

УДК 378; 004.89

Гибридная интеллектуальная информационная система управления знаниями о науке и образовании

Андреев А.М.¹, Березкин Д.В.¹,
Козлов И.А.^{1,*}

*kozlovilya89@yandex.ru

¹МГТУ им. Н.Э. Баумана, Москва, Россия

В рамках международного научного конгресса "Наука и инженерное образование. SEE-2016", II международная научно-методическая конференция «Управление качеством инженерного образования. Возможности вузов и потребности промышленности» (23-25 июня 2016 г., МГТУ им. Н.Э. Баумана, Москва, Россия).

В работе рассмотрена задача управления разнородными данными и знаниями о науке и образовании. Предложена архитектура гибридной информационной системы управления знаниями, содержащей элементы искусственного интеллекта. Предлагаемая система способна сочетать различные модели данных, знаний и методы для их анализа. Рассмотрены особенности реализации отдельных подсистем предлагаемой системы. Приведены примеры использования системы для решения отдельных задач управления ядерными знаниями.

Ключевые слова: управление знаниями, гибридная интеллектуальная система

Введение

Для современных информационных систем (ИС), выполняющих хранение и обработку научных и технических данных, характерен стремительный рост объемов информации в связи с постоянным появлением новых источников данных. Также проблемой является разнородность накапливаемых данных и необходимость обработки как структурированной, так и текстовой информации. Кроме того, увеличивается разнообразие и сложность задач по обработке данных, решаемых информационными системами.

Потребность в сложных аналитических операциях над разнородными данными приводит к тому, что традиционные для направления Data Mining методы агрегирования данных становятся недостаточными и требуется решать задачи извлечения, хранения и обработки знаний.

Отмеченные тенденции характерны и для ИС, используемых в образовании, значение которых постоянно растет. Информационные технологии становятся неотъемлемым звеном инфраструктуры образовательной деятельности, их интеграция приводит к созданию

единой информационно-образовательной среды, что является основополагающим направлением комплексной информатизации сферы образования [1]. В последние годы ведутся работы международных профессиональных консорциумов по унификации технических решений в области информационных технологий для образования, ориентированных на использование web-технологий и обеспечение глобальной интероперабельности информационных ресурсов образовательных учреждений. Разрабатываются специальные стандарты для таких информационных систем, например, пул стандартов IEEE 1484, который определяет требования для разных элементов системы и, в первую очередь, определяет требования по работе с метаданными [2-4].

Для единой информационно-образовательной среды важнейшей задачей является управление знаниями, которые необходимы для организации образовательного процесса. Для ее решения в настоящей статье предлагается создание гибридной интеллектуальной информационной системы (ГИИС), способной сочетать различные модели данных и знаний и методы для их анализа.

1. Постановка задачи

Характерной тенденцией, оказывающей сильное влияние на архитектуру современных ИС, является бурное развитие глобальной сети Интернет, где информация содержится, как правило, в плохо структурированном виде. Очевидно, что это отражает особенность человека, который привык выражать свои мысли и делиться своими знаниями в виде текстов. В частности, большое количество информации, связанной с наукой и образованием, представлено научными статьями, сообщениями СМИ, нормативными документами об образовании.

Преобразование этой информации в форму хорошо структурированных данных фактографических баз данных или же извлечение из этих документов знаний и перевод их на языки формальных онтологий, например, в соответствии с требованиями пула стандартов Semantic WEB [5], представляется весьма проблематичным и пока реализовано в очень узких областях. Кроме этого, существует большое число документальных систем, в которых информация традиционно представлена в слабоструктурированном виде. С развитием сети Интернет появляется возможность подключаться к ним и получать необходимую информацию.

При этом, помимо информации, представленной в текстовой форме, ИС должны обрабатывать и структурированную информацию о науке и образовании, в частности, статистические показатели студенческой успеваемости и характеристики лабораторного оборудования.

Еще одной важной особенностью ИС управления научными знаниями является необходимость управления жизненным циклом (ЖЦ) сложных научных и технических проектов. На разных этапах ЖЦ применяются различные информационные системы, как правило, содержащие большие объемы данных, справочной информации, знаний в некоторой предметной области. В связи с этим актуальна задача интеграции различных средств

управления разнородными научными знаниями. Для этого обычно используют нейтральную модель интеграции стандарта ISO 15926 или некоторые ее модификации. В работе [6] рассмотрены методологические вопросы интеграции разнородных баз знаний путем решения задачи онтологического мэппинга знаний. Авторами была предложена аналогичная технология интеграции разнородных знаний, описанная в работе [7].

Отмеченные особенности определяют требования к архитектуре разрабатываемой ИС. Предлагается реализовывать ее в виде ГИИС, что обеспечивает дополнительный синергетический эффект при взаимодействии различных интеллектуальных систем, входящих в ее состав [8,9].

Чтобы иметь возможность эффективно решать задачи управления знаниями о науке и образовании, она должна быть способна манипулировать информацией как в текстовой, так и в структурированной форме, причем текстовые документы могут иметь различные наборы знаков. Для этого ГИИС должна объединять в себе различные модели данных и знаний и обеспечивать их интеграцию.

2. Предлагаемая технология управления знаниями о науке и образовании

Предлагаемая технология управления разнородными научными знаниями основывается на следующих предпосылках:

- знания содержатся в электронных документах, имеющих разные форматы (включая тексты и таблицы) и наборы метаданных;
- документы распределены по различным разнородным источникам (открытым Интернет-ресурсам и источникам ограниченного доступа);
- имеются эксперты, которые могут определить интересующие источники информации, сформулировать темы, подлежащие мониторингу и определить принципы систематизации знаний.

Предлагаемая технология охватывает все основные этапы жизненного цикла работы со знаниями современных интеллектуальных информационных систем, среди которых можно выделить следующие:

- 1) автоматический сбор информации из разнородных источников, включая публикации в сети Интернет, документальные базы данных, табличные данные, реляционные СУБД;
- 2) очистка данных, исправление ошибок и устранение противоречий;
- 3) извлечение структурированной информации из текстов и формирование баз знаний;
- 4) организация хранения данных и знаний;
- 5) выполнение различных видов анализа над хранилищами данных и знаний, включая аннотирование, кластеризацию, рубрикацию, полнотекстовый и семантический поиск, выявление трендов и прогноз развития ситуаций, географическая привязка информации;

- 6) интеграция разнородных знаний, относящихся к разным предметным областям при помощи операции онтологического мэппинга с использованием нейтральной модели стандарта ISO 15926;
- 7) организация обмена данными и знаниями между информационными системами различного назначения.

Обобщенная структура системы, реализующая предлагаемую технологию управления разнородными знаниями, приведена на рис. 1. Многие из подсистем, показанные на рис. 1, весьма сложны по своей структуре и часто используют различные модели и представления данных и знаний, а также большой набор методов их обработки и анализа. Поэтому предлагаемая ГИИС решает задачу интеграции этих средств управления данными и знаниями в единую информационную среду.

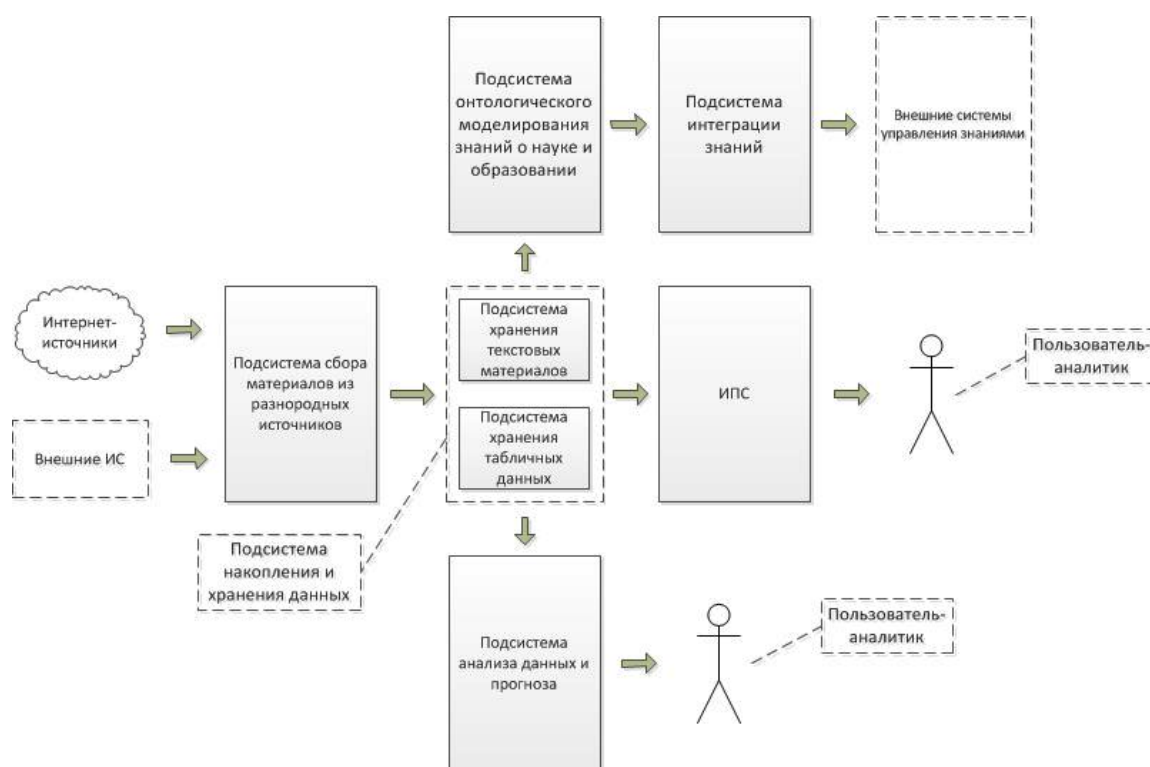


Рис. 1. Обобщенная структура системы управления научными знаниями

Рассмотрим более подробно особенности реализации отдельных подсистем предлагаемой гибридной интеллектуальной системы управления знаниями.

3. Подсистема сбора материалов из разнородных источников

Подсистема сбора выполняет автоматическую периодическую загрузку разнородных данных с Интернет-сайтов и из специализированных источников. Подсистема загружает текстовые материалы, такие как новости науки и образования и научные статьи, при этом выполняется извлечение текстовой информации и метаданных документов. Также осуществляется сбор табличных материалов и извлечение из них структурированной информации. Управление подсистемой сбора осуществляется посредством веб-интерфейса, поз-

воляющего изменять настройки автоматического сбора данных, а также запускать сбор вручную.

При сборе текстовых документов выполняется контроль качества загружаемой информации. Качество обеспечивается посредством своевременного обнаружения сбоев в работе подсистемы, обусловленных изменением верстки сайта источника [10]. Обнаружение этих сбоев очень важно для правильной работы подсистемы, поскольку в результате изменения верстки подсистема сбора может начать извлекать документы некорректно: терять некоторые фрагменты текстов, либо ошибочно извлекать фрагменты из других участков сайта, например, комментарии пользователей.

Обнаружение сбоев выполняется в два этапа. На первом этапе каждый загружаемый документ подвергается оперативному анализу: выполняется бинарная классификация документа с целью определения его корректности. В основе классификации лежит проверка схожести анализируемого документа с сообщениями, составляющими обучающую выборку, с точки зрения различных статистических характеристик. Если на первом этапе обнаружен потенциально некорректный документ, выполняется отложенная проверка, при которой анализу подвергается группа из нескольких последних загруженных сообщений. Решение о наличии или отсутствии сбоя принимается на основе определения схожести законов распределения статистических характеристик проверяемого и эталонного наборов документов. Сочетание оперативной и отложенной проверки корректности документов позволяет обеспечить быстроту реакции на сбой и высокое качество анализа.

Также для обеспечения качества накапливаемой текстовой информации выполняется обнаружение и удаление нечетких дубликатов ранее загруженных сообщений [11]. Для этого новое сообщение сравнивается с каждым из полученных ранее документов. Результат сравнения представляет собой вектор, каждый элемент которого отражает расстояние между документами по одному из учитываемых критериев. Построенный вектор подвергается интерпретации с помощью метода опорных векторов (SVM), в результате чего пара документов относится к классу «дубликатов» или «не-дубликатов». Если документ был признан дубликатом хотя бы одного из ранее загруженных сообщений, он удаляется.

Таким образом, в результате работы подсистемы сбора возможно обнаружить сбои в источниках информации, сделать крайне маловероятной загрузку искаженной информации и исключить дублирующие документы. Кроме того, модели документов, построенные в целях контроля качества сбора, отражают особенности их структуры и содержания и могут быть использованы на этапе анализа данных.

4. Подсистема накопления и хранения данных

Подсистема накопления и хранения данных выполняет функцию хранения разнородных данных с использованием различных СУБД.

Для хранения текстовых документов возможно использование объектной СУБД ODB-Jupiter [12], а также реляционных СУБД, таких как PostgreSQL и Microsoft SQL Server. Подсистема выполняет автоматическое индексирование текстовой информации, создание

и ведение тематического рубрикатора, а также предоставляет возможность полнотекстового поиска информации. Для осуществления полнотекстового поиска используются различные информационно-поисковые системы, такие как «ODB Text» [13], Sphinx и ElasticSearch [14]. Для выполнения поиска запрос может быть задан на естественном языке (например, «Инновационные технологии в образовании»), на формализованном языке запросов (например, «(инженерное NEAR/1 образование) | (подготовка NEAR/1 инженерных NEAR/3 кадров)»), либо на проблемно-ориентированном языке поиска по реквизитам (например, «@timestamp>"01.06.2016",@timestamp<"30.06.2016"»).

Информация, извлекаемая из сети Интернет, а также от некоторых специализированных ИС может содержать большое число цифровых данных, представленных в различных форматах, причем эти форматы могут быть разными даже у одного источника информации. В таких случаях выполняется извлечение цифровых данных, преобразование их к единому формату и хранение их в реляционных СУБД. Были проработаны различные варианты такого хранения, наиболее востребованным оказался вариант хранения в СУБД Microsoft SQL Server. При этом исходные данные также сохраняются, но в объектной СУБД ODB-Jupiter, что дает возможность иметь доступ к оригиналам документов.

При хранении цифровых данных в реляционной СУБД интерпретация их семантики возлагается на те аналитические системы, которые должны их анализировать. Сами данные имеют только информацию об их положении в исходной таблице: идентификатор источника информации, идентификатор таблицы, номер строки и столбца. Такой подход весьма типичен для такого направления как NoSQL [15]. Он позволяет упростить реализацию хранилищ данных и достичь всех вытекающих отсюда преимуществ, но переносит основную работу на аналитика, который должен очень хорошо представлять данные, с которыми он работает, чтобы написать правильные запросы. Так как в нашем случае подход NoSQL имитируется в реляционной СУБД, то аналитические запросы к числовым данным должны быть записаны на языке SQL.

В настоящий момент ведется работа по созданию технологии гибридного хранения разнородных данных, позволяющей накапливать текстовые и структурированные данные в едином хранилище и выполнять гибридные запросы для поиска этих данных.

5. Подсистема анализа данных и прогноза

Подсистема анализа данных и прогноза выполняет аналитическую обработку накапливаемых данных. Используемые методы анализа определяются спецификой решаемых задач, связанных с обработкой научных и технических данных.

Были проработаны и реализованы основные способы анализа текстовых документов, такие как автоматическая кластеризация и рубрикация, квазиреферирование, выделение понятий. Одной из основных функций подсистемы является анализ активности тем, связанных с наукой и образованием. Активность темы определяется частотой встречаемости документов, удовлетворяющих соответствующему поисковому запросу. Это позволяет использовать методы анализа временных рядов для прогнозирования динамики активности заданной темы в будущем.

Архитектура подсистемы анализа данных и прогноза позволяет выполнять подключение нескольких сервисов построения прогноза, реализующих разные методы. Сервис визуализации данных позволяет собирать результаты прогноза, поступающие от различных сервисов и отображать их в Веб-клиенте пользователя-аналитика. В идеале данная подсистема должна сама выбирать или рекомендовать аналитику оптимальный метод прогноза, таким образом, целью дальнейших исследований является создание ГИИС с адаптивной гибридизацией.

В настоящее время для решения задачи прогнозирования динамики активности тем по текстовым сообщениям нами активно исследуются и реализуются различные модели и методы прогноза, такие как авторегрессионная модель, модели на основе эволюционных вычислений и искусственные нейронные сети. Также предусмотрена возможность выполнять прогноз и отображать его результаты средствами внешних аналитических систем.

Актуальной задачей является анализ связей между темами: это позволяет выявлять скрытые зависимости и наличие влияния одной темы на другую. Для выполнения такого анализа подсистема осуществляет построения графа тем, в котором узлы соответствуют темам, а ребра - связям между ними. При этом цвет и размер элементов графа соответствует текущей активности соответствующих тем и связей.

Еще одной ключевой функцией подсистемы является анализ и прогноз развития ситуаций на основе текстовой информации. В его основе лежит выявление в потоке текстовых сообщений событий, относящихся к заданным темам, связанным с наукой и образованием. На основе обнаруженных событий выполняется отслеживание развития ситуаций и построение возможных сценариев их дальнейшего развития.

В подсистеме анализа данных и прогноза реализованы различные способы представления результатов. Это различные аналитические отчеты, структура которых настраивается пользователями, графики реального и прогнозируемого развития ситуации, графы связей между темами и списки обнаруженных событий и ситуаций.

Для решения целого ряда аналитических задач бывает важным учитывать географическую привязку интересующей информации. Для этого предусмотрен сервис гео-привязки, который выполняет извлечение географической информации из текстов, построение распределения документов по географическому принципу и визуализацию результатов извлечения. В качестве простейшей гео-информационной системы, в которой выполняется отображение информации на трехмерной карте, используется программа Google Earth. Наличие в этой программе средства отображения динамических изменений информации дает возможность объединить результаты анализа текстов во времени и по территориальному признаку. В частности, это позволяет сравнить динамику развития некоторых образовательных технологий в различных странах.

6. Подсистема онтологического моделирования знаний о науке и образовании

Подсистема онтологического моделирования знаний о науке и образовании включает в свой состав средства автоматического извлечения структурированной информации из текстов с использованием методов, относящихся к направлению Information Extraction

[16]. Извлечение основано на правилах, оперирующих морфологическими и предметно-ориентированными признаками и формируемых с использованием машинного обучения.

Для решения задачи построения онтологических справочников предусмотрена возможность создания каталожных описаний, которые могут содержать структурированную информацию об объектах, в частности – о научном оборудовании. Построенные каталожные описания могут быть автоматически преобразованы к виду формальных онтологий, для редактирования которых используется редактор онтологий Protégé. Хранение онтологий в виде RDF-триплетов возможно в различных СУБД с использованием системы Jena, доступ к ним осуществляется на языке SPARQL.

С помощью специальной программы управления онтологией осуществляется создание и наполнение онтологии. Так как в данной системе выполняется онтологическое моделирование знаний, поступающих из различных источников, создается локальная, поименованная область онтологии, связанная с каждым типом источника информации. Программой создаётся схема онтологии (т.е. описание классов и свойств области онтологии) и осуществляется загрузка информационных объектов. В результате формируется заполненная онтология, содержащая несколько поименованных областей (по количеству типов источников данных).

В рамках подсистемы онтологического моделирования знаний также решается задача семантического контроля текстовых документов, которая заключается в обнаружении и исправлении семантических ошибок. Для этого извлекаемая из слабоструктурированных источников данных информация проверяется на соответствие некоторым эталонным знаниям, представленным в виде формальных онтологий.

7. Подсистема интеграции знаний

Подсистема интеграции знаний обеспечивает выполнение операции онтологического мэппинга разнородных баз знаний, представленных в виде формальных онтологий в соответствии с рекомендациями пула стандартов Semantic WEB консорциума W3C. В результате этого появляется возможность получить доступ к данным и знаниям о науке и образовании из различных внешних интеллектуальных систем, и решать задачи управления ЖЦ сложных научных и технических проектов, как это предложено в работе [7].

8. Пример использования системы

Разработанная система управления знаниями о науке и образовании была использована в рамках работы с ядерными знаниями, выполняемой для корпорации «Росатом». В частности, технология анализа и прогнозирования динамики активности тем применялась для построения прогноза активности темы «Ядерная безопасность». Для оценки достоверности прогноз строился за определенный период в прошлом, и его результаты сравнивались с реальной информацией. Результат прогнозирования приведен на рис. 2. Зеленая линия соответствует реальной активности темы, красная – прогнозируемой.

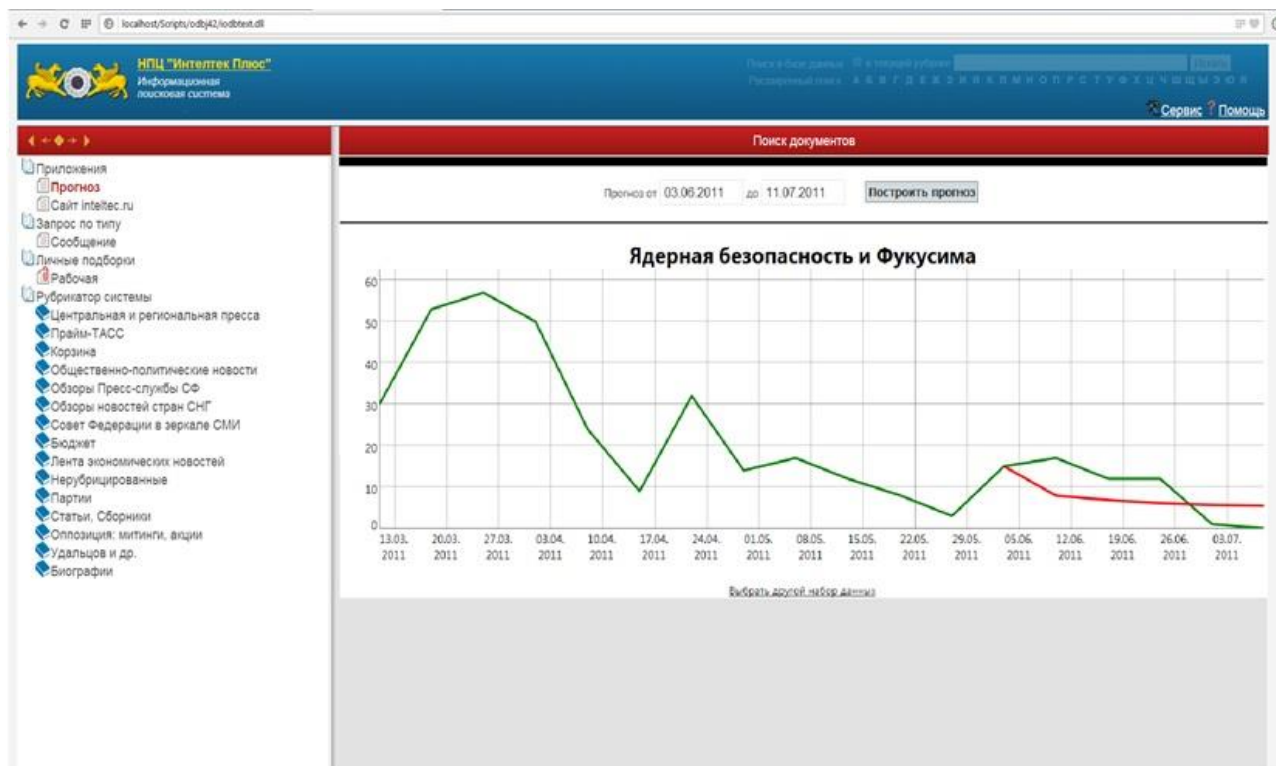


Рис. 2. Результат прогнозирования активности темы «Ядерная безопасность»

Также выполнялось онтологическое моделирование ядерных знаний. Были подготовлены каталожные описания оборудования АЭС, на основе которых было автоматически построено онтологическое представление объектов предметной области. Построенное каталожное описание и сформированный на его основе фрагмент онтологии показаны на рис. 3.

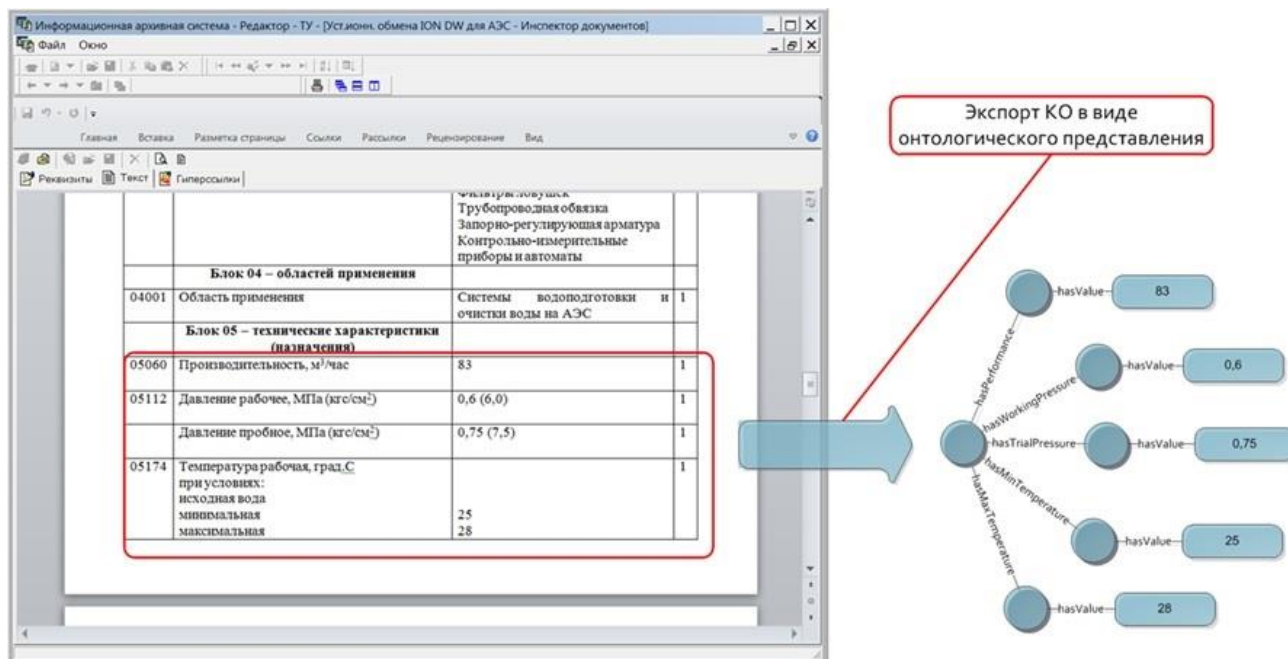


Рис. 3. Автоматическое формирование онтологического представления ядерных знаний

9. Направления дальнейших исследований

Наличие в системе как числовых, так и текстовых данных, относящихся к одному объекту или событию, открывает возможности для их совместного использования для анализа и прогнозирования. На данный момент анализ текстовой и числовой информации выполняется независимыми аналитическими системами. Предполагается их последующая интеграция в рамках единой ГИИС на основе сервис-ориентированной архитектуры, что позволит решать задачи анализа разнородной информации. Одной из таких задач, является, например, прогнозирование последствий внедрения в учебный процесс тех или иных образовательных технологий и методов.

В качестве дальнейшего развития технологии гео-привязки предполагается создание «интеллектуальных» карт, которые будут содержать не только картографическую информацию, но и семантику некоторой предметной области, а также правила логического вывода. В результате появится возможность задавать вопросы на языке, близком к естественному, или на некотором диалекте языка SPARQL и получать ответы в виде новых объектов на карте, формируемых в ходе выполнения семантического анализа.

Заключение

Представленная гибридная интеллектуальная информационная система обеспечивает возможность интеграции различных средств управления разнородными данными и знаниями о науке и образовании в единую информационно-образовательную среду ВУЗа. Система может быть использована для выявления тенденций и перспективных направлений в развитии науки и образования, поиска инновационных методов и технологий, поддержки принятия решений в области управления образованием.

Состав блоков ГИИС может меняться в зависимости от конкретной области ее применения, а сами структурные элементы подлежат настройке для выполнения определенных задач.

Список литературы

- [1]. Башмаков А.И., Старых В.А. Принципы и технологические основы создания открытых информационно-образовательных сред. / Под редакцией академика РАО Тихонова А.Н. М.: Бином. Лаборатория знаний. 2010. 720 с.
- [2]. 1484.12.1-2002. IEEE Standard for Learning Object Metadata. Standard by IEEE, 09/06/2002. New York: IEEE. 2002. 40 p.
- [3]. 1484.12.3-2005. IEEE Standard for Learning Technology – Extensible Markup Language (XML) Schema Definition Language Binding for Learning Object Metadata. Standard by IEEE, 11/10/2005. New York: IEEE. 2005. 60 p.
- [4]. 1484.20.1-2007. IEEE Standard for Learning Technology – Data Model for Reusable Competency Definitions. Standard by IEEE, 01/25/2008. New York: IEEE. 2008. 32 p.
- [5]. W3C Semantic Web Activity. / Сайт W3C. 2013. Режим доступа: <https://www.w3.org/2001/sw/> (дата обращения 29.07.2016).

- [6]. West M. Developing High Quality Data Models. San Francisco, CA: Morgan Kaufmann, Elsevier. 2011. 408 p.
- [7]. Березкин Д.В. Построение онтологического справочника отраслевого уровня с учетом рекомендаций стандарта ISO 15926 // Гибридные и синергетические интеллектуальные системы: теория и практика. Сборник научных трудов 1-го Международного симпозиума (г. Светлогорск, 29 июня – 2 июля 2012 г.). В 2 т. Т. 2. М.: Физматлит. 2012. С. 245-254.
- [8]. Колесников А.В., Кириков И.А., Листопад С.В. Гибридные интеллектуальные системы с самоорганизацией: координация, согласованность, спор. М.: ИПИ РАН. 2014. 189 с.
- [9]. Rajeswari P.V.N., Prasad T.V. Hybrid Systems for Knowledge Representation in Artificial Intelligence // International Journal of Advanced Research in Artificial Intelligence (IJARAI). 2012. Т. 1. №8. С. 31-36.
- [10]. Андреев А.М., Березкин Д.В., Козлов И.А., Симаков К.В. Подход к автоматизированному контролю работы системы извлечения данных с веб-сайтов // Информатика и её применения. 2013. Т. 7. №3. С. 2-13.
- [11]. Андреев А.М., Березкин Д.В., Козлов И.А., Симаков К.В. Многокритериальный метод выявления нечетких дубликатов в потоке текстовых сообщений // Системы и средства информатики. 2015. Т. 25. №1. С. 34-53.
- [12]. Когаловский М.Р. Энциклопедия технологий баз данных. М.: Финансы и статистика. 2005. 800 с.
- [13]. ИПС ODB Text. Режим доступа: <https://www.inteltec.ru/odbtext/odbtext.shtml> (дата обращения 30.07.2016).
- [14]. Elasticsearch. Search & Analyze Data in Real Time. Режим доступа: <https://www.elastic.co/products/elasticsearch> (дата обращения 30.07.2016).
- [15]. Cattell R. Scalable SQL and NoSQL data stores // Acm Sigmod Record. 2011. Vol. 39. Is. 4. С. 12-27. DOI:[10.1145/1978915.1978919](https://doi.org/10.1145/1978915.1978919)
- [16]. Poibeau T., Saggion H., Piskorski J., Yangarber R. Multi-source, Multilingual Information Extraction and Summarization. Series: Theory and Applications of Natural Language Processing. Berlin Heidelberg: Springer-Verlag. 2013. 375 с. DOI:[http://dx.doi.org/10.1007/978-3-642-28569-1](https://doi.org/10.1007/978-3-642-28569-1)