Научная статья
Original article

Экспериментальная интеграция университетских баз знаний на основе технологий семантического веба

Тельнов В.П.[1,*] и Одинцов К.В.[2]
[1]*Национальный исследовательский ядерный университет МИФИ, Обнинск, Россия*
[2]*Московский государственный университет имени М.В. Ломоносова, Москва, Россия*
[*]*Автор-корреспондент: telnov@bk.ru*

Аннотация: Исследуются оптимальные алгоритмы классификации текстового сетевого контента на русском и английском языках в интересах его интеграции с существующими базами знаний. Внешний сетевой контент может быть представлен в форматах RDF, RDFS, OWL, XML, HTML, JSON, CSV, в виде реляционных, графовых баз данных или вовсе не структурирован. Тестирование алгоритмов осуществляется методом кросс-валидации. Новизна представленного исследования состоит в применении принципа оптимальности по Парето для многокритериальной оценки и ранжирования изучаемых алгоритмов при условии отсутствия априорной информации о сравнительной значимости критериев. Предлагаемые программные решения основаны на облачных вычислениях с использованием сервисных моделей DBaaS и PaaS для обеспечения масштабируемости хранилищ данных и сетевых сервисов.

Ключевые слова: семантический веб, база знаний, машинное обучение, обработка естественных языков.

Для цитирования: Тельнов В.П., Одинцов К.В. Экспериментальная интеграция университетских баз знаний на основе технологий семантического веба. Умная цифровая экономика. 2023. Т.3, № 2, с. 15-23

Experimental Integration of University Knowledge Bases Founded on Semantic Web Technologies

Telnov V.P.[1,*] and Odintsov K.V.[2]
[1]*National Research Nuclear University MEPhI, Obninsk, Russia*
[2]*Lomonosov Moscow State University, Moscow, Russia*
[*]*Corresponding author: telnov@bk.ru*

Abstract: Optimum algorithms for classifying text network content in Russian and English are investigated in the interests of its integration with existing knowledge bases. External network content can be presented in RDF, RDF, OWL, XML, HTML, JSON, CSV formats, in the form of relational, graph databases, or not structured at all. The novelty of the research consists in the application of the Pareto's optimality principle for multi-criteria evaluation and ranking of the algorithms under study. The software solutions are based on cloud computing using DBaaS and PaaS service models to ensure scalability of data warehouses and network services.

Keywords: semantic web, knowledge base, machine learning, text classification.

For citation: Telnov V.P., Odintsov K.V. Experimental Integration of University Knowledge Bases Founded on Semantic Web Technologies. Smart Digital Economy. 2023. Vol. 3, №2. pp. 15-23

Введение

Классификация и интеграция слабо структурированных знаний на основе онтологий есть нетривиальная задача информатики. Человеку обычно нетрудно понять, связаны ли две или более сущности на основе когнитивных ассоциаций, в то время как с использованием программных механизмов не всегда легко это сделать. Наделение информационно-аналитических систем здравым смыслом и пониманием предметных областей было и остается важной задачей в области искусственного интеллекта. По состоянию на 2023 год образовательные веб-порталы университетов и открытые корпоративные системы управления знаниями обособлены друг от друга, не используют возможности семантической паутины и методы машинного обучения.

Обсуждаемый здесь проект [1] представляет собой шаг к виртуальной интеграции распределенных знаний, первоначально в части университетских баз знаний по компьютерным наукам и программированию. Проект призван объединить в себе те преимущества, которые дает совместное применение технологий семантического веба [12], методов машинного обучения [4] и оптимизация по Парето [11]. Чтобы осуществить виртуальную интеграцию внешних объектов с уже существующими базами знаний, нужно каким-то образом разумно «встроить» внешние данные в имеющиеся онтологии. Конкретно, преподаватель или инженер по знаниям, работая в редакторе онтологий, должен понимать, в какой класс (классы) онтологии поместить новый объект и как его связать с существующими объектами. При этом новые интегрированные объекты не должны нарушать такие свойства онтологии, как выполнимость (satisfiability) и непротиворечивость (consistency). Существуют алгоритмы машинного обучения и метрики, которые с той или иной точностью позволяют делать это. Одно из важных наблюдений состоит в том, что на разных графах знаний оказываются эффективными различные алгоритмы. Не существует единого метода классификации контента, наилучшего для всех баз знаний. Основной элемент научной новизны рассматриваемого проекта заключается в регулярном применении принципа оптимальности по Парето. Осуществляя виртуальную интеграцию знаний, мы каждый раз программно выбираем наилучший метод классификации контента, выполняя многокритериальную оптимизацию алгоритмов при условии, что отсутствует априорная информация о сравнительной важности критериев. Программно генерируются рекомендации и подсказки относительно того, как человеку «встраивать» новые данные в существующие базы знаний. Потенциальные пользователи результатов проекта – это студенты, преподаватели, руководители, эксперты и специалисты в области компьютерных наук, программирования и управления знаниями.

На вопросах развития технологий семантического веба, машинного обучения и обработки естественных языков концентрируются научные группы из Стэндфордского университета, Массачусетского Института Технологий, Университета Бари, Университета Лейпцига, Университета Манчестера. Мировые гиганты ИТ-индустрии активно развивают

модели представления знаний и технологии машинного обучения, среди них *IBM Watson Studio*, *Google AI & Machine Learning*, *Amazon Comprehend NLP*, *AWS Machine Learning*, Yandex DataSphere и др. Программные средства для исследований в области искусственного интеллекта и обработки естественных языков предлагают *MATLAB*, *Stanford NLP*, *Scikit-learn*, *Weka*, др. В России профильные исследования осуществляют в Центре компетенций НТИ МФТИ, Университете ИТМО, на факультете ВМК МГУ, в Институте системного программирования РАН им. В.П. Иванникова.

В 2023 году в рамках проекта [1] решаются задачи по созданию и подключению двух новых компонентов веб-портала: 1) агента «Семантическая классификация», который реализует адаптивные алгоритмы классификации текстового сетевого контента на русском и английском языках для наполнения и бесшовной виртуальной интеграции баз знаний по компьютерным наукам и программированию; 2) агента «Редактор онтологий *WebProtege*», который обеспечивает удаленный доступ к базам знаний для преподавателей и инженеров по знаниям и предоставляет инструментарий для совместного редактирования онтологий. Бесшовная виртуальная интеграция не подразумевает физическую консолидацию данных, но означает возможность навигации по внешним источникам данных при помощи браузера *RDF* [9] так, как если бы эти данные были частью одной онтологии. Адаптивность интеграции означает, что для каждой предметно-ориентированной базы знаний и для каждого корпуса текстов индивидуально подбирается и программно настраивается оптимальный по Парето метод классификации внешних данных. При этом учитываются как минимум следующие критерии качества работы метода: метрики *Accuracy*, *Precision*, *Recall*, *F1-score*, индивидуальные параметры настройки алгоритма. Изначально предполагается, что отсутствует какая-либо априорная информация о сравнительной значимости упомянутых критериев.

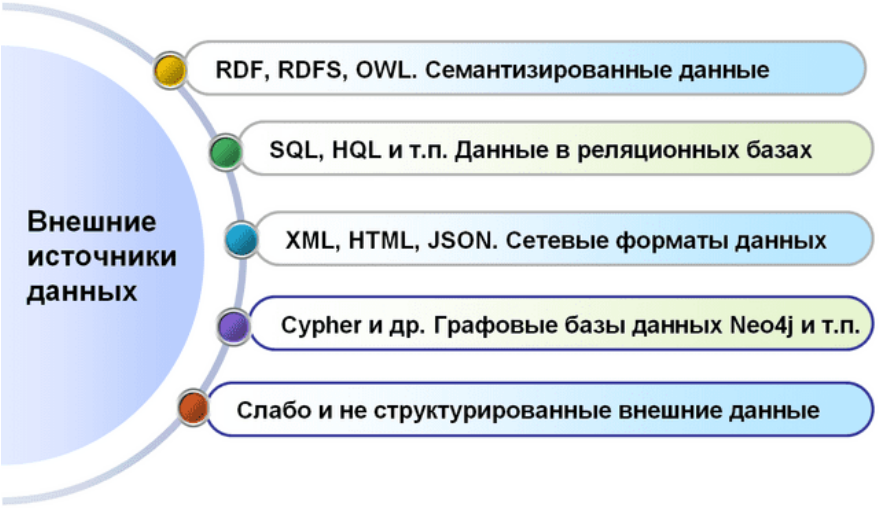

Рисунок 1 – Форматы данных для виртуальной интеграции с базами знаний

Разрабатываются алгоритмы интеграции баз знаний с внешними данными в форматах, показанных на Рис. 1. В процессе тестирования алгоритмов в роли контрольных и обучающих

множеств выступают семь графов знаний по компьютерным наукам и программированию, которые созданы авторами статьи на основе преподаваемых дисциплин:

- Технологии программирования.
- Объектно-ориентированное программирование.
- Веб-программирование на стороне клиента.
- Парадигмы и паттерны программирования.
- Облачные сервисы и технологии.
- Семантический веб.
- Объединенный граф знаний.

Для решения задач классификации текстового сетевого контента и интеграции его в базы знаний исследуются и применяются следующие методы машинного обучения:

- Классификатор на основе модели логистической регрессии и максимальной энтропии *Maxent Classifier (Softmax)*.
- Классификатор на основе метода опорных векторов *SVM Classifier with SGD*.
- Наивный байесовский классификатор *Naive Bayes Classifier*.
- Классификатор на основе терминологических деревьев решений *Decision Terminological Tree.*
- Классификатор с использованием метода ближайших соседей *Nearest Neighbors Classifier*.

В ходе решения вышеуказанных задач используется общедоступное программное обеспечение (*Apache Jena*, *Stanford NLP*, *Scikit-learn*, *Weka*) и оригинальный программный код.

Методика исследований

Методы исследования включают в себя проектирование, программную реализацию, настройку и тестирование алгоритмов. Проект реализуется на облачной платформе *Jelastic* в средах выполнения *Java* и *Python*. Основу разработки составляют стандарты и технологии семантического веба *RDFS*, *OWL*, *SPARQL*, а также дескрипционные логики *ALC* и *SROIQ(D)* [10]. Полнотекстовые учебные объекты и медийный контент размещаются в сети Интернет в произвольных удаленных хранилищах данных и на видео хостингах. Выбор конкретного удаленного хранилища не принципиален, могут использоваться любые репозитории, оснащенные средствами отображения контента (*Google Drive*, *Яндекс.Диск*, *YouTube*, др.).

Для тестирования алгоритмов классификации текстов на русском и английском языках применялся скользящий контроль (cross-validation). Каждое обучающее множество три раза разбивалось случайным образом на три выборки примерно одинакового размера. Каждая из трех выборок поочерёдно объявлялась контрольной выборкой, остальные две выборки объединялись в обучающую выборку. Алгоритм классификации текста настраивался по обучающей выборке и затем классифицировал объекты контрольной выборки. Описанная процедура повторялась три раза для каждого алгоритма классификации текста и для каждого графа знаний. Для оценки качества алгоритмов классификации использовались макро средние значения общепринятых метрик машинного обучения *Accuracy*, *Precision*, *Recall*, *F1-score*. Конкретные значения метрик усреднялись по всем классам независимо от числа объектов в

этих классах. Метрика *Accuracy* показывает долю правильно классифицированных текстов. Метрика *Precision* характеризует способность алгоритма отличать классы друг от друга. Метрика *Recall* показывает способность алгоритма обнаруживать конкретный класс вообще. Метрика *F1-score* является производной от двух предыдущих метрик и вычисляется как их среднее гармоническое. Она информативна в тех случаях, когда значения других метрик значительно разнятся между собой.

Имея результаты тестирования пяти алгоритмов классификации на семи графах знаний на русском и английском языках, возможно вычислить множество оптимальных по Парето алгоритмов, которые являются наилучшим по совокупности всех проведенных вычислительных экспериментов. Оптимизационная задача формулируется следующим образом. Требуется выбрать наилучший метод классификации с учетом всех вычисленных показателей качества, не делая никаких априорных предположений о сравнительной важности этих показателей. Для этого в классе транзитивных антирефлексивных бинарных отношений рассматривается отношение Парето в евклидовом пространстве. Данное отношение также называют отношением доминирования. Суть этого отношения состоит в следующем. Говорят, что некоторый элемент x из некоторого множества доминирует другой элемент y из этого же множества, если x не хуже y по всем аспектам (критериям) и минимум по одному аспекту превосходит y. Множество всех недоминируемых элементов называют множеством Парето. Бинарное отношение Парето обеспечивает универсальную математическую модель многокритериального контекстно-независимого выбора в евклидовом пространстве. На основе отношения Парето строится функция выбора, которая и генерирует множество элементов, наилучших с учетом всех вычисленных метрик, без каких-либо априорных предположений о сравнительной важности этих метрик. Соответствующие математические выкладки представлены в работах [3, 7].

Результаты исследований

В ходе реализации проекта [1] на семи корпусах специализированных текстов по компьютерным наукам и программированию исследована эффективность относительно простых, интуитивно понятных методов машинного обучения для решения задач наполнения из Интернета и интеграции баз знаний без непосредственного участия человека. В ходе тестирования методов машинного обучения использовались умеренные объемы данных. Каждый из семи задействованных графов знаний содержал около одной тысячи объектов и не более ста классов. Полученные в ходе вычислительных экспериментов результаты позволяют сделать следующий вывод. Среди пяти протестированных методов классификации текстов на естественных языках неожиданным лидером оказывается метод *Naive Bayes Classifier*. Он всегда входит в множество оптимальных по Парето алгоритмов. Метод *Nearest Neighbors Classifier* немногим ему уступает. Методы *Maxent Classifier (Softmax)* и *SVM Classifier with SGD* выглядят аутсайдерами на исследованных корпусах текстов.

Создаваемый семантический веб-портал по компьютерным наукам и программированию спроектирован в соответствии с архитектурным паттерном *Model-View-Controller* [8]. Алгоритмы обработки и представления данных отделены друг от друга и от собственно данных (учебных объектов). Учебные объекты объединены посредством

онтологий в графы знаний, которые размещаются в семантических репозиториях на облачной платформе. На рис. 2 ниже приведена укрупнённая диаграмма компонентов семантического образовательного веб-портала по компьютерным наукам и программированию. Вновь создаваемые программные компоненты выделены стрелками и цветом.

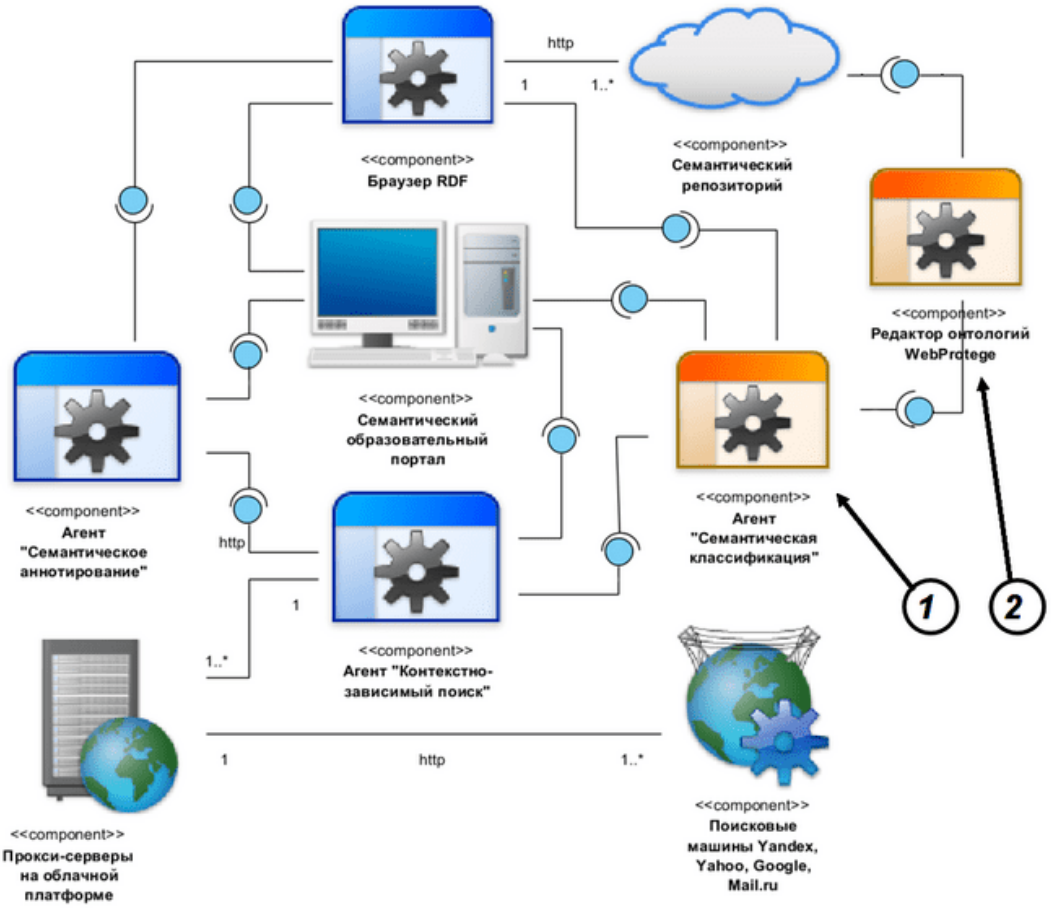

Рисунок 2 – Диаграмма компонентов семантического веб–портала в нотации *UML*. Новые компоненты выделены стрелками и цветом: 1 – агент «Семантическая классификация»; 2 – «Редактор онтологий *WebProtege*»

Общая структура программного продукта включает перечисленные ниже компоненты.

- Семантический репозиторий как хранилище графов знаний, снабженный движком *Apache Jena*.
- Поисковые виджеты для быстрого погружения в графы знаний. Их интерфейс аналогичен тому, как работает строка поисковых запросов в популярных поисковых машинах.
- Интеллектуальный браузер *RDF* для интерактивной навигации по графам знаний. Интуитивно понятный визуальный способ навигации напоминает компьютерную игру-приключение типа «бродилка», не требует специальных навыков и доступен начинающим пользователям.
- Компонент для интерактивного контекстно-зависимого поиска и селекции контента в Интернете. Использует возможности популярных поисковых машин.

- Компонент для семантического аннотирования сетевого контента в интересах наполнения и актуализации графов знаний.
- Новый программный компонент «Семантическая классификация» для классификации сетевого контента в интересах наполнения, актуализации и интеграции баз знаний.
- Подключаемый программный компонент «Редактор онтологий *WebProtege*» для удалённого доступа к графам знаний со стороны преподавателей, инженеров по знаниям и для совместного редактирования онтологий.
- Публичные точки доступа к международным базам знаний *Wikidata* и *DBpedia*.

Обсуждение

Для верификации результатов, представленных в предыдущем разделе статьи, было выполнено их сопоставление с данными, которые были получены независимыми исследователями на других корпусах текстов с применением «продвинутых» методов глубокого машинного обучения. В недавнем обзоре [5] в табл. 1 на стр. 27 приведены результаты тестирования ряда алгоритмов машинного обучения для решения задач классификации текстов. В частности, *Naive Bayes Classifier* в нашем исследовании показал среднюю точность приблизительно 96%, в то время как тот же классификатор на корпусе текстов *SST-2* дал точность 81,80%. Как показано в обзоре [5], алгоритмы глубокого машинного обучения на корпусе текстов *SST-2* дают среднюю точность около 91%, что не лучше той точности, которую показывает *Naive Bayes Classifier* на тестовых данных в настоящем исследовании. Данное наблюдение позволяет утверждать, что в настоящее время методы *Naive Bayes Classifier* и *Nearest Neighbors Classifier* обеспечивают достаточную компетентность семантических баз знаний как систем искусственного интеллекта.

В процессе разработки и тестирования программного обеспечения существующие графы знаний расширяются учебными объектами из внешних источников [2,6]. Первоначальное наполнение и последующая актуализация баз знаний, их интеграция со сторонними образовательными ресурсами есть прерогатива преподавателей университетов и инженеров по знаниям. Образовательный портал возможно тиражировать без ограничений, адаптируя содержание баз знаний под новые виды образовательных программ или уровни обучения. Программный продукт доступен из любой точки мира, где есть доступ в Интернет. Он может применяться при традиционном и дистанционном обучении как основной и дополнительный репозиторий лекционного материала, литературы для изучения, практических заданий, средств контроля знаний и т.д.

Заключение

Созданный прототип базы знаний в области компьютерных наук и программирования в настоящее время используется в учебном процессе в НИЯУ МИФИ в рамках магистерских программ. В случае внедрения подобных интегрированных баз знаний в образовательную практику университетов можно ожидать расширения профессионального кругозора студентов и повышения уровня компетенции преподавателей благодаря появлению нового унифицированного канала доступа к учебным материалам сторонних университетов.

Предлагаемый образовательный продукт предоставляет преподавателям университетов инструментарий «авторинга», способствующий созданию новых и обновлению существующих учебных курсов в области компьютерных наук и программирования. Интеллектуальный браузер *RDF* обеспечивает студентам, преподавателям и всем заинтересованным лицам возможность бесшовной интерактивной навигации по базам знаний многих университетов мира. Это ожидаемо повысит привлекательность университетского образования в глазах абитуриентов и третьих лиц, а также создаст предпосылки для расширения области применения баз знаний как систем искусственного интеллекта.

Благодарности

Исследование выполнено за счет гранта Российского научного фонда № 22-21-00182

Список литературы

1. Графы знаний по компьютерным дисциплинам. Интеллектуальные поисковые агенты. URL: http://vt.obninsk.ru/s/ (дата обращения 12.06.2023).
2. Национальный Открытый Университет «ИНТУИТ». URL: https://intuit.ru/ (дата обращения 12.06.2023).
3. Тельнов В.П., Коровин Ю.А. Применение методов машинного обучения для наполнения и актуализации баз ядерных знаний // Известия вузов. Ядерная энергетика. 2022. № 4. С. 122-133. DOI: 10.26583/npe.2022.4.11.
4. Geron A. Hands–On Machine Learning with Scikit-Learn, Keras, and TensorFlow, O'Reilly Media, Inc. Sebastopol, 2nd edition. 2019. 856 p.
5. Minaee S., Kalchbrenner N., Cambria E. et al. Deep Learning Based Text Classification: A Comprehensive Review // ACM Computing Surveys. 2022. Vol. 54, No. 3. Article 62. P. 1 – 40. DOI: 10.1145/3439726.12 W3C Semantic Web. URL: http://www.w3.org/standards/semanticweb (дата обращения 12.06.2023).
6. Massachusetts Institute of Technology OpenCourseWare. URL: http://ocw.mit.edu/search/?s=department_course_numbers.sort_coursenum (дата обращения 12.06.2023).
7. Telnov V. P., Korovin Y. A., Odintsov K. V. On the Issue of Optimum Machine Learning Methods for Filling and Updating Nuclear Knowledge Graphs // Lobachevskii J. Math. 2023. Vol. 44, No. 1. P. 227 – 236. DOI:10.1134/S1995080223010419.
8. The Model View Controller Pattern. URL: http://griffon-framework.org/tutorials/5_mvc_patterns.html (дата обращения 12.06.2023).
9. Telnov V., Korovin Y. Semantic Web and Interactive Knowledge Graphs as Educational Technology // In Cloud Computing Security / Ed. Dinesh G. Harkut. 2020. IntechOpen, London. DOI: 10.5772/intechopen.92433.
10. Telnov V., Korovin Y. Semantic web and knowledge graphs as an educational technology of personnel training for nuclear power engineering // Nuclear Energy and Technology. 2019. Vol. 5, No.3. P. 273 – 280. DOI: 10.3897/nucet.5.39226.

11. Vilfredo Pareto. URL: http://www.newworldencyclopedia.org/entry/Vilfredo_Pareto (дата обращения 12.06.2023).

12. W3C Semantic Web. URL: http://www.w3.org/standards/semanticweb (дата обращения 12.06.2023).