УДК: 004.89

Название: Подход к созданию исследовательской информационной системы с документально подтверждаемой информацией

Автор(ы):

Сидорова Е.А. (Институт систем информатики СО РАН),

Серый А.С. (Институт систем информатики СО РАН)

Аннотация: Данная статья посвящена проблемам интеллектуального доступа к информационным ресурсам. Приводится анализ информационных систем, поддерживающих работу с данными, либо представленными текстовыми документами и их фрагментами, либо заданными объектной моделью на основе онтологии предметной области. Авторами предложен подход к созданию системы, которая бы поддерживала оба вышеперечисленных способа представления информации, описана ее архитектура и схема хранилища данных.

Ключевые слова: информационная система, интеллектуальный доступ к данным, онтология, текст, аннотирование.

1. Введение. Современные информационные системы предоставляют пользователю информацию в виде «готового знания» — набора определенных фактов о действительности [5]. Однако в некоторых сферах деятельности, таких как юриспруденция, делопроизводство, научные исследования и т.д., требуется документальное подтверждение любой информации или факта. Это означает, что любой факт должен сопровождаться ссылкой на источник. Таким источником может быть как документ, так и отдельно взятый фрагмент документа, подтверждающий данный факт. Важной функцией в вышеупомянутых сферах деятельности становится также и проверка данных при поступлении в базу данных, контроль достоверности информации, ее актуальности во времени.

Работа с документами и их текстовыми фрагментами поддерживается системами класса корпус-менеджер [1,6], которые обеспечивают аннотирование или разметку текста, комплексную возможность просмотра контекстов в виде конкордансов, поиск в корпусе текстов, фильтрацию по различным основаниям, сбор статистики и т.п. Основное назначение таких систем — проведение комплексного исследования на обширном материале, представленном в корпусе документов.

На сегодняшний день существует множество различных систем для работы с текстами и корпусами. Одни из них предоставляют инструменты для разметки текста и поиска по корпусам, но, в то же время, не позволяя, например, выявить в тексте референциальные отношения. Другие создаются для извлечения информации и выделения терминов и кореферентных выражений, являясь узкоспециализированными в фиксированных

предметных областях. В качестве примера первых можно привести системы **Bonito** и **Xaira** (http://projects.oucs.ox.ac.uk/xaira). Первая представляет собой графический пользовательский интерфейс для корпуса-менеджера Manatee, основным назначением которого является обработка различных поисковых запросов и визуализация результатов поиска. Вторая предназначена для поиска в тексте на основе XML-разметки и не зависит от языка, т.к. в ее основе лежит объектная модель, позволяющая описывать сущности, а также задавать методы представления и поиска различных лингвистических ресурсов.

Представителями систем, реализующих другой подход, являются, к примеру, создаваемые с целью разработки и оценки методов извлечения информации и data mining аннотированные корпусы текстов (преимущественно англоязычных) по биомедицине и молекулярной биологии **GENIA** (http://www-tsujii.is.s.u-tokyo.ac.jp/GENIA/) и **BioInfer** (http://www.it.utu.fi/BioInfer). **GENIA** — это информационный ресурс на базе аннотированного биомедицинского корпуса, в котором выделены термины, отношения, ситуации, кореферентные выражения, а **BioInfer** — открытый ресурс для ручной разметки корпуса и связанных ресурсов и для извлечения информации в биомедицинской предметной области; элементами разметки здесь являются термины и связи между ними.

Данные системы эффективны при решении конкретных задач в специализированных (фиксированных) предметных областях, однако сфера их применения не может быть обобщена или расширена. Возможность многослойной лингвистической разметки корпусов текстов, а также визуализация разметки и средства фильтрации и поиска реализованы в системе UAM CorpusTool [13]. Она позволяет задавать проекты в виде набора файлов, к которому может быть применена единая система признаков разметки. Каждому признаку соответствует слой текстовой разметки. UAM CorpusTool позволяет аннотировать фрагменты текста без ограничения на вложенность, пересечение и разрывность. Система позволяет анализировать, фильтровать размеченные фрагменты корпуса и осуществлять поиск по размеченному корпусу. Разметка представлена в формате XML, а также хранится отдельно от текста в виде аннотаций [2,9]. Несмотря на то, что система находится в открытом доступе, она работает только с английскими словарями, что делает невозможным лексический анализ текстов на русском языке.

Наша работа, следуя курсу наиболее актуальных направлений исследований в области обеспечения интеллектуального доступа к информационным ресурсам, направлена на создание системы, которая бы совмещала стандартные способы представления информации на основе модели предметной области и представления этой же информации в виде совокупности текстовых фрагментов из различных документальных источников, в которых

она упоминается. Система будет снабжена инструментами как для информационного поиска, основанного на фактографическом представлении информации, так и для проведения исследований и анализа имеющейся информации. Достоверность информации, представленной в системе, может быть подтверждена ссылками на текстовые источники, в которых она упоминается.

Работа выполняется при финансовой поддержке Российского фонда фундаментальных исследований (грант №12-07-31216).

2. Общее описание подхода. Разрабатываемая система содержит информацию, представленную двумя типами данных: тексты на естественном языке и набор формально описанных фактов, упоминаемых в этих текстах или введенных пользователем вручную (или иным способом). При этом связи между фактом и текстами, являющимися его источниками, должны сохраняться. В этом и состоит идея «документального подтверждения данных», когда для некоторого факта можно подобрать множество текстов и указать конкретные цитаты, где упоминается данный факт. В сравнении с информационно-поисковыми системами, осуществляющими фактографическое индексирование текстов, предлагаемый подход имеет существенное отличие: один и тот же факт может быть представлен в различных документах. Можно рассматривать связь факта и текстов как ссылки на первоисточники. В качестве примера ситуации, в которой необходима ссылка на источник, можно привести наличие факта назначения срока наказания за некое противоправное деяние, подтверждением которого будет цитирование соответствующей статьи уголовного кодекса.

Фактически контент системы имеет двойственную природу и требует разработки методов представления, хранения и анализа знаний в едином информационном пространстве. В рамках предлагаемого подхода представление структурированных ресурсов базируется на описании предметной области в виде онтологии, определяющей основные понятия и отношения рассматриваемой области знаний. Информация, содержащаяся в интегрируемых текстовых ресурсах — документах и корпусах — порождает и дополняет объектную структуру представления данных.

При наличии возможности оценить достоверность того или иного факта на основе анализа его источников, а также возможности удобного и содержательного доступа к данным, такая информационная система будет полезна как аналитикам, так и лицам, принимающим решения. Включение в систему развитых средств анализа и визуализации информации из различных корпусов документов в виде конкорданса, т.е. совокупности контекстов фактов, окажется полезно лингвистам для проведения исследований семантических свойств текста, а также инженерам знаний, которые получат инструментарий для автоматизированного

создания лингвистических ресурсов и систем анализа текстов, в том числе для автоматического наполнения информационных систем. Для долговременного функционирования и развития такой системы необходимо обеспечить поддержание логической целостности и достоверности информации. Целостность контента обеспечивается онтологией, дающей полное и целостное описание предметное области, а также организацией хранилища данных и методов доступа к нему. Поддержание информации в актуальном состоянии обеспечивает специально разработанный механизм, на основе анализа поступающих текстовых данных и их экспертной оценки. Устаревшие и более не считающиеся достоверными факты автоматически исключаются из системы.

Создание единой инструментальной среды, включающей инструменты как хранения и поиска информации, так и поддержки исследований на основе текстовых материалов, позволяет говорить о создаваемом программном продукте как об *исследовательской информационной системе*.

3. Аннотация документа. Как было сказано выше, контент системы неоднороден. С одной стороны он описывается онтологией предметной области, с другой — формируется множеством семантически аннотированных текстовых документов (пример такой семантической аннотации можно посмотреть в [12]). Данные представлены в виде информационных объектов со своими атрибутами, связями и текстовыми источниками — фрагментами документов, в которых упоминаются объекты. Хранение этих двух по сути различных видов контента организовывается по-разному. Оригиналы документов (представляющие собой файлы определенного формата) хранятся в специальной директории файловой системы сервера. Тем не менее, важно, чтобы документы и информационные объекты были связаны друг с другом в единой базе данных ИИС. Структурой, организующей связывание текстовых фрагментов и информационных объектов, является аннотация.

3.1. Требования к хранилищу данных. Основные требования к базе данных ИИС формулируются нами следующим образом:

- база данных должна совмещать хранение онтологии предметной области, лингвистической онтологии (системы дополнительных лингвистических признаков для разметки текста) и контента в виде информационных объектов и аннотированных документов;
- база данных должна обеспечивать хранение настроек визуализации как для отображения информационных объектов, так и для раскраски информации в тексте;
- в базе данных должны быть так или иначе представлены документы, имеющиеся в системе, причем тексты документов должны входить в данное представление (текст

дополняется ссылкой на оригинал, находящийся в файловой системе), а документы — снабжаться метаинформацией, аналогично представлению информационного объекта;

- в базе данных должны быть представлены аннотации, обеспечивающие связывание фрагментов текста с информационными объектами;
- необходимо обеспечить хранение корпусов документов; каждый корпус может иметь свою собственную направленность (по теме и/или жанру) и индивидуальный набор лингвистических признаков для лингвистического аннотирования; корпуса создаются и поддерживаются пользователями-экспертами, обладающими соответствующими правами; права могут раздаваться индивидуально для каждого корпуса;
- хранение данных о зарегистрированных пользователях: логины, пароли, права доступа и права редактирования.

3.2. Хранилище аннотаций. База данных состоит из пяти концептуальных блоков: онтологии, блока данных (информационных объектов), лингвистического блока, блока аннотаций и блока пользовательских настроек. На Рис. 1 изображен фрагмент схемы БД, подробно описывающий блок аннотаций.

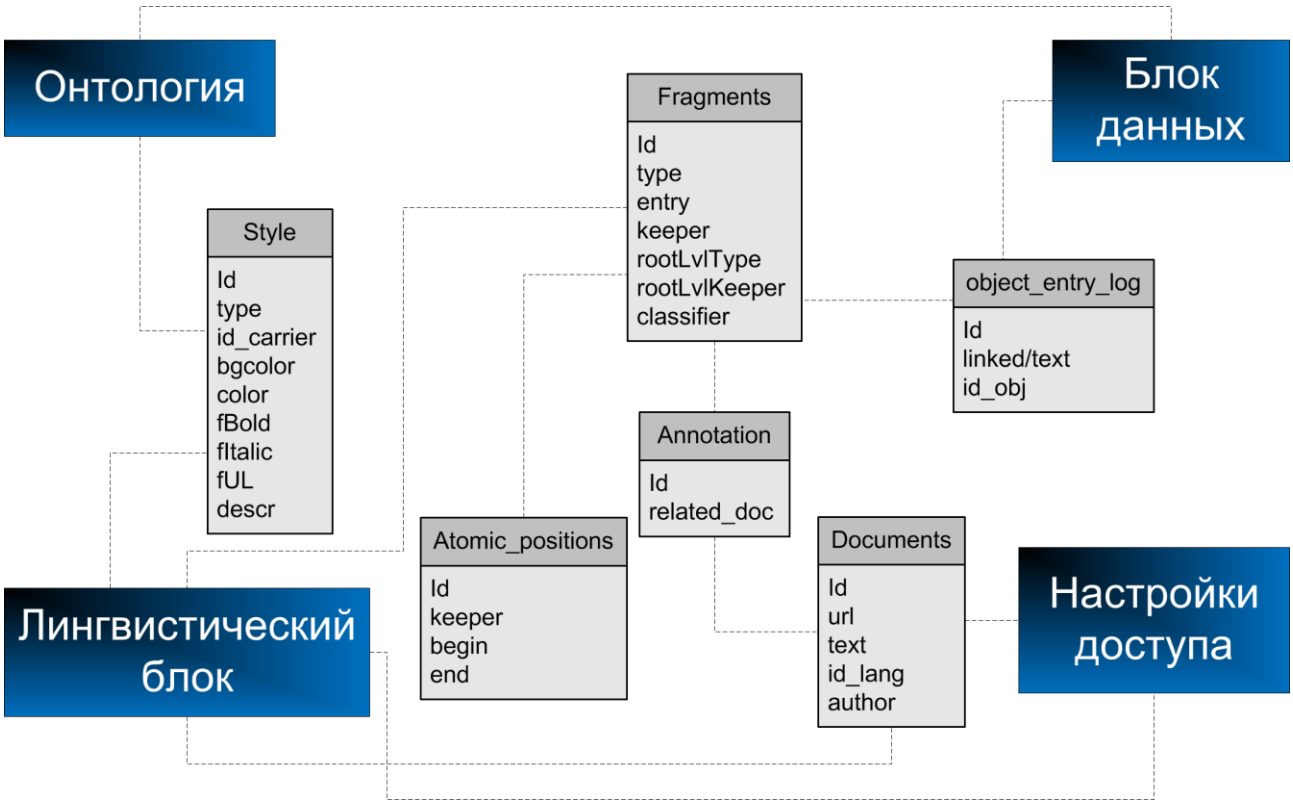

Рис. 1. Блок аннотаций базы данных системы

Рассмотрим таблицы, входящие в блок аннотаций.

3.3. Таблица документов. Таблица *Documents* содержит информацию об имеющихся документах, их тексты и ссылки на оригиналы документов, находящиеся в репозитории. Ниже перечислены столбцы данной таблицы.

id — уникальный идентификатор (ключ) документа,

url — ссылка на оригинал документа,

text — текст документа,

id_lang — язык документа,

author — автор документа (идентификатор пользователя системы, от имени которого был добавлен документ).

3.4. Таблица аннотаций. Таблица *Annotation* необходима для сведения информации обо всех размеченных фрагментах всех имеющихся документов. Она состоит всего из двух столбцов.

id — уникальный идентификатор (ключ),

related_doc — идентификатор документа, аннотированного данной аннотацией.

Аннотация документа может быть многоуровневой, т.е. содержать как лингвистическую разметку, в случае чего она будет связана с лингвистическим блоком, так и семантическую разметку информационных объектов, связь с которыми осуществляется через вхождение объекта в текст.

3.5. Таблица вхождений объектов. Таблица *object_entry_log* описывает все вхождения информационных объектов в текст документов.

id — уникальный идентификатор (ключ),

linked/text — идентификатор документа, содержащего данное вхождение,

id_obj — идентификатор объекта, участвующий в данном вхождении.

Вхождение объекта соответствует каждому упоминанию объекта в тексте, соответственно, в каждом тексте может быть несколько вхождений одного и того же объекта. Одно вхождение объекта может включать несколько текстовых фрагментов, которые соответствуют разным атрибутам объекта.

3.6. Таблица фрагментов. Таблица *Fragments* — таблица размеченных фрагментов. Фрагмент — это либо ссылка на ранее размеченный фрагмент, но уже с новыми свойствами, либо множество атомарных позиций, задающих текстовый интервал. Соответственно, выделяется непосредственный владелец фрагмента, т.е. признак или атрибут объекта, сопоставляемый с данным текстовым фрагментом, и владелец нулевого уровня, содержащий данный фрагмент не как ссылку, а как набор атомарных позиций (см. определение фрагмента). Ниже приведен список столбцов данной таблицы.

id — уникальный идентификатор (ключ),

type— тип непосредственного владельца данного фрагмента,

entry — вхождение (аннотация) объекта, в котором задействован данный фрагмент,

keeper — непосредственный владелец фрагмента,

rootLvlType — тип владельца нулевого уровня,

rootLvlKeeper — владелец нулевого уровня,

classifier — принимает значения *true* или *false* в зависимости от того, является ли данный фрагмент обозначением классификатора — типа сущности или непосредственным ее значением.

Параметры **type** и **rootLvlType** определяют таблицы, в которых задаются элементы лингвистической или предметной онтологии, а **keeper** и **rootLvlKeeper** — индексы строк в этих таблицах. Параметр entry является ссылкой на элемент таблицы *object_entry_log* и может быть не определен, если владельцем фрагмента не является объект. Данное поле определяет свойства объекта, упомянутые в конкретном вхождении, если эти свойства были аннотированы.

3.7. Таблица атомарных позиций. Таблица *Atomic_positions* содержит атомарные позиции или атомы, т.е. неразрывные текстовые фрагменты. Для атомарной позиции можно однозначно определить координаты начала и конца в тексте. Более сложные фрагменты состоят из множества атомарных позиций. Таблица включает четыре столбца.

id— уникальный идентификатор (ключ),

keeper — фрагмент, содержащий данную позицию,

begin — начало атома,

end — конец атома.

3.8. Таблица стилей. Таблица *Style* — это таблица стилей, используемых для визуализации разметки текста. Стиль сопоставляется либо элементам онтологии предметной области, либо лингвистическим понятиям.

id — уникальный идентификатор (ключ),

type — тип элемента, которому соответствует данный стиль,

id_carrier — ссылка на этот элемент,

bgcolor — цвет фона текста,

color — цвет текста,

fBold — полужирный текст,

fItalic — курсив,

fUL — подчеркнутый текст,

descr — описание стиля.

Как видно, значения кортежей таблицы задают различные стили форматирования текста в зависимости от типа размечаемого элемента. Параметр **type** определяет таблицу, в которой задается элемент онтологии, а **id_carrier** — индексы строк в этой таблице.

4. Архитектура системы. Для доступа к контенту предполагается разработка нескольких видов интерфейса: программного, пользовательского и интерфейса редактора. Программный интерфейс (API) позволит использовать содержимое системы другими программами и приложениями для решения задач информационного поиска, обработки текстов документов, построения лингвистических ресурсов и др. Пользовательский интерфейс обеспечит доступ к содержимому системы рядовым пользователям и предоставит удобные средства поиска фактов и их контекста в документах, обеспечит навигацию по системе признаков и в корпусе документов. Интерфейс редактора предоставит удобные средства добавления информации и разметки текстов, формирования корпуса, согласованного с общим содержанием системы, создания системы лингвистических признаков. На рис. 2 показана общая архитектура системы.

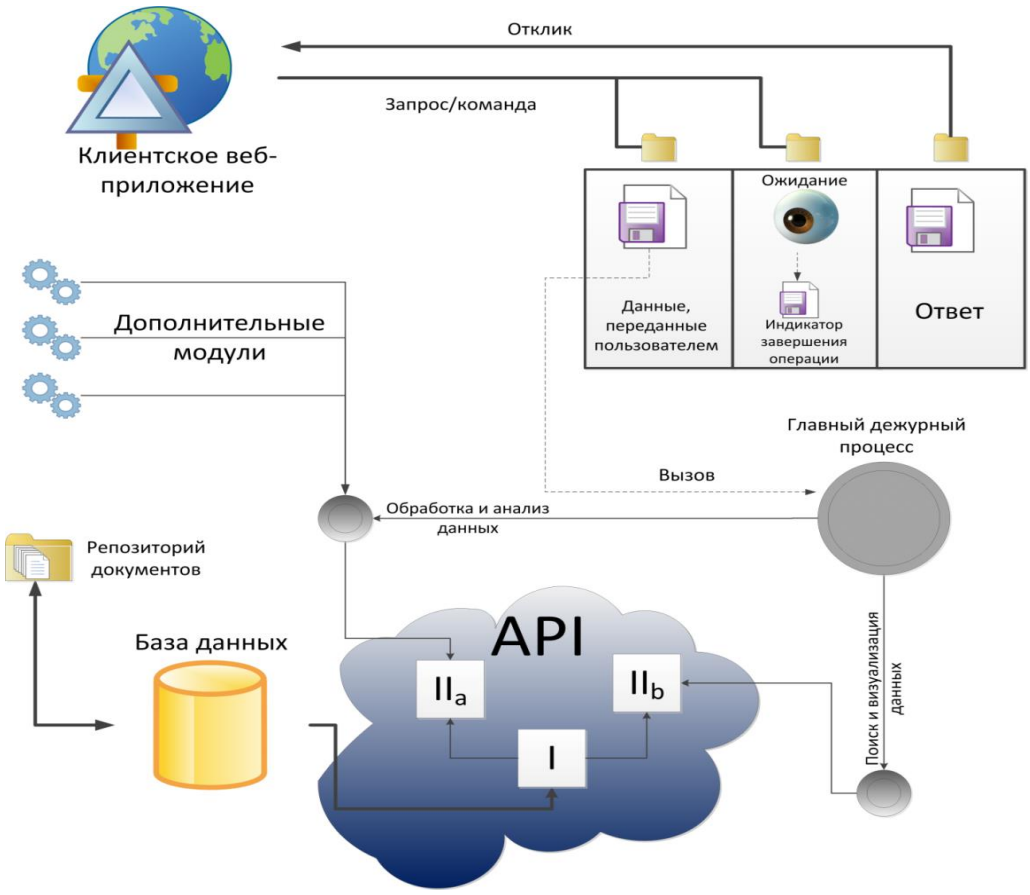

Рис. 2. Архитектура системы

Фактически работа с хранилищем данных в системе включает три контура (API). I — уровень ядра, инкапсулирует операции генерации SQL-запросов и обмена информацией с базой данных, реализует оптимизационные алгоритмы и механизмы обеспечения корректности и целостности данных. II_a — аналитический уровень, предоставляет инструменты для поиска данных и редактирования БД в терминах онтологии. API данного уровня оперирует понятиями онтологии, информационными объектами, документами и их аннотациями. II_b — уровень визуализации, снабжает пользовательский интерфейс информацией, необходимой для визуализации объектов аналитического уровня — информационных объектов и аннотаций. API данного уровня расширяет объекты уровня II_a набором методов для настройки визуализации. Как можно видеть на рис. 2, конечный пользователь системы работает с системой через веб-приложение. Для исполнения пользовательских запросов необходимо дополнительное звено. Таким звеном должен стать дежурный процесс, следящий за поступлениями новых запросов и обеспечивающий обмен информацией между веб-приложением и базой данных, не нарушая при этом принятой концепции. Для обмена информацией было решено использовать язык JSON (Java Script Object Notation). Единичную операцию поискового обращения к БД можно составить из следующих пунктов:

- пользователь средствами веб-интерфейса формулирует запрос и отправляет его на сервер;
- сервер, получив запрос, присваивает ему код, конвертирует его в файл формата JSON и дает сигнал об этом дежурному процессу;
- дежурный процесс извлекает из файла запроса всю необходимую информацию и вызывает соответствующие функции API базы данных;
- получив результат, дежурный процесс формирует файл-ответ в формате JSON, помечая его тем же кодом, которым был помечен запрос;
- веб-сервер после получения запроса и передачи его в дежурный процесс проверяет наличие ответа на него в течение заданного времени с заданной периодичностью; в случае отсутствия ответа по истечению срока пользователю отправляется сообщение вида not responded;
- если во время очередной проверки веб-сервер обнаруживает ответ на присланный запрос, он забирает его и отправляет пользователю.

Кроме того, в системе функционируют инструментальные модули, реализующие средства разметки текстов, проверки поступающего потока данных, контроля достоверности и пр.; эти средства, разрабатываемые по отдельности, интегрируются в единую среду.

5. Контроль целостности контента. Задача интеграции структурной и текстовой информации предполагает не только ее сбор с помощью ручной разметки текста экспертом или автоматического аннотирования документов какой-либо системой анализа, но и разработку методов и средств контроля поступающей информации, отслеживания ее достоверности и актуальности. Контроль достоверности данных предполагает слежение, в автоматическом или автоматизированном режиме, за поступающей информацией и ее оценку на основании упоминающих ее источников-документов. Соответственно, чем большее число источников упоминает некоторый факт, и чем выше авторитет этих источников, тем больше доверия вызывает данный факт. Даже если на момент поступления в систему информация имела статус заслуживающей доверия, со временем она может устареть, потерять актуальность и даже стать ложной. Косвенным признаком утери актуальности факта может послужить длительное отсутствие его упоминаний в новых документах. Накапливаясь в системе, подобные неактуальные данные занимают компьютерные ресурсы, вносят беспорядок и затрудняют поиск. Одним из видов средств, облегчающих работу с базой данных и обеспечивающих рациональное использование компьютерных ресурсов, являются разработанные оригинальные методы корректного пополнения базы данных фактами, полученными из текста. Основная идея подхода заключается в трехуровневой проверке информации, поступающей в базу данных. Рассмотрим каждый из этих уровней более подробно.

1. Установление референциальных связей между информационными объектами. Данный метод базируется на существующих методах разрешения анафоры в документах на русском и английском языках [4, 8, 11], а также на методах сравнения информационных объектов, предлагаемых в [3]. Он включает в себя установление степени сходства объектов, построение множества гипотетических эквивалентов для каждого объекта и объединение объектов, признанных кореферентными. Данный уровень подробно описан в [8]. Основной целью поиска референциально тождественных объектов является сокращение числа ИО, представляющих одну сущность, что, в свою очередь, повышает вероятность их успешной идентификации.
2. Идентификация информационных объектов — разрешения контекстной омонимии, являющейся одним из побочных эффектов обработки текста. Контекстная омонимия проявляется в наличии двух и более вариантов отождествления полученных из текста информационных объектов с объектами базы данных информационной системы. Ключевым для метода идентификации данных является понятие фокусного множества. Фокусное множество включает все экземпляры отношений, с помощью которых текущий

объект непосредственно связан с другими входными объектами. При этом множество отношений разбивается на подмножества связей с идентифицированными и требующими идентификации объектами. Основой метода является сопоставление фокусных множеств найденных в тексте объектов с фокусными множествами объектов, уже содержащихся в базе данных информационной системы.

3. Разрешение противоречий между содержащимися в базе данных и вновь поступившими фактами посредством вычисления специального параметра, количественно выражающего достоверность того или иного атрибута или связи. Фактически данный пункт заключает в себя слежение, как за достоверностью поступающих данных, так и за актуальностью уже имеющихся. Для контроля данных были разработаны подходы, включающие элементы теории вероятности. Были введены вероятностные характеристики документов и содержащихся в них фактов, а в основу модели жизненного цикла факта в информационной системе положен неоднородный марковский случайный процесс.

6. Заключение. Основной отличительной чертой данной работы является интеграция различных методов и подходов, существующих независимо, под одной исследовательской информационной оболочкой. В рамках данной системы знания извлекаются из текстов и визуализируются в виде базирующихся на онтологии структур (объектов, отношений). Процесс добавления новой информации может осуществляться пользователем либо через специализированный редактор, наследуемый от редактора данных из порталов знаний [5], либо при аннотировании текстов через подсистему семантической разметки корпуса текстов.

Технология разметки корпуса текстов в терминах объектов и связей предметной области была апробирована при создании корпуса текстов по катализу [7]. В рамках этого проекта были разработаны методы и средства семантического аннотирования двух типов: терминологическая разметка, которая предназначалась для фиксации в текстах имен понятий предметной области (терминологическая разметка была использована для создания предметного словаря по катализу), и разметка ситуаций (химических реакций), представляющих собой многоместные отношения, в которых размеченные сущности выступают в определенных семантических ролях.

Специфической особенностью исследовательской информационной системы является постоянное накопление документальных ресурсов и информации, корректность и достоверность которых необходимо постоянно отслеживать. С этой точки зрения потребуются дополнительные исследования ситуаций добавления разными экспертами противоречивой информации. Также в дальнейшем планируется подключать модули

автоматического анализа текста, извлечение информации в которых будет опираться на экспертные знания, выраженные с помощью средств аннотирования документов.

Список литературы

1. Апресян Ю. Д., Богуславский И. М. и др. Синтаксически и семантически аннотированный корпус русского языка: современное состояние и перспективы // Национальный корпус русского языка: 2003–2005. М.:Индрик, 2005. С.193–214.
2. Биряльцев Е.В., Елизаров А.М., Жильцов Н.Г., Иванов В.В., Невзорова О.А., Соловьев В.Д. Модель семантического поиска в коллекциях математических документов на основе онтологий // Труды 12-й Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» – RCDL'2010. Казань, 2010. С. 296–300.
3. Васильев И.А. Оценка семантической близости объектов с использованием дескриптивной логики // Материалы 5-й научно-практической конференции «Современные средства и системы автоматизации». Томск: ТУСУР, 2004. С. 160–163.
4. Ермаков А.Е. Референция обозначений персон и организаций в русскоязычных текстах СМИ: эмпирические закономерности для компьютерного анализа. // Компьютерная лингвистика и интеллектуальные технологии: Труды международной конференции «Диалог 2005». М.: Наука, 2005. С. 131–135.
5. Загорулько Ю.А., Боровикова, О.И. Подход к построению порталов научных знаний // Автометрия. 2008 № 1, Т. 44, С. 100–110.
6. Захаров В.П., Богданова С.Ю. Корпусная лингвистика // Учебник для студентов гуманитарных вузов. Иркутск: ИГЛУ, 2011. 161 с.
7. Кононенко И.С., Сидорова Е.А. Система семантической разметки корпуса текстов как инструмент извлечения экспертных знаний (на материале текстов по катализу) // Труды международной конференции «Корпусная лингвистика – 2011». СПб, 2011. С. 193–198.
8. Серый А.С., Сидорова Е.А. Поиск референциальных отношений между информационными объектами в процессе автоматического анализа документов // Труды XIV Всероссийской научной конференции RCDL-2012 Электронные библиотеки: перспективные методы и технологии, электронные коллекции. – Переславль-Залесский, 2012. С.206–212
9. Blanco X. Using Noo J for Multipurpose Analysis of Romance Languages Corpora // Труды межд. конф. «Корпусная лингвистика–2008». СПб., 2008. С.40–44.
10. Caroline V. Gasperin Statistical anaphora resolution in biomedical texts. // In Proc. of the 22nd International Conference on Computational Linguistics. Manchester. UK. 2008. P. 257–264.

11. Heeyoung Lee, Yves Peirsman, Angel Chang, Nathanael Chambers, Mihai Surdeanu, and Dan Jurafsky. 2011. Stanford's Multi-Pass Sieve Coreference Resolution System at the CoNLL-2011 Shared Task. // In Proc. of the 15th Conference on Computational Natural Language Learning: Shared Task. Portland. Oregon. USA. 2011. P.28–34.
12. Kim J.D., Ohta T., Tsujii J. Corpus annotation for mining biomedical events from literature // BMC Bioinformatics. 2008. 9:10.
13. O'Donnell M. The UAM CorpusTool: Software for corpus annotation and exploration // In Bretones Callejas, Carmen M. et al. (eds) Applied Linguistics Now: Understanding Language and Mind / La Lingüística Aplicada Hoy: Comprendiendo el Lenguaje y la Mente. Almería: Universidad de Almería. 2008. p. 1433–1447.

UDK: 004.89

Title: An approach to designing an information system with documented information.

Author(s):

Sidorova E.A. (A.P. Ershov Institute of Informatics Systems),

Seryj A.S. (A.P. Ershov Institute of Informatics Systems)

Abstract: The authors discuss an intelligent access to information resources. An analysis of information systems which support textual or ontology-based data representation is provided. The authors propose an approach to develop an information system that would support both of the aforementioned ways to represent knowledge. A possible architecture and database schema for such a system are described.

Keywords: information system, intelligent access to data, ontology, text annotation.