные научные центры, а также исследовательские структуры крупных коммерческих организаций [2, 6, 18, 25, 32, 33].

Авторы разрабатываемого комплексного подхода к изучению контекстного знания активно развивают данную проблематику, а результаты их исследований представлены в научной литературе [4, 24, 26–30].

3. Информационные системы лингвистического анализа и обработки текстов

В настоящее время разработано и используется достаточное число программных продуктов, ориентированных на предоставление специализированных услуг по обработке неструктурированной текстовой и нетекстовой информации.

Программы лингвистического анализа текста. Эти программы нацелены на решение конкретных, чаще всего весьма специальных задач по анализу текста. Не все из них обеспечивают полнотекстовый поиск. Их перечень и подробное описание представлено на портале http://asknet.ru/analytics/programms.htm. Среди основных из них можно отметить следующие:

- поисковая система AskNet. Семантическая вопросно-ответная система, реализующая семантический анализ текста на русском и английском языках. Существуют коробочные версии (корпоративная, сайтовая и персональная системы);
- семейство систем RCO (Russian Context Optimizer). В частности, библиотека для разработчика информационно-поисковых систем RCO Text Categorization Engine.. Позволяет на основании лексических профилей определять принадлежность текста к заданному множеству категорий; для каждого термина из лексических профилей, обнаруженного в тексте, получить количество его вхождений в текст, а также позиции терминов в тексте;
- программные продукты семейства Ontos (OntosMiner, Light Ontos for Workgroups, Ontos SOA, TAIS Ontos). Семантический поиск, анонсированный разработчиками, в действительности таковым не является, поскольку сводится к поиску по ключевым словам с использованием тематических синонимов;
- программно-аппаратный комплекс Google Mini (для корпоративной сети, в России не представлен) и программа Google Desktop (для персонального пользователя). Реализуется поиск по ключевым словам для текстов на многих языках. Нет полнотекстового поиска;
- программа специализированного http-сервера Яndex.Server. Она позволяет индексировать и обеспечивать поиск одновременно для одного или нескольких сайтов или компьютеров пользователя. Поиск работает с учетом морфологии русского, украинского и английского языков. Результат поиска — совокупность документов, упорядоченных по релевантности или дате. Синтаксический и семантический анализ текста не реализованы;
- Galaktika-ZOOM. Программа позволяет выявлять значимые слова и словосочетания документа, проводить поиск документов по вводимым пользователем ключевым словам с учетом их синонимов, а также формировать отчеты по частоте встречаемости слов в документах;
- кластеризующие поисковые системы (Vivisimo, Nigma). Метапоисковые системы с кластеризацией результатов поиска обеспечивают возможность выделения слов, часто встречающихся

совместно со словами поискового запроса. Однако использование только кластерного анализа не дает существенных преимуществ метапоисковым системам данного типа. Улучшение качества поиска, особенно при обработке запросов на естественном языке возможно только на основе использования синтаксического и семантического анализа.

Анализ возможностей существующих информационных систем позволяет сделать вывод, что наиболее близкими по функционалу к информационной системе, предполагаемой к использованию при исследовании контекстного знания, являются программы Яndex.Server и Nigma.

4. К постановке задачи разработки комплексного подхода изучения контекстного знания

Хотя термин «контекстное знание» не является концептуальной инновацией, но востребованность данной тематики прослеживается как в научных кругах, так и на практике — в бизнесе. Несмотря на высокую востребованность, анализ отечественных и зарубежных публикаций демонстрирует наличие пробелов в изучении и трактовке как самого термина, так и недостаток технологий поиска, выделения и анализа контекстного знания. Таким образом, исследовательская повестка как с теоретической точки зрения, так и с научно-практической не исчерпана.

Синтетический метод извлечения контекстного знания из научных текстовых массивов русскоязычной социально-гуманитарной сферы, предлагаемый авторами, предполагает использование методов и технологий:

- сочетание абзацно-ориентированных полнотекстовых запросов, позволяющих эксплицировать «горизонтальный» контекст (микроконтекст) употребления поисковых терминов в составе авторского абзаца, и частотно-ранжированных запросов, позволяющих эксплицировать предметную область документа или совокупности документов (вертикальный контекст, или макроконтекст);
- гибридный поиск, то есть одновременный поиск по каталогу и по полным текстам; позволяет соотнести объекты каталога, в том числе нетекстовые (графические образы, аудио и видеоресурсы, найденные по описаниям в каталоге), и релевантные фрагмента текста, найденные по полнотекстовой библиотеке;
- мультимодальный поиск, объединяющий поиск по мультимодальной (например, музейной) коллекции и полнотекстовый поиск фрагментов текста, релевантных объектам коллекции;
- кластеризация результатов абзацно-ориентированного поиска с возможностями управлением параметрами кластеризации, что позволяет выявить кластеры контекстного знания, соотнесенные с термин-концептами;
- семантическое картирование результатов запроса (визуализация кластеров результатов запроса).

Новизна предлагаемого подхода определяется тем, что впервые разрабатывается комплексное исследование, направленное на описание контекстного знания в русскоязычной социогуманитарной сфере с использованием систем продвинутого полнотекстового и мультимодального поиска. Также впервые в автоматизированном режиме будет изучена типология контекстного знания, будет произведено его семантическое картирование и структурированное описание как в дисциплинарном (история, философия, культурология, политология, экономика и т.д.), так и в междисциплинарном плане (выявление контекстного знания, связывающего различные дисциплины).

На основе разрабатываемого подхода предполагается разработать и апробировать новые технологии полнотекстового поиска (например, модуляционно-имитационное изучение контекстов гуманитарного и социального знаний: экспликация контекстуального знания в произведениях одного автора на основе структурированных контекстов другого и наоборот; или экспликация понятийно-тематических трендов в материалах электронных СМИ по тематике «электронное правительство», «технологии информационного общества», «компьютерные игры», «архитектурный и инженерный подходы в государственном управлении»).

Это касается также гибридного мультимодального поиска, позволяющего эксплицировать разнообразные текстовые и релевантные им нетекстовые контексты (графика. аудио, видео).

Немаловажной задачей представляется также разработка специализированных тезаурусов для экспликации контекстного знания, дальнейшая их апробация (проверка на эффективность и релевантность) в рамках запросов в создаваемой информационно-поисковой системе.

При этом основной целью является разработка эффективной научной методики и технологии автоматизированного извлечения и изучения русскоязычного контекстного знания на релевантном массиве информационных ресурсов как текстовой, так и нетекстовой модальности (графика, аудио, видео). Для этого предлагается решить целый ряд конкретных задач:

- разработать технологию мультимодального поиска, обеспечив функциональную интеграцию электронного каталога, мультимодальной коллекции и полнотекстовой библиотеки для извлечения контекстного знания в гуманитарной и социальной сфере, как в локальной, так и в распределенной информационной среде;
- обеспечить на этой основе автоматизированное извлечение различных типов, видов и уровней контекстного знания на материальной базе полнотекстовых документов социально- гуманитарной направленности и мультимодальных (музейных) коллекций;
- разработать типологию выявленного контекстного знания, дать его структурированное описание.

5. Методы и технологии комплексного изучения контекстного знания русскоязычной социально-гуманитарной сфере

В разрабатываемом подходе предлагается анализировать текст и нетекстовые модальности информации сначала на более высоком (и более

абстрактном) уровне — уровне обобщенных структурных инвариантов контекстуального знания, а затем редуцировать и специфицировать их применительно к конкретным дисциплинарным обстоятельствам и междисциплинарным отношениям. В традиционном контент-анализе первичными являются целевая функция и категории анализа, вторичными — получаемые «обобщенно-текстовые» единицы анализа; в подходе, характерном для настоящего проекта, первичен «обобщенный текст» (с элементами мультимодальной информации), вторичен получаемый «контент», т.е. структурированное описание контекстуального знания. Можно сказать, что традиционный контент-анализ и предлагаемый «нетрадиционный» анализ контекстного знания являются дополнительными друг другу методами и технологиями изучения содержательных и смысловых информационных контекстов.

«Обобщенный текст» (текст + мультимодальная информация) в этом случае является генератором эксплицируемых контекстов и, соответственно, структур контекстного знания. Инструментом генерации являются гибкие функциональные структуры мультимодальных запросов.

Решение поставленных задач в рамках реализации предлагаемого исследования предполагает использование следующих методов и технологий:

- метод и технология сочетания абзацно-ориентированных полнотекстовых запросов, позволяющих эксплицировать «горизонтальный» контекст (микроконтекст) употребления поисковых терминов в составе авторского абзаца, и частотно-ранжированных запросов, позволяющих эксплицировать предметную область документа или совокупности документов (вертикальный контекст, или макроконтекст);
- метод и технология гибридного поиска, то есть одновременного поиска по каталогу и по полным текстам; позволяет соотнести объекты каталога, в том числе нетекстовые (графические образы, аудио и видеоресурсы, найденные по описаниям в каталоге), и релевантные фрагмента текста, найденные по полнотекстовой библиотеке;
- метод и технология мультимодального поиска, объединяющая поиск по мультимодальной (например, музейной) коллекции и полнотекстовый поиск фрагментов текста, релевантных объектам коллекции;
- метод и технология кластеризации результатов абзацно-ориентированного поиска с управлением параметрами кластеризации. Позволяет выявить кластеры контекстного знания, соотнесенных с термин-концептами;
- метод и технология семантического картирования результатов запроса (визуализация кластеров результатов запроса).

Предлагаемые методы и технологии опираются также на следующие конкретные методики и инструменты поиска:

- методика сочетания абзацно-ориентированных и частотно-ориентированных запросов, позволяющий объединять в исследовательских целях эксплицируемые «горизонтальные» микроконтексты (в пределах авторского абзаца) и «вертикальные» макроконтексты (в пределах документа или их совокупности);

- инструмент автоматической кластеризации результатов абзацно-ориентированного запроса с обратной связью с поисковым запросом (позволяет осуществлять кластеризацию запроса и управлять ее параметрами);
- инструмент многослойного тематического запроса с вариацией используемых слоев, позволяющий выделять аспекты эксплицируемой темы
  (от 2 до 8 аспектов);
- инструмент фокусировки полнотекстового запроса (позволяющий задавать расстояние между поисковыми терминами в искомом абзаце, и находить оптимальное соотношение между полнотой и точностью поиска);
- инструмент каскадного поиска (результаты одного запроса автоматически входят в поисковый образ другого запроса; позволяет осуществлять структурную модуляцию запроса для выявления новых элементов контекстного знания;
- инструмент гибридного квазисемантического поиска (одновременно по описаниям ресурса, взятым из каталога, и по полным текстам; используется для мультимодального поиска — например, по описаниям музейных артефактов и по полным текстам библиотеки;
- инструмент тезаурусного поиска (абзацный поиск с автоматическим включением в поисковый образ функциональной структуры тезауруса; используется для автоматического расширения культурного контекста в ходе выполнения запроса).

В качестве инструментального ядра исследования предлагается использовать программную среду «Humanitariana»). Эта информационная система может функционировать в режиме локальной сети и в режиме распределенной информационной среды с возможностью обращения ко всем ресурсам с любого из серверов организаций-участниц. В ней реализованы следующие функции, необходимые для реализации проекта: 1) модуль сбора и загрузки данных (текстовых массивов); 2) модуль, реализующий несколько видов тематического контекстного поиска и извлечения знаний: частотно-ориентированный поиск, абзацно-ориентированный поиск (простой («однослойный») контекстный запрос или расширенный («многослойный») контекстный запрос); 3) модуль создания тематических коллекций материалов на основе кластеризации результатов запросов, анализа горизонтальных и вертикальных микро- и макро-контекстов.

В ходе проведения исследования предполагается расширить возможности полнотекстового поиска (в том числе разработать гибридный поиск: запрос одновременно по каталогу и полным текстам; мультимодальный поиск: запрос одновременно по мультимодальной коллекции и полным текстам и др.). Все основные виды запросов и работа с их результатами возможны как в локальной среде участников проекта, так и в распределенной среде.

6. Материально-техническая база исследования

Предлагаемый подход разрабатывается для практической реализации с использованием в качестве ядра информационной системы «Гуманитариана», созданной в 2014-2016 гг. на базе Университета ИТМО. Система разработана в многозвенной клиент-серверной Интернет/Интранет архитектуре: Web-browser / Web-server + ApplicationServer / Relational DBMS, с протоколами HTTP, CGI, PIPE API, ODBC.

Функционирует в среде Windows: на сервере — Windows 2000/2003/XP/Vista/Windows 7, СУБД MySQL, Веб-сервер Apache, сервер приложения.На серверах может функционировать также операционная система Linux. Серверприложения охватывает всю бизнес-логику системы.

На клиентском месте — любая из версий Windows. Клиентской программой для ЭБ «Humanitariana» является стандартный Веб-браузер (поддерживаются MS InternetExplorer, MozillaFirefox, Opera, AppleSafari, GoogleChrome, Яндекс.Браузер).

С учетом тенденций развития современного информационного пространства была выбрана модель децентрализованной среды под управлением пользовательского браузера и с ориентацией на Веб-сервисы и Интернет-протоколы. Браузер обращается к множеству независимых серверов, находящихся в ведении различных организаций. Прямого взаимодействия серверов при этом не требуется.

«Humanitariana» реализует инструментарий контекстного поиска в русскоязычных и англоязычных текстах, анализ горизонтальных и вертикальных макро и микроконтекстов и построение терминограмм. Преимущество «Humanitariana» — возможность объединения ресурсной базы нескольких организаций. Виртуальный информационно-ресурсный центр «Humanitariana» может функционировать в режиме локальной сети и в режиме распределенной информационной среды с возможностью обращения ко всем ресурсам с любого из серверов организаций-участниц [27]. Это позволяет привлекать к исследованию необходимое число участников — как исследователей, так и исследовательских организаций. В первую очередь организаций-владельцев крупных информационных массивов и тематических коллекций материалов в гуманитарной и социальной сферах (музеи, вузы, информационные центры, библиотеки).

«Humanitariana» включает в себя следующие функции: 1) модуль сбора и загрузки данных (текстовых массивов); 2) модуль, реализующий несколько видов тематического контекстного поиска и извлечения знаний: частотно-ориентированный поиск, абзацно-ориентированный поиск (простой («однослойный») контекстный запрос или расширенный («многослойный») контекстный запрос). 3) модуль создания тематических коллекций материалов на основе на основе кластеризации результатов запросов, анализа горизонтальных и вертикальных макро и микроконтекстов. Информационная система «Humanitariana» является постоянно развиваемой, поэтому может гибко подстраиваться под возникающие задачи и исследования.

Предполагается использовать ресурсы и сервисы созданного в рамках вышеназванного проекта виртуального информационно-сервисного центра для

изучения контекстного знания в гуманитарной сфере. Предлагаемый подход дополняется технологиями гибридного и мультимодального поиска — т.е. поиска одновременно по каталогу, мультимодальным (музейным) коллекциям и полным текстам в рамках функционально интегрированной информационной системы.

Таким образом, развивается методология, методика составления и технология осуществления тематизируемых полнотекстовых и мультимодальных запросов к информационным ресурсам для экспликации содержательных контекстов различного типа, вида и уровня. Информационные ресурсы размещены как в локальной сети Университета ИТМО, так и в распределенной среде проекта «Гуманитариана», в которой участвуют специализированные информационные ресурсы и сервера нескольких организаций сферы образования, науки и культуры (около 10000 документов на декабрь 2016 г.). За счет развития ресурсной базы предполагается в дальнейшем увеличение в распределенной среде базы текстами из различных областей гуманитарного знания: истории, философии, культурологии, социологии, экономики, лингвистики психологии, политологии, искусствоведении, религиоведении.

7. Заключение

Апробация разрабатываемого подхода позволит получить практические результаты за счёт пополнения информационно-ресурсной базы проекта «Контекстное знание», развития виртуального ресурсно-сервисного центра «Гуманитариана» (Университет ИТМО) и масштабирования последнего в распределенной среде. При этом к ожидаемым практически результатам можно отнести:

- разработка методики структурированного описания контекстного знания, получаемого в результате запросов в информационно-поисковой среде проекта;
- экспликация и описание массива единиц контекстного знания для их дальнейшей типологизации;
- развитие программной среды проекта: развитие подсистемы кластерного анализа в абзацно-ориентированном запросе; разработка 2-х новых видов полнотекстового запроса; разработка поиска по каталогу в распределенной среде участников проекта; гибридный поиск одновременно ко каталогу и полным текстам в локальной среде;
- корректировка и совершенствование методики структурированного описания контекстного знания, получаемого в результате запросов в информационно-поисковой среде проекта;
- разработка типологии контекстного знания (рабочий вариант) на основе эксплицированных контекстов;
- разработка методики организации НИР в вузе с использованием сервисов «Humanitariana», позволяет повысить управляемость и интенсифицировать проведение научных исследований, в первую очередь аналитического характера за счет создания собственных тематических хранилищ и возможности доступа к тематическим

хранилищам и коллекциям других организаций, сокращения трудоёмкости обработки информационных источников и ресурсов.

При этом сервисы развиваемой в рамках исследования информационной среды обеспечат использование методов полнотекстового и мультимодального поиска, анализа контекстного знания в научной и образовательной деятельности.

Литература

[1] Domingue J., Fensel D., Hendler J.A. Handbook of Semantic web Technologies. Heidelberg; Dordrecht; London; N.Y.: Springer, 2011. 1077 p.

[2] EventCube: multi-dimensional search and mining of structured and text data. In Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining (KDD '13)

[3] Koltsov S., Koltsova O., Nikolenko S. I. Latent Dirichlet Allocation: Stability and Applications to Studies of User-Generated content // Proceedings of WebSci '14 ACM Web Science Conference, Bloomington, IN, USA, June 23 - 26, 2014. NY : ACM, 2014. Pp. 161-165.

[4] Kononova, O., Liapin S. Using the Contextual Search for the Organization Scientific Research Activities // Communications in Computer and Information Science. 2016. Vol. 674. Pp. 392-399. DOI: 10.1007/978-3-319-49700-6_38

[5] Melucci M. Contextual Search: A Computational Framework // Foundations and Trends in Information Retrieval: 2012. Vol. 6. No. 4–5. P. 257-405. DOI: 10.1561/1500000023

[6] Saifa H., Heb Y., Fernandeza M., Alani H. Contextual semantics for sentiment analysis of Twitter. Information Processing & Management, Volume 52, Issue 1, 2016. Pages 5–19. DOI: 10.1016/j.ipm.2015.01.005

[7] Smart P.R., Sieck W.R., Shadbolt N.R. Using Web-Based Knowledge Extraction Techniques to Support Cultural Modeling // Proceedings of 4th International Conference, SBP 2011, MD, USA, March 29-31, 2011. Springer, London – New York. 2011.Pp.113–120. DOI 10.1007/978-3-642-19656-0

[8] Szetela D. Effective Contextual Search Management. URL: https://www.seroundtable.com/archives/018064.html (дата обращения: 12.03.2017)

[9] Tao F., Lei K.H., Han J., Zhai C., Cheng X., Danilevsky M., Desai N., Ding B., Ge Ge J., Ji H., Kanade R., Kao A., Li Q., Li Y., Lin C., Liu J., Oza N., Srivastava A., Tjoelker R., Wang C., Zhang D., Zhao B. EventCube: multi-dimensional search and mining of structured and text data // Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining (KDD '13). ACM, New York, NY, USA. 2013. Pp. 1494-1497. DOI: 10.1145/2487575.2487718

[10] Turney P.D., Pantel P. et al. From frequency to meaning: Vector space models_of semantics // Journal of Artificial Intelligence Research. 2010. 37 (1). Pp. 141–188

[11] Агеев М.С., Добров Б.В., Журавлев С.В., Лукашевич Н.В., Сидоров А.В., Юдина Т.Н. Технологические аспекты организации доступа к разнородным информационным ресурсам в университетской информационной системе Россия // Электронные библиотеки. 2002. Том 5. Выпуск 2. С. 1-13.

[12] Аюшеева Н.Н. Исследование и разработка моделей и методов поиска информационных образовательных ресурсов в электронной библиотеке: автореф. дис. ... канд.техн.наук: 05.13.11. – Улан-Удэ, 2004. 228 с.
[13] Бородкин Л.И. Приоритеты современной исторической информатики: технологии e-Science // Круг идей: междисциплинарные подходы в исторической информатике. Труды X конференции Ассоциации "История и компьютер" / Под редакцией Л.И.Бородкина, И.М.Гарсковой. М., 2008. С.5-15.
[14] Вербицкий А.А. Контекстное обучение в компетентностном формате (Компетентностный подход как новая образовательная парадигма) // Проблемы социально-экономического развития Сибири. 2011. № 4 (6). С. 67-73. URL: http://brstu.ru/static/unit/journal_2/docs/number6/67-73.pdf (дата обращения: 12.03.2017).
[15] Воронцов К.В. Аддитивная регуляризация тематических моделей коллекций текстовых документов // Доклады РАН. 2014. Т. 455. №3. С. 268–271.
[16] Воронцов К.В. Потапенко А.А. Регуляризация вероятностных тематических моделей для повышения интерпретируемости и определения числа тем // Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной Международной конференции «Диалог» (Бекасово, 4–8 июня 2014 г.) 2014. Вып.13 (20). С.676–687.
[17] Демин И.С. Поиск научной и учебной информации в сети Интернет // Вестник Тамбовского университета. Серия: Гуманитарные науки. 2008. Вып. 9. С. 446–450.
[18] Ермаков А. Е. Эксплицирование элементов смысла текста средствами синтаксического анализа-синтеза // Компьютерная лингвистика и интеллектуальные технологии: труды Международной конференции Диалог’2003. М.: Наука, 2003.
[19] Жижимов О.Л., Молородов Ю.И., Пестунов И.А., Смирнов В.В., Федотов А.М. Интеграция разнородных данных в задачах исследования природных экосистем // Вестник НГУ. Сер.: Информационные технологии. 2011. Т.9. № 1. С. 67-74.
[20] Жукова Н.В. Изменение культуры и новая образовательная парадигма // Вестник Уральского института экономики, управления и права. 2012. №3 (20). С.60-65.
[21] Карпенко О., Шпаковская Л. Л., Кольцова Е. Ю., Торчинский Ф. И., Дубровский Д. В. Язык вражды в русскоязычном Интернете: Материалы исследования по опознаванию текстов ненависти. СПб.: изд-во Европейского университета в Санкт-Петербурге, 2003
[22] Кольцова О.Ю. Чем дышит блогосфера? К методологии анализа больших текстовых данных для социологических задач / Онлайн исследования в России 3.0 / Под редакцией Шашкина А.В., Девятко И.Ф., Давыдова С.Г. М.: Издательский дом «Кодекс», 2012. С. 163-186.
[23] Кольцова О.Ю., Маслинский К.А. Выявление тематической структуры российской блогосферы: автоматические методы анализа текстов // Социология 4М. 2013. № 36. С. 113-139.

[24] Кононова, О.В., Крутько Е.А., Ляпин С.Х. Технологии извлечения знаний на службе научно-исследовательской деятельности в вузе // Информационное общество. 2016. № 6. С. 25-37.
[25] Куршев Е. П., Осипов Г. С., Рябков О. В., Самбу Е. И., Соловьева Н. В., Трофимов И. В. Интеллектуальная метапоисковая система // Труды международного семинара Диалог'2002. Компьютерная лингвистика и интеллектуальные технологии. М.: Наука, 2002. С. 320–330.
[26] 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, 22-24 июня 2016 г. С. 70-86.
[27] Ляпин С.Х, Куковякин А.В. Тематические коллекции полнотекстовых запросов для изучения контекстного знания (проект Humanitariana) // Сборник научных статей XVIII Объединенной конференции «Интернет и современное общество» IMS-2015, Санкт-Петербург, 23-25 июня 2015 г., Университет ИТМО, Санкт-Петербург, 2015. С. 216 – 224.
[28] Ляпин С.Х., Куковякин А.В. Кластеры полнотекстового поиска в распределенной информационной среде: технология и проекты // Материалы XXII Межд. конференции "Крым 2015" (эл. версия), ГПНТБ, Москва, 2015. URL: http://www.gpntb.ru/win/Inter-Events/crimea2015/disk/003.pdf (дата обращения: 26.01.2017.)
[29] 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/abrau/2016/9.pdf (дата обращения: 12.03.2017).
[30] Ляпин С.Х., Куковякин А.В. Обработка неструктурированных данных (извлечение контекстного знания) с помощью сервисов полнотекстового поиска в электронной библиотеке // Труды XVII Международной конференции DAMDID/RCDL' 2015 «Аналитика и управление данными в областях с интенсивным использованием данных», Обнинск, 13-16 октября 2015, Обнинск, 2015. С.164 – 167.
[31] Манаев О.Т. Контент-анализ как метод исследования // Социология: энциклопедия. М., 2003. URL: http://psyfactor.org/lib/content-analysis3.htm (дата обращения: 12.03.2017).
[32] Осипов Г. С., Куршев Е. П., Кормалев Д. А., Трофимов И. В., Рябков О. В., Тихомиров И. А. Семантический поиск в среде интернет. Переславль-Залесский, ИПС РАН, 2003.
[33] Осипов Г. С., Тихомиров И. А., Смирнов И. В. Интеллектуальный поиск в глобальных и локальных вычислительных сетях и базах данных // Труды международной конференции. Программные системы: теория и приложения., ИПС РАН, г. Переславль-Залесский, май 2004 / Под редакцией С. М. Абрамова. В двух томах. М.: Физматлит, 2004. Т. 1. С. 21-23.

[34] Терещенко Е. А., Ефимова Т. Г. Обнаружение скрытой тематической структуры в блогах: на примере Живого Журнала // Избранные тезисы докладов IV Студенческой социологической межвузовской конференции / Отв. ред.: М. Р. Демин. СПб.: НИУ ВШЭ (Санкт-Петербург), 2013.
[35] Черний А.В., Тузовский А.Ф. Развитие информационной системы организации с использованием семантических технологий // Знания–Онтологии–Теория: Матер. Всеросс. конф. с междунар. участием. Новосибирск, 20–22 октября 2009. Новосибирск: ЗАО «РИЦ Прайс-Курьер», 2009. Т. 2. С. 52–59.
[36] Шуклин А. Кому нужен контекстный поиск? URL: http://www.cnews.ru/reviews/free/BI2012/articles/articles2.shtml (дата обращения: 12.03.2017)

Synthetic Method of Contextual Knowledge Extraction in the Russian socio-humanitarian sphere: an integrated approach

O.V. Kononova [1], S.Kh. Lyapin [1], D.E. Prokudin [2, 1]
[1] ITMO University, [2] Saint-Petersburg State University

The authors propose a synthetic method for the extraction of contextual knowledge as an integrated approach that focuses on the problem of explication of context knowledge in Russian texts of Humanities and social orientation and subsequent typology of contextual knowledge in individual disciplines and interdisciplinary relations, as well as on General scientific and General cultural term-concepts, combining different disciplines.

The problem has scientific and cultural significance, as contextual knowledge is an important part of the explicit and implicit knowledge that a person operates in the scientific, educational and cultural activities.

The proposed approach to solving the problem of the explication of contextual knowledge is empowering technology thematizing flexible full-text and multi-modal search, implemented and tested in the framework of the virtual resource and service center, established in 2014-2016 on the basis of the ITMO University in the framework of the project RFH № 14-03-12017 (project "Humanitariana").

Keywords: synthetic method, full-text search, multimodal search, context knowledge, micro context, macro context, terminogramme, semantic mapping, search thesaurus