ИНЖИНИРИНГ ОНТОЛОГИЙ

УДК 004.89 *Научная статья* DOI: 10.18287/2223-9537-2022-12-3-336-352

Подход к автоматизированному наполнению графов знаний сущностями на основе анализа таблиц

© 2022, Н.О. Дородных✉, А.Ю. Юрин

Институт динамики систем и теории управления имени В.М. Матросова СО РАН, Иркутск, Россия

Аннотация

Использование технологий *Semantic Web*, в том числе онтологий и графов знаний, является широко распространённой практикой при разработке современных интеллектуальных систем информационного поиска, рекомендательных и вопросно-ответных систем. Процесс разработки онтологий и графов знаний включает использование различных источников информации (например, баз данных, документов, концептуальных моделей). Таблицы являются одним из наиболее доступных и широко распространённых способов хранения и представления информации, а также ценным источником знаний в предметной области. В данной работе предлагается автоматизировать процесс извлечения конкретных сущностей (фактов) из табличных данных для последующего наполнения целевого графа знаний. Для этого разработан новый подход, ключевой особенностью которого является семантическая интерпретация (аннотирование) отдельных элементов таблицы. Приведено описание его основных этапов, показано применение подхода при решении практических задач создания предметных графов знаний, в том числе в области экспертизы промышленной безопасности нефтехимического оборудования и технологических комплексов. Выполнена экспериментальная оценка качества аннотирования на тестовом наборе табличных данных. Полученные результаты показали целесообразность использования предлагаемого подхода и разработанного программного обеспечения для решения задачи извлечения фактов из табличных данных для последующего наполнения целевого графа знаний.

Ключевые слова: *semantic web, граф знаний, семантическая интерпретация таблиц, аннотирование таблиц, извлечение сущностей, таблица.*

Цитирование: *Дородных Н.О., Юрин А.Ю.* Подход к автоматизированному наполнению графов знаний сущностями на основе анализа таблиц // Онтология проектирования. 2022. Т.12, №3(45). С.336-352. DOI:10.18287/2223-9537-2022-12-3-336-352.

Благодарности: авторы выражают признательность рецензентам и членам редколлегии журнала «Онтология проектирования» за ценные замечания и рекомендации по усовершенствованию данной статьи.

Финансирование: работа выполнена при финансовой поддержке Совета по грантам Президента России (проект СП-978.2022.5) и госзадания Минобрнауки России по проекту «Методы и технологии облачной сервис-ориентированной цифровой платформы сбора, хранения и обработки больших объёмов разноформатных междисциплинарных данных и знаний, основанные на применении искусственного интеллекта, модельно-управляемого подхода и машинного обучения» (№ государственной регистрации: 121030500071-2).

Конфликт интересов: авторы заявляют об отсутствии конфликта интересов.

Введение

Технологии семантической сети (*Semantic Web*) широко используются в различных предметных областях (ПрО) в качестве основы для структурирования и формализации зна-

ний. Основными элементами этих технологий являются онтологии и графы знаний (ГЗ). ГЗ – это граф, предназначенный для накопления и передачи знаний о реальном мире, узлы которого представляют сущности, а рёбра – отношения между этими сущностями [1]. ГЗ позволяют объединять большие объёмы информации, полученные из различных источников, и обеспечивать её представление с помощью стандартизированных средств моделирования знаний. В настоящее время существует множество крупных ГЗ (например, *Google Knowledge Graph*, *Probase*, *DBpedia*, *Wikidata*, *YAGO*), которые могут хранить различные факты о реальном мире. Современные поисковые системы, такие как *Google* и *Bing*, используют ГЗ для улучшения результатов поиска. Кроме того, существуют корпоративные ГЗ, которые являются внутренними для компании и применяются для различных коммерческих целей, в частности, для торговли (например, *Amazon*, *Uber*, *eBay*, *Airbnb*), социальных сетей (например, *VK*, *LinkedIn*) и финансов (например, *Bloomberg*, *Wells Fargo*, *Accenture*, *Capital One*), в других ПрО (промышленность, оценка рисков, реклама, бизнес-аналитика и др.),

В ГЗ сущности, их атрибуты и отношения между ними типизируются на основе онтологии, представляющую собой модель ПрО [2]. Поскольку, в отличие от онтологий, конкретные сущности (экземпляры) являются центральными элементами ГЗ, наполнение ими ГЗ является важной частью процесса разработки. Такое наполнение может осуществляться с использованием различных источников информации помимо экспертов ПрО (например, баз данных, документов, концептуальных моделей). Таблицы являются одним из доступных, простых и понятных способов представления, хранения и обмена данными, а также ценным источником знаний о ПрО. В сети фиксируются миллионы таблиц, которые по некоторым оценкам содержат миллиарды ценных фактов [3], которые могут быть извлечены и использованы для построения и расширения различных ГЗ. Однако, как правило, таблицы не сопровождаются явной семантикой, необходимой для машинной интерпретации содержания. Накапливаемая в них информация часто является неструктурированной и не стандартизированной. Указанные факторы затрудняет использование таких табличных данных на практике.

В данной работе предлагается новый подход к автоматическому извлечению конкретных сущностей (фактов) из таблиц и наполнению ими целевого ГЗ. Особенностью подхода является возможность поддержки автоматизированного восстановления семантики таблиц на основе модели ПрО. Это позволяет задавать явную семантическую аннотацию для отдельных элементов таблицы (столбцов и связей между ними) и извлекать конкретные сущности из их ячеек. Предлагаемый подход реализован в виде прототипа веб-ориентированного средства, который совместно с плагином *PKBD.Onto* [4] был использован при разработке предметного ГЗ (ПрГЗ) для системы поддержки принятия решений (СППР) в задачах *экспертизы промышленной безопасности* (*ЭПБ*), в частности, при диагностике и оценке технического состояния нефтехимического оборудования и технологических комплексов. Данный подход и средство применялись в научно-исследовательских работах для платформы *TALISMAN* Института системного программирования имени В.П. Иванникова Российской академии наук (ИСП РАН). В рамках данного проекта была получена экспериментальная количественная оценка предлагаемого подхода.

1 Состояние вопроса

Таблицы могут иметь различные компоновки, стили и содержание. Их используют для различных приложений, в том числе для:

- *построения ГЗ* – процесса заполнения ГЗ информацией, извлечённой из документов, таблиц и других информационных источников [5];

- *пополнения ГЗ* – обнаружения новых фактов о сущностях из большого корпуса текстов или таблиц и пополнения ГЗ этими фактами [6];
- *расширения ГЗ* – создания новых экземпляров отношений с использованием табличных данных и обновления ГЗ извлечённой информацией [7].

Существуют специальные подходы, предметно-ориентированные языки и программные инструменты (например, *RDF123*, *csv2rdf4lod*, *Datalift*, *Spread2RDF*, *Sheet2RDF*, *XSPARQL*, *SPARQL-Generate*), предназначенные для прямого преобразования отдельных элементов таблицы в структуру целевого ГЗ или онтологии в формате *RDF* и *OWL*. В публикациях содержатся частные решения для определённых типов макетов таблиц и извлечения табличных данных без анализа и трактовки их смыслового содержания. Активно развиваются методологические основы *семантической интерпретации таблиц* и извлечения знаний из семантически аннотированных табличных данных. В последние годы появилось большое количество работ, предлагающих новые решения для связывания табличных данных с внешними понятиями, содержащимися в ГЗ. Большая часть из них сосредоточена на анализе содержимого таблиц на естественном языке и их контекста.

Все существующие подходы к семантической интерпретации (аннотированию) табличных данных можно разделить на группы:

- *аннотирование узкоспециализированных табличных данных*, учитывающее их стили оформления и структуру для определённой ПрО (например, различные измерительные данные в таблицах по естественным наукам) [8];
- *полуавтоматические подходы*, ориентированные на ручное задание соответствий между элементами исходных таблиц и целевого ГЗ [9];
- *автоматические подходы* с дополнительным участием пользователя в процессе аннотирования, которые направлены на сопоставление ячеек, столбцов и связей между столбцами с сущностями, классами и свойствами из целевого ГЗ. Эта группа включает в себя ведущие программные инструменты: *T2K Match* [10], *TableMiner+* [11], *FactBase lookup* [12], *Meimei* [13], *ColNet* [14], *Sherlock* [15], *TAKCO* [16], *MantisTable* [17], *TURL* [18]; *JHSTAbEL* [19], *DAGOBAH* [20], *MTab* [21];
- *комбинированные подходы*, объединяющие автоматический и полуавтоматический режим аннотирования таблиц [22].

Количественные сравнения, проведённые в 2021 году в рамках конкурса *SemTab*[1] на международной научной конференции (*International Semantic Web Conference – ISWC'20*), показали, что качество семантической интерпретации таблиц существующими решениями остаётся недостаточным для работы с реальными данными. Как правило, существующие инструменты решают ограниченный круг задач. Например, *JHSTabEL* [19] ориентирован только на семантическое аннотирование ячеек таблицы, а *ColNet* [14] предназначен для семантического аннотирования столбцов. Следует отметить, что существующие решения не ориентированы на непрограммирующих пользователей. Они не имеют графического пользовательского интерфейса и требуют дополнительной настройки. Более того, многие из представленных инструментов не находятся в свободном доступе.

Названные особенности и ограничения существующих исследований определяют актуальность задачи создания новых методов и программных средств, обеспечивающих построение ГЗ на основе табличных данных, включая поддержку возможности пополнения и расширения уже существующих ГЗ новыми фактами.

[1] https://www.cs.ox.ac.uk/isg/challenges/sem-tab/

2 Предлагаемый подход

2.1 Постановка задачи

В данном разделе приводятся основные предположения и некоторые формальные определения, связанные с задачей семантической интерпретации таблиц и извлечения сущностей из семантически аннотированных табличных данных. В статье приняты следующие предположения и термины.

Предположение 1. *Исходная (входная) таблица* – это реляционная таблица в третьей нормальной форме (3НФ), представляющая набор однотипных сущностей, где:

- *категориальный столбец* содержит названия (упоминания) некоторых именованных сущностей;
- *литеральный столбец* содержит литеральные значения (например, даты, числа);
- *сущностный (тематический) столбец* – это категориальный столбец, оцениваемый как потенциальный ключ и определяющий смысловое содержание исходной таблицы;
- остальные (*не сущностные*) столбцы представляют свойства сущностей, в том числе их отношения с другими сущностями.

Таким образом, каждая строка в исходной таблице содержит описание некоторого факта.

Предположение 2. Первая строка исходной таблицы является *заголовком*, содержащим имена столбцов.

Предположение 3. Значения ячеек столбца в исходной таблице имеют одинаковые синтаксические (*типы данных*) и семантические типы (*сущности*).

Предположение 4. Исходная таблица может быть представлена в формате *CSV* или *JSON*.

Предположение 5. Исходные таблицы обрабатываются независимо друг от друга.

Основной особенностью предлагаемого подхода является поддержка семантической интерпретации (аннотирования) отдельных элементов исходной таблицы.

Семантическая интерпретация (аннотирование) таблицы – процесс распознавания и связывания табличного содержания с внешними понятиями из ГЗ, включающий три основные задачи:

1) *аннотирование ячеек* – сопоставление между значениями ячеек таблицы и сущностями (конкретными объектами) из ГЗ;
2) *аннотирование столбцов* – сопоставление между столбцами или заголовками (если они доступны) и понятиями (классами или типами данных) из ГЗ;
3) *аннотирование отношений между столбцами* – поиск отношений, используя свойства, между основным сущностным столбцом и всеми остальными столбцами.

Целью данной работы является семантическое аннотирование столбцов и отношений между ними с последующим извлечением конкретных сущностей из ячеек таблицы. Данное аннотирование может осуществляться либо с использованием целевого ГЗ общего назначения (например, *DBpedia*, *Wikidata*, *YAGO*), либо с использованием целевой модели ПрО, представленной в виде онтологической схемы (онтологии на терминологическом уровне – *TBox*).

2.2 Основные этапы подхода

Предлагаемый подход может быть представлен в виде последовательности этапов (см. рисунок 1).

Этап 1. Преобразование таблиц однотипных сущностей из формата *CSV* в *JSON* представление с использованием формата «*ключ-значение*». При этом осуществляется очистка

значений ячеек: восстанавливаются некорректные символы *Unicode* и теги *HTML*, удаляются множественные пробелы и различные «*мусорные*» символы.

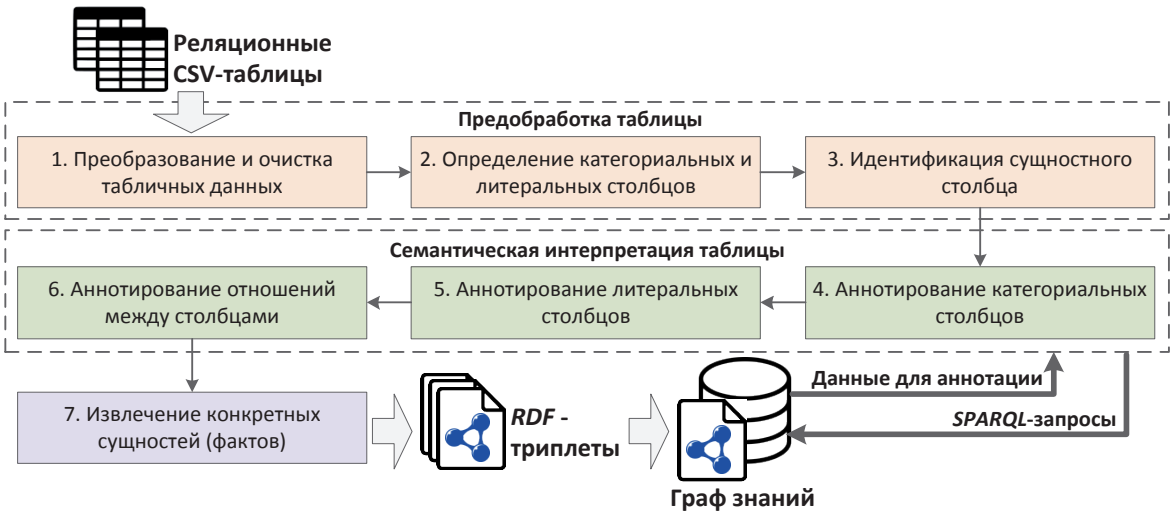


Рисунок 1 – Основные этапы предлагаемого подхода

Этап 2. Определение категориальных и литеральных столбцов в исходной таблице. Для этой цели используется библиотека для обработки естественного языка – *Stanford CoreNLP* и, в частности, *Java*-реализация распознавателя – *Stanford Named-Entity Recognizer* (*Stanford NER*)[2], который позволяет распознавать в тексте вхождения именованных сущностей (персон, компаний, местоположений и др.). *Stanford NER* определяет множество классов именованных сущностей. Эти классы присваиваются каждой ячейке в исходной таблице, характеризуя данные, которая она содержит. В зависимости от присвоенного *NER*-класса ячейка может быть как категориальная (*categorical*), так и литеральная (*literal*). В таблице 1 представлены соответствия между *NER*-классами и соответствующими типами ячеек.

Дополнительно к *NER*-аннотатору используется механизм регулярных выражений для уточнения *NER*-класса – *NONE* и *CARDINAL* (см. таблицу 2).

Таким образом, тип столбца определяется исходя из общего количества категориальных и литеральных ячеек.

Этап 3. Идентификация сущностного (тематического) столбца среди категориальных столбцов. Для решения этой задачи применяются специальные эвристики:

- *доля пустых ячеек* (*fraction of empty cells* – ***emc***) – количество пустых ячеек, отнесённое к количеству строк в столбце;
- *доля ячеек с акронимами* (*fraction of cells with acronyms* – ***acr***) – количество ячеек, содержащих акронимы, отнесённое к количеству строк в столбце;
- *доля ячеек с уникальным содержимым* (*fraction of cells with unique content* – ***uc***) – количество ячеек с уникальным текстовым содержимым, отнесённое к количеству строк в столбце;
- *среднее количество слов в ячейке* (*average number of words* – ***aw***) – вычисляется как среднее количество слов в ячейках каждого категориального столбца, при этом лучшим кандидатом является столбец с наибольшим средним количеством слов в ячейке;
- *расстояние от первого категориального столбца* (*distance from the first NE-column* – ***df***) – рассчитывается смещение на сколько столбцов текущий категориальный столбец находится от первого категориального столбца слева;

[2] https://nlp.stanford.edu/software/CRF-NER.html.

Таблица 1 – Соответствие между *NER*-классами и типами ячеек

***NER*-класс**	**Тип ячейки**	**Описание**
NONE	*categorical*	Всё, что не смог определить *NER*-аннотатор.
LOCATION	*categorical*	Локации, не относящиеся к *GPE*, например горные хребты, водоёмы и т.п.
GPE	*categorical*	Страны, города, штаты и т.п.
NORP	*categorical*	Национальности, религиозные или политические группы.
PERSON	*categorical*	Люди, в том числе вымышленные.
PRODUCT	*categorical*	Транспортные средства, оружие, продукты питания и т.п. (не услуги).
FACILITY	*categorical*	Здания, аэропорты, шоссе, мосты и т.п.
ORG	*categorical*	Компании, агентства, учреждения и т.п.
EVENT	*categorical*	Названия ураганов, сражений, войн, спортивных событий и т.п.
WORK OF ART	*categorical*	Названия фильмов, книг, песен и т.п.
LAW	*categorical*	Именованные документы, законы и т.п.
DATE	*literal*	Абсолютные или относительные даты или периоды.
TIME	*literal*	Время (меньше суток).
PERCENT	*literal*	Проценты (включая символ – %).
MONEY	*literal*	Денежное выражение, включая единицы измерения.
QUANTITY	*literal*	Измерения по весу или расстоянию.
ORDINAL	*literal*	Порядковые номера: «первый», «второй» и т.д.
CARDINAL	*literal*	Цифры, не относящиеся к другим типам.

Таблица 2 – Соответствия между дополнительными *NER*-классами и типами ячеек

***NER*-класс**	**Тип ячейки**	**Описание**
POSITIVE INTEGER	*literal*	Целое положительное число.
NEGATIVE INTEGER	*literal*	Целое отрицательное число.
FLOAT	*literal*	Число с плавающей точкой.
BOOLEAN	*literal*	Логическое значение: «*true*», «*false*» и т.д.
MAIL	*literal*	Почтовый индекс.
EMAIL	*literal*	Адрес электронной почты.
ISSN	*literal*	Номер *ISSN*.
ISBN	*literal*	Номер *ISBN*.
IP V4	*literal*	*IP*-адрес версии 4.
BANK CARD	*literal*	Номер банковской карты.
COORDINATES	*literal*	Координаты долготы и широты.
PHONE	*literal*	Номер мобильного телефона.
COLOR	*literal*	Номер цвета в 16 бит.
TEMPERATURE	*literal*	Температура в градусах Цельсия или Фаренгейта.
URL	*literal*	*URL*-адрес.
EMPTY	*literal*	Пустое значение.

- *название предлогов в заголовке столбца* (*name of prepositions in column header* – ***hpn***) – если название заголовка является предлогом, то столбец, вероятно, не является сущностным, а, скорее всего, образует отношение с сущностным столбцом.

uc и ***aw*** являются основными показательными характеристиками сущностного столбца в исходной таблице. При этом столбец, содержащий пустые ячейки (***emc***) или акронимы (***acr***), а также столбец с заголовком предлогом (***hpn***), подвергаются штрафу, а общая оценка для столбца нормализуется по его расстоянию от самого левого категориального столбца (***df***).

Таким образом, применяя все эти эвристики, можно вычислить общую (агрегированную) оценку, определяющую итоговую вероятность того, что определённый категориальный столбец является наиболее подходящим для сущностного (тематического) столбца:

$$subcol(c_j) = \frac{w_1 \times uc(c_j) + w_2 \times aw(c_j) - w_3 \times emc(c_j) - w_4 \times arc(c_j) - w_5 \times hpn(c_j)}{\sqrt{df(c_j)+1}},$$

где w_1 - w_5 – это весовые коэффициенты, которые определяют важность отдельной эвристики. По умолчанию $w_1 = 2$, а все остальные коэффициенты равны 1.

Этап 4. Семантическое аннотирование категориальных столбцов таблицы. Данная задача решается в два последовательных подэтапа:

1) *поиск и формирование набора классов кандидатов* для каждого категориального столбца, включая сущностный столбец, на основе целевого ГЗ или онтологической схемы;
2) *выбор наиболее подходящего класса из набора кандидатов* в качестве релевантной аннотации для назначения категориальному столбцу, включая сущностный столбец.

Предполагается, что целевой ГЗ или онтологическая схема для некоторой ПрО, к которой принадлежит исходная таблица, сформированы заранее и предоставляют полную и корректную информацию. Поиск классов кандидатов осуществляется либо по точному совпадению значения заголовка столбца с названием класса кандидата, либо по совпадению отдельных *N*-грамм заголовка и нечёткому лексическому сопоставлению. Для этой цели применяется метрика *расстояния редактирования* (*Левенштейна*). Запросы к целевому ГЗ или онтологической схеме осуществляются с использованием стандартизированного языка запросов – *SPARQL*[3]. При этом набор классов кандидатов ранжируется по убыванию вероятности совпадения класса кандидата с заголовком столбца. Класс кандидат с наивысшей оценкой вероятности определяется релевантным по умолчанию. Однако окончательный выбор наиболее подходящего класса из данного набора осуществляется пользователем (экспертом).

Этап 5. Автоматическое связывание типов данных из целевого ГЗ или онтологической схемы с литеральными столбцами в исходной таблице. В качестве целевых типов данных используется стандарт консорциума *W3C – XML-схема* (*XML Schema Datatypes*)[4]. *XML*-схема имеет множество встроенных простых типов данных. Встроенные типы включают примитивные и производные типы. Примитивные типы данных не являются производными от других типов. Например, числа с плавающей запятой (*float*) – это математическое понятие, которое не является производным от других типов. Производные типы данных определяются в терминах существующих типов данных. Например, целочисленный тип данных (*integer*) является частным случаем, производным от десятичного типа (*decimal*).

Для автоматического связывания литеральных столбцов с типами данных из *XML*-схемы предполагается использовать информацию о ранее распознанных именованных сущностях в столбцах (см. этап 2). Каждому литеральному *NER*-классу ставится соответствие типа данных из *XML*-схемы (см. таблицу 3). Таким образом, наиболее подходящий тип данных определяется исходя из общего количества определённых типов данных из *XML*-схемы для каждой ячейки.

Этап 6. Семантическое аннотирование отношений между столбцами. Данная задача решается в два последовательных подэтапа:

1) *поиск и формирование набора свойств (предикатов) кандидатов* на основе целевого ГЗ или онтологической схемы для каждой пары столбцов: (*S, C*) и (*S*, *L*), где *S* – сущностный столбец, *C* – категориальный столбец, *L* – литеральный столбец;
2) *выбор наиболее подходящего свойства из набора кандидатов* в качестве релевантной аннотации для обозначения связи между парой столбцов.

Поиск свойств кандидатов происходит на основе определённых релевантных классов, назначенных для столбцов (см. этап 4) с использованием языка запросов *SPARQL*. Финальный выбор наиболее подходящего свойства из набора кандидатов осуществляется пользователем (экспертом).

[3] https://www.w3.org/TR/rdf-sparql-query/.

[4] http://www.w3.org/TR/xmlschema-2/.

Таблица 3 – Соответствия между литеральными *NER*-классами и типами данных *XML*-схемы

***NER*-класс**	**Тип данных**	**Описание**
DATE	*xsd:date*	Представляет календарную дату. Шаблон: *CCYY-MM-DD* (здесь необязательна часть, представляющая время).
TIME	*xsd:time*	Представляет конкретное время дня. Шаблон: *hh:mm:ss.sss* (долевая часть секунд необязательна).
PERCENT	*xsd:nonNegativeInteger*	Представляет целое число, большее или равное нулю. Шаблон: [0, 1, 2, …].
MONEY	*xsd:decimal*	Представляет произвольное число.
QUANTITY	*xsd:nonNegativeInteger*	Представляет целое число, большее или равное нулю. Шаблон: [0, 1, 2, …].
ORDINAL	*xsd:positiveInteger*	Представляет целое число, большее нуля. Шаблон: [1, 2, 3, …].
CARDINAL	*xsd:decimal*	Представляет произвольное число.
POSITIVE INTEGER	*xsd:nonNegativeInteger*	Представляет целое число, большее или равное нулю. Шаблон: [0, 1, 2, …].
NEGATIVE INTEGER	*xsd:negativeInteger*	Представляет целое число, меньшее нуля. Этот тип данных получен из *nonPositiveInteger*.
FLOAT	*xsd:float*	Представляет 32-битовое число с плавающей запятой одиночной точности.
BOOLEAN	*xsd:boolean*	Представляет логическое значение, которое может быть «*true*» или «*false*».
MAIL	*xsd:decimal*	Представляет произвольное число.
EMAIL	*xsd:string*	Представляет символьную строку.
ISSN	*xsd:string*	Представляет символьную строку.
ISBN	*xsd:string*	Представляет символьную строку.
IP V4	*xsd:string*	Представляет символьную строку.
BANK CARD	*xsd:decimal*	Представляет произвольное число.
COORDINATES	*xsd:string*	Представляет символьную строку.
PHONE	*xsd:decimal*	Представляет произвольное число.
COLOR	*xsd:string*	Представляет символьную строку.
TEMPERATURE	*xsd:string*	Представляет символьную строку.
URL	*xsd:anyURI*	Представляет *URI* как определено в *RFC* 2396. Значение *anyURI* может быть абсолютно или относительно, и может иметь необязательный идентификатор фрагмента.
EMPTY	*xsd:string*	Представляет символьную строку.

Этап 7. Извлечение конкретных сущностей (фактов) из таблицы и наполнение этими сущностями целевого ГЗ или онтологической схемы. Сущности извлекаются из каждой строки исходной таблицы (*row-to-instance extraction*) согласно определённой аннотации столбцов и связей между ними. Для каждой ячейки сущностного (тематического) столбца формируется конкретная сущность со ссылкой (*rdf:type*) на определённый релевантный класс и соответствующим свойством (предикатом), связывающим её с другой сущностью (для категориальных столбцов) или литералом (для литеральных столбцов). Таким образом, на выходе генерируется документ в формате *RDF*, содержащий извлечённые конкретные сущности со ссылками на их классы и свойства. Извлечённые *RDF*-триплеты могут пополнить целевой ГЗ или онтологическую схему на аксиоматическом уровне – *ABox*.

2.3 Программная реализация подхода

Предлагаемый подход реализован в виде прототипа веб-ориентированного программного средства с клиент-серверной архитектурой (см. рисунок 2) на языках *Python* и *PHP*. Взаимодействие с программным средством осуществляется через открытый программный веб-

интерфейс – *REST API*. Связывание исходных табличных метаданных с целевым ГЗ или онтологической схемой выполняется пользователем посредством клиентской части средства.

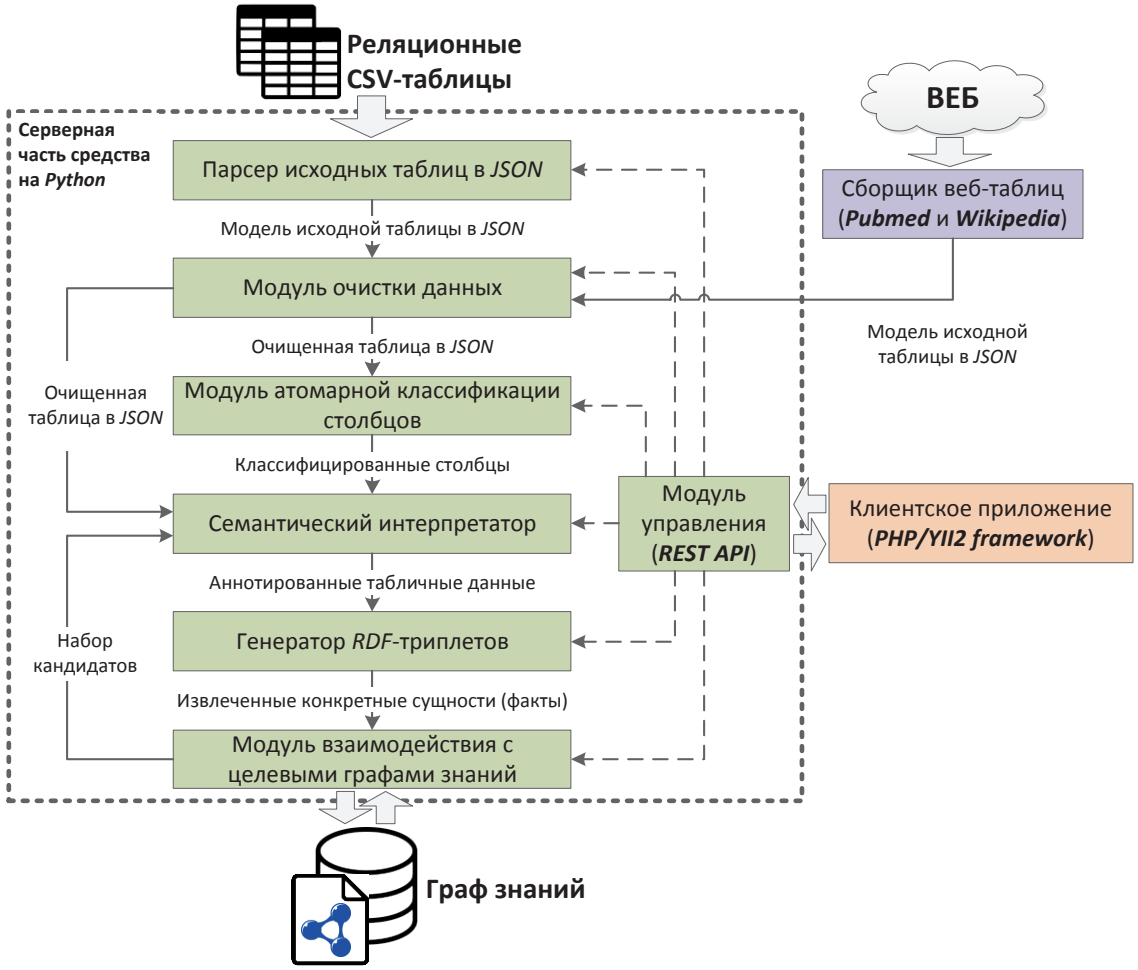


Рисунок 2 – Архитектура разработанного веб-ориентированного программного средства

Основные модули программного средства и их назначение:

- *парсер исходных таблиц в JSON* обеспечивает преобразование исходных таблиц из формата *CSV* в модель представления таблиц в формате *JSON*;
- *модуль очистки данных* обеспечивает очистку табличных данных перед основной обработкой (семантическим аннотированием);
- *модуль атомарной классификации столбцов* обеспечивает типизацию столбцов в исходной таблице на категориальные и литеральные, используя *NER*-классы, а также определяет сущностный столбец среди категориальных;
- *семантический интерпретатор* (*аннотатор*) связывает табличные данные с классами и свойствами целевого ГЗ;
- *генератор RDF-триплетов* извлекает конкретные сущности (факты) из ячеек исходной таблицы и сериализует их в формате *RDF* на основе аннотированных табличных данных;
- *модуль взаимодействия с целевыми ГЗ* обеспечивает взаимодействие с целевым ГЗ общего назначения (например, *DBpedia*) или онтологической схемой через открытую точку доступа (например, *DBpedia SPARQL Endpoint*) или специальный *GraphQL*-интерфейс с целью получения наборов кандидатов классов, типов данных и свойств; обеспечивает

наполнение целевого ГЗ или онтологическую схему недостающими извлечёнными *RDF*-триплетами;

- *модуль управления* является ядром системы, обеспечивающим вызов функций остальных модулей с использованием интерфейса взаимодействия с клиентскими приложениями через открытый *REST API* на основе микро-фреймворка *Flask*;
- *клиентское приложение* представляет собой веб-сайт, разработанного на основе *PHP*-фреймворка *Yii2*, который позволяет запрашивать аннотацию исходной таблицы и генерацию *RDF*-триплетов по *REST API*;
- *сборщик веб-таблиц* выполняет поиск по ключевым словам и извлечение веб-таблиц в формате *JSON* из архива публикаций по биомедицине *PubMed* и статей *Wikipedia*.

3 Пример практического применения

Разработанный подход и программное средство использовались при решении задачи автоматизированного создания ПрГЗ в рамках двух научно-исследовательских проектов: ИСП РАН и Иркутского научно-исследовательского и конструкторского института химического и нефтяного машиностроения (ИркутскНИИхиммаш).

3.1 Экспериментальная оценка

В рамках первого проекта проводился эксперимент с целью показать принципиальную возможность использования предлагаемого подхода и программного средства для наполнения ПрГЗ на основе табличных данных. Для этого были подготовлены три набора тестовых таблиц. Наборы формировались на основе сборщика данных (см. рисунок 2). Однако только для одного из этих наборов удалось создать релевантный целевой ГЗ, который использовался для дальнейшего семантического аннотирования таблиц. Целевой ГЗ был создан на платформе *TALISMAN*[5] и представляет собой семантический ориентированный граф, доступ к которому предоставляется через интерфейс *GraphQL*. Для тестирования был выбран и зарегистрирован набор данных – *wiki-UKU-49: United Kingdom Universities from Wikipedia*[6]. Таблицы для формирования этого набора извлекались из статей *Wikipedia* по категории «*Университеты Великобритании*». Всего было извлечено около 200 таблиц, 49 таблиц однотипных сущностей были отобраны вручную для эксперимента. Данные из отобранных таблиц аннотировались при помощи разработанного программного средства. На основе установленных аннотаций из таблиц было извлечено 1080 сущностей, которые пополнили тестовый ГЗ на платформе *TALISMAN*.

Экспериментальная количественная оценка была получена в процессе обработки таблиц из данного набора на этапах 2, 3, 5 и 7 (см. раздел 2.2).

На этапе классификации столбцов на категориальные и литеральные типы (*этап 2*) использовалась простая метрика точности:

$$accuracy = \frac{C}{N},$$

где C – количество правильно классифицированных столбцов; N – общее количество столбцов в исходной таблице.

Точность также рассчитывалась на этапе идентификации сущностного (тематического) столбца (*этап 3*). Если сущностный столбец в исходной таблице определялся правильно, то точность была равна 1, в противном случае – 0.

[5] http://talisman.ispras.ru.

[6] https://data.mendeley.com/datasets/33v9tk6jjb/1.

На этапе семантического аннотирования литеральных столбцов (*этап 5*) точность вычислялась по формуле:

$$accuracy = \frac{LC}{LN},$$

где *LC* – количество правильно аннотированных литеральных столбцов; *LN* – общее количество литеральных столбцов в исходной таблице.

На последнем этапе извлечения сущностей (фактов) из семантически аннотированных табличных данных (*этап 7*) точность вычислялась по формуле:

$$accuracy = \frac{E}{CV},$$

где *E* – количество правильно выделенных сущностей; *CV* – общее количество ячеек с уникальным содержимым для всех категориальных столбцов в исходной таблице.

Результаты экспериментальной оценки представлены в таблице 4.

Расчёт оценки производился только для полностью автоматических этапов, где не требовалось участие пользователя. Поэтому этапы семантического аннотирования столбцов (*этап 4*) и отношений между ними (*этап 6*) были пропущены. Необходимо отметить, что не все элементы исходных таблиц были проаннотированы, так как построенный целевой ГЗ оказался не полным. Это сказалось на уменьшении оценки на этапе извлечения сущностей (*этап 7*).

Таблица 4 – Экспериментальная оценка для набора данных «*wiki-UKU-49: United Kingdom Universities from Wikipedia*»

Этапы подхода	**Accuracy**
Атомарная классификация столбцов *(этап 2)*	0.95
Идентификация сущностного столбца *(этап 3)*	0.94
Аннотирование литеральных столбцов *(этап 5)*	0.96
Извлечение конкретных сущностей *(этап 7)*	0.73

3.2 Применение в задачах ЭПБ

В рамках второго проекта применение подхода проводилось при конструировании ПрГЗ для СППР в области ЭПБ [23]. ЭПБ представляет собой процедуру, которая заключается в подтверждении соответствия технического объекта требованиям промышленной безопасности. Для поддержки процедуры ЭПБ разрабатывается система, которая предназначена для диагностики и оценки технических состояний нефтехимического оборудования и технологических комплексов, в частности, путём интерпретации условий и параметров функционирования технических систем, а также параметров их технического состояния. В связи с этим, система должна располагать знаниями о техническом объекте и о различных воздействующих факторах. Ранее была разработана онтологическая схема, описывающая основные понятия и взаимосвязи для задач ЭПБ [4]. Цель текущего исследования состояла в наполнении этой схемы конкретными сущностями (фактами) для получения ПрГЗ с использованием таблиц в качестве основного источника информации.

Исходные таблицы были представлены в формате *CSV* и получены путём преобразования таблиц с произвольной компоновкой, извлечённых из отчётов по ЭПБ формы «*2.04-01/03-08*» АО ИркутскНИИхиммаш. В результате был сформирован и зарегистрирован набор данных (корпус) – *ISI-167E: Entity Spreadsheet Tables*. Набор содержит 167 реляционных таблиц однотипных сущностей. Пример фрагмента исходной таблицы с описанием фитингов из этого набора представлен в таблице 5. Все исходные таблицы из набора *ISI-167E* были семантически аннотированы на основе разработанной онтологической схемы (см. рисунок 3). Процесс аннотирования выполнялся с использованием разработанного программного средства. На рисунке 4 представлен фрагмент страницы веб-приложения с атомарными типами столбцов и аннотациями для таблицы с описанием фитингов.

Таблица 5 – Фрагмент исходной таблицы с описанием фитингов

Обознач. по чер-жу	Наименование	Кол., шт.	Условный переход, мм	Условное давление, МПа	Материал Марка
Б	Выход остатка продукта патрубок 159х4,5-180	1	150	25	Сталь 20
В	Выход паров продукта патрубок 273х8-200	1	250	25	Сталь 20
Г 1-3	Вход теплоносителя патрубок 159х4,5-190	3	150	25	Сталь 20
Е	Дренаж патрубок 57х4-110	1	50	25	Сталь 20
Ж	Люк-лаз патрубок 480х10-200	1	450	25	09Г2С
3 1-2	Бобышка регулятора уровня	2	40	-	Сталь 20
К 1-3	Люк монтажный патрубок 219х6-258	3	200	25	Сталь 20
Поз.9	Штуцер ввода трубн.пучка патрубок (1,2,3)700х36-335	3	700	25	09Г2С

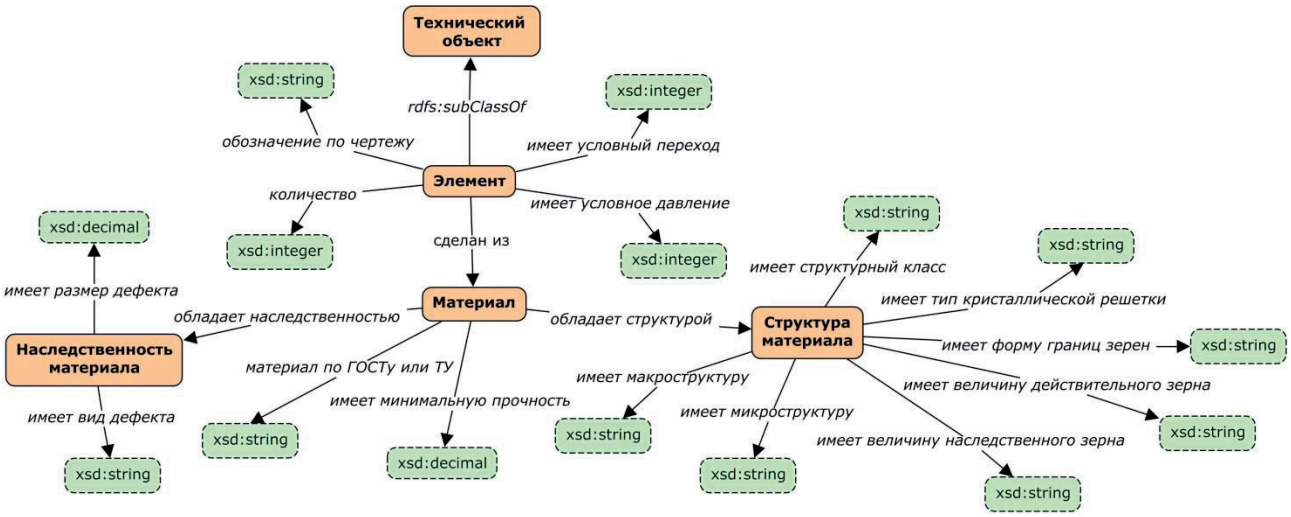


Рисунок 3 – Фрагмент онтологической схемы

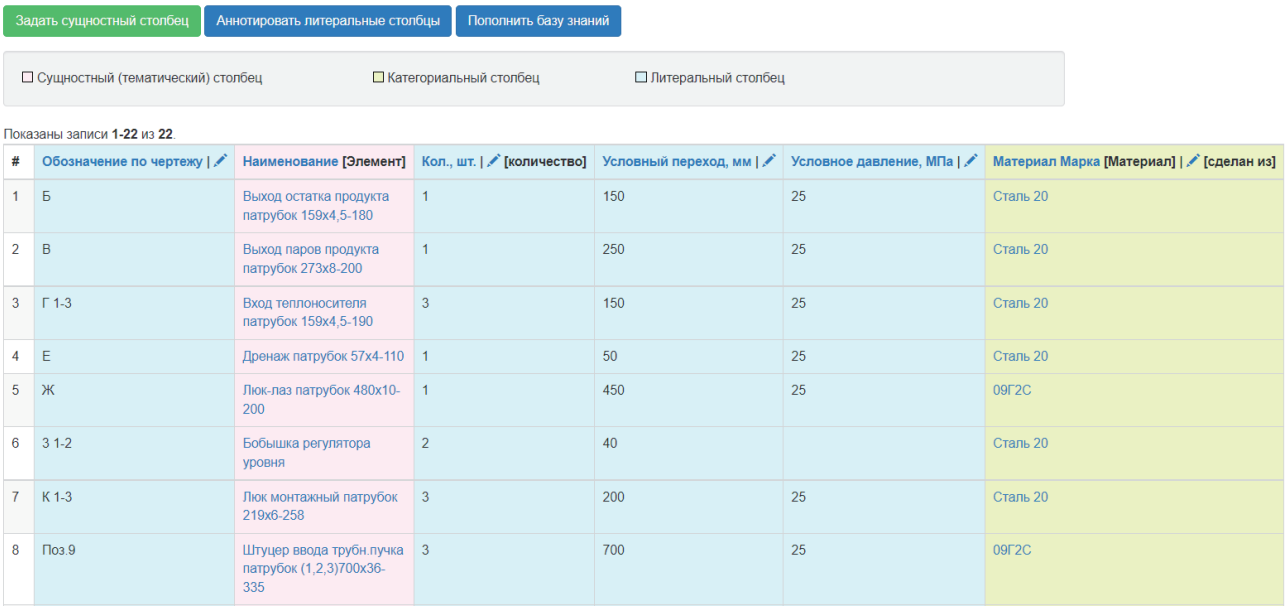


Рисунок 4 – Фрагмент страницы веб-приложения с обработанной таблицей

На рисунке 5 представлена схема полного семантического аннотирования исходной таблицы с описанием фитингов, включая извлечённые сущности (факты) из первой строки.

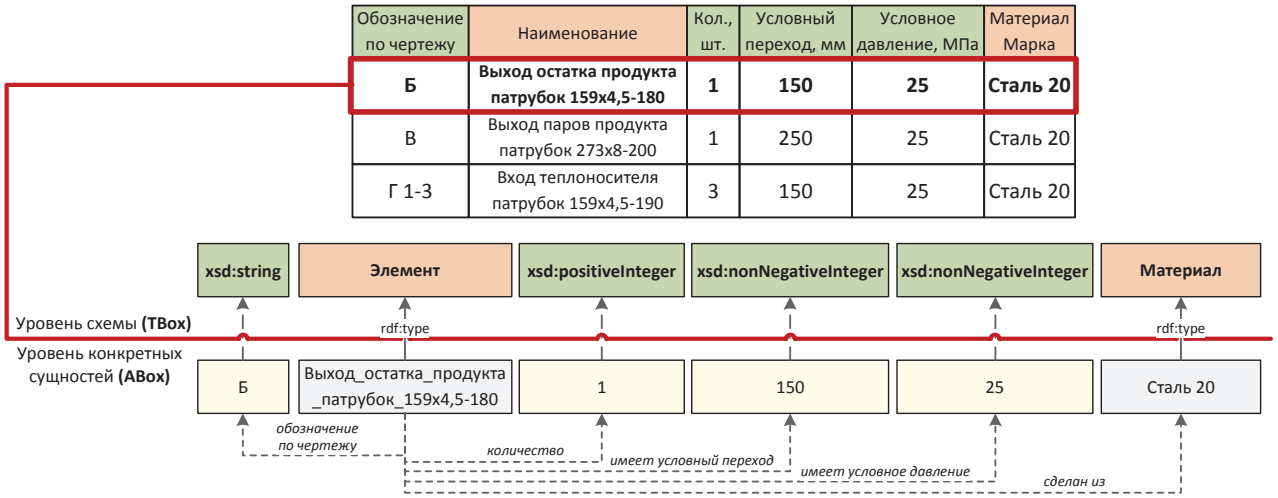


Рисунок 5 – Пример аннотированных табличных данных с извлечёнными сущностями из первой строки исходной таблицы с описанием фитингов

На основе установленных аннотаций для всех таблиц из набора *ISI-167E* было извлечено 1036 уникальных сущностей, которые пополнили онтологическую схему на аксиоматическом уровне *ABox*.

Заключение

В статье представлен подход для автоматизированной семантической интерпретации реляционных таблиц однотипных сущностей и извлечения конкретных сущностей (фактов) из аннотированных табличных данных. Извлечённые сущности могут пополнить целевой ГЗ или онтологическую схему на уровне конкретных данных (*ABox*). Предлагаемый подход реализован в форме веб-ориентированного программного средства.

Применение предлагаемого подхода и разработанного средства осуществлено в рамках научно-исследовательских проектов ИСП РАН и АО «ИркутскНИИхиммаш». Получена экспериментальная оценка подхода на тестовом наборе данных, которая показала его перспективность для задач наполнения целевого ГЗ конкретными сущностями, извлечёнными из таблиц. Решена практическая задача наполнения конкретными сущностями ранее разработанной онтологической схемы для СППР в области ЭПБ нефтехимического оборудования и технологических комплексов.

Список источников

[1] ***Hogan A., Blomqvist E., Cochez M., d'Amato C., Melo G.D., Gutierrez C., Gayo J.E.L., Kirrane S., Neumaier S., Polleres A., Navigli R., Ngomo A.-C.N., Rashid S.M., Rula A., Schmelzeisen L., Sequeda J., Staab S., Zimmermann A.*** Knowledge Graphs. *ACM Computing Surveys*. 2021. Vol. 54(4). P.1-37.

[2] ***Villazon-Terrazas B., Garcia-Santa N., Ren Y., Srinivas K., Rodriguez-Muro M., Alexopoulos P., Pan J.Z.*** Construction of Enterprise Knowledge Graphs (I). Exploiting Linked Data and Knowledge Graphs in Large Organisations. Springer, Cham. 2017.

[3] ***Lehmberg O., Ritze D., Meusel R., Bizer C.*** A large public corpus of web tables containing time and context metadata. In: Proc. of the 25th Int. Conf. Companion on World Wide Web, 2016. P.75-76.

[4] ***Видия А.В., Дородных Н.О., Юрин А.Ю.*** Подход к созданию онтологий на основе преобразования электронных таблиц с произвольной компоновкой. *Онтология проектирования*. 2021. Т. 11. № 2(40). С.212-226. DOI: 10.18287/2223-9537-2021-11-2-212-226.

[5] ***Ré C., Sadeghian A.A., Shan Z., Shin J., Wang F., Wu S., Zhang C.*** Feature engineering for knowledge base construction. *IEEE Data Engineering Bulletin*. 2014. Vol. 37. P.26-40.

[6] ***Balog K.*** Populating knowledge bases. In: Entity-Oriented Search. The Information Retrieval Series. Springer, Cham. 2018. Vol. 39. P.189-222.

[7] ***Zhang S., Balog K.*** Web table extraction, retrieval, and augmentation: A survey. *ACM Transactions on Intelligent Systems and Technology*. 2020. Vol. 11(2). P.1-35.

[8] ***De Vos M., Wielemaker J., Rijgersberg H., Schreiber G., Wielinga B., Top J.*** Combining information on structure and content to automatically annotate natural science spreadsheets. *International Journal of Human-Computer Studies*. 2017. Vol. 103. P.63-76.

[9] ***Maguire E., González-Beltrán A., Whetzel P.L., Sansone S.A., Rocca-Serra P.*** On-toMaton: A bioportal powered ontology widget for Google Spreadsheets. *Bioinformatics*. 2013. Vol. 29(4). P.525-527.

[10] ***Ritze D., Bizer C.*** Matching web tables to DBpedia - A feature utility study. In: Proc. of the 20th Int. Conf. on Extending Database Technology (EDBT'17). 2017. P.210-221.

[11] ***Zhang Z.*** Effective and efficient semantic table interpretation using TableMiner+. *Semantic Web*. 2017. Vol. 8(6). P.921-957.

[12] ***Efthymiou V., Hassanzadeh O., Rodriguez-Muro M., Christophides V.*** Matching web tables with knowledge base entities: From entity lookups to entity embeddings. In: Proc. of the 16th Int. Semantic Web Conf. (ISWC'2017). 2017. P.260-277.

[13] ***Takeoka K., Oyamada M., Nakadai S., Okadome T.*** Meimei: An efficient probabilistic approach for semantically annotating tables. Proc. of the AAAI Conf. on Artificial Intelligence. 2019. Vol. 33(1). P.281-288.

[14] ***Chen J., Jimenez-Ruiz E., Horrocks I., Sutton C.*** ColNet: Embedding the semantics of web tables for column type prediction. Proc. of the AAAI Conf. on Artificial Intelligence. 2019. Vol. 33(1). P.29-36.

[15] ***Hulsebos M., Hu K., Bakker M., Zgraggen E., Satyanarayan A., Kraska T., Demiralp Ç., Hidalgo C.*** Sherlock: A Deep Learning Approach to Semantic Data Type Detection. In: Proc. of the 25th ACM SIGKDD Int. Conf. on Knowledge Discovery & Data Mining. 2019. P.1500-1508.

[16] ***Kruit B., Boncz P., Urbani J.*** Extracting Novel Facts from Tables for Knowledge Graph Completion. Proc. of the 18th Int. Semantic Web Conf. (ISWC'2019). Lecture Notes in Computer Science. 2019. Vol. 11778. P.364-381.

[17] ***Cremaschi M., Paoli F.D., Rula A., Spahiu B.*** A fully automated approach to a complete Semantic Table Interpretation // Future Generation Computer Systems. 2020. Vol. 112. P.478-500.

[18] ***Deng X., Sun H., Lees A., Wu Y., Yu C.*** TURL: Table Understanding through Representation Learning. Proc. of the VLDB Endowment. 2020. Vol. 14(3). P.307-319.

[19] ***Xie J., Lu Y., Cao C., Li Z., Guan Y., Liu Y.*** Joint Entity Linking for Web Tables with Hybrid Semantic Matching. Proc. of the Int. Conf. on Computational Science. Lecture Notes in Computer Science. 2020. Vol. 12138. P.618-631.

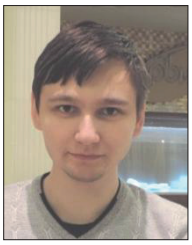
[20] ***Huynh V.-P., Liu J., Chabot Y., Deuzé F., Labbé T., Monnin P., Troncy R.*** DAGOBAH: Table and Graph Contexts for Efficient Semantic Annotation of Tabular Data. Proc. of the 20th Int. Semantic Web Conf. (ISWC'2021). SemTab. 2021. P.19-31.

[21] ***Nguyen P., Yamada I., Kertkeidkachorn N., Ichise R., Takeda H.*** SemTab 2021: Tabular Data Annotation with MTab Tool. Proc. of the 20th Int. Semantic Web Conf. (ISWC'2021). SemTab. 2021. P.92-101.

[22] ***Vu B., Knoblock C.A., Szekely P., Pham M., Pujara J.*** A Graph-Based Approach for Inferring Semantic Descriptions of Wikipedia Tables. Proc. of the 20th Int. Semantic Web Conf. (ISWC'2021). Lecture Notes in Computer Science. 2021. Vol. 12922. P.304-320.

[23] ***Берман А.Ф., Кузнецов К.А., Николайчук О.А., Павлов А.И., Юрин А.Ю.*** Информационно-аналитическая поддержка экспертизы промышленной безопасности объектов химии, нефтехимии и нефтепереработки. *Химическое и нефтегазовое машиностроение*. 2018. № 8. С.30-36.

Сведения об авторах



Дородных Никита Олегович, 1990 г. рождения. Окончил Иркутский национальный исследовательский технический университет (ИрНИТУ) (2012), к.т.н. (2018). Старший научный сотрудник Института динамики систем и теории управления им. В.М. Матросова Сибирского отделения РАН (ИДСТУ СО РАН). В списке научных трудов около 70 работ в области автоматизации создания интеллектуальных систем и баз знаний, получения знаний на основе преобразования концептуальных моделей и электронных таблиц. ORCID: 0000-0001-7794-4462; Author ID (RSCI): 979843; Author ID (Scopus): 57202323578; Researcher ID (WoS): E-8870-2014. *tualatin32@mail.ru.* ✉.

Юрин Александр Юрьевич, 1980 г. рождения. Окончил ИрНИТУ (2002), к.т.н. (2005). Заведующий лабораторией Информационных технологий исследования природной и техногенной безопасности ИДСТУ СО РАН, доцент Института информационных технологий и анализа данных ИрНИТУ. Член РАИИ и Ассоциации вычислительной техники. Член редколлегии международного научного журнала «Computer, Communication & Collaboration». В списке научных трудов более 100 работ в области разработки систем поддержки принятия решений, экспертных систем и баз знаний, использования прецедентного подхода и семантических технологий при проектировании интеллектуальных диагностических систем. ORCID: 0000-0001-9089-5730; Author ID (RSCI): 174845; Author ID (Scopus): 16311168300; Researcher ID (WoS): A-4355-2014. *iskander@icc.ru*.

Поступила в редакцию 02.06.2022, после рецензирования 24.07.2022. Принята к публикации 11.08.2022.

Scientific article DOI: 10.18287/2223-9537-2022-12-3-336-352

An approach for automated knowledge graph filling with entities based on table analysis

© 2022, N.O. Dorodnykh✉, A.Yu. Yurin

Matrosov Institute for System Dynamics and Control Theory of the Siberian Branch of Russian Academy of Sciences (ISDCT SB RAS), Irkutsk, Russia

Abstract

The use of Semantic Web technologies including ontologies and knowledge graphs is a widespread practice in the development of modern intelligent systems for information retrieval, recommendation and question-answering. The process of developing ontologies and knowledge graphs involves the use of various information sources, for example, databases, documents, conceptual models. Tables are one of the most accessible and widely used ways of storing and presenting information, as well as a valuable source of domain knowledge. In this paper, it is proposed to automate the extraction process of specific entities (facts) from tabular data for the subsequent filling of a target knowledge graph. A new approach is proposed for this purpose. A key feature of this approach is the semantic interpretation (annotation) of individual table elements. A description of its main stages is given, the application of the approach is shown in solving practical problems of creating subject knowledge graphs, including in the field of industrial safety expertise of petrochemical equipment and technological complexes. An experimental quantitative evaluation of the proposed approach was also obtained on a test set of tabular data. The obtained results showed the feasibility of using the proposed approach and the developed software to solve the problem of extracting facts from tabular data for the subsequent filling of the target knowledge graph.

Key words: *semantic web, knowledge graph, semantic table interpretation, table annotation, entity extraction, table.*

For citation: *Dorodnykh NO, Yurin AYu.* An approach and web-based tool for automated knowledge graph filling with entities from tables [In Russian]. *Ontology of designing*. 2022; 12(3): 336-352. DOI:10.18287/2223-9537-2022-12-3-336-352.

Acknowledgment: We express our gratitude to reviewers and members of the Editorial Board who made valuable comments and recommendations for improving this paper.

Financial Support: The reported study was supported by the Council for Grants of the President of Russia (grant No. SP-978.2022.5) and the Ministry of Education and Science of the Russian Federation (Project no. 121030500071-2 "Methods and technologies of a cloud-based service-oriented platform for collecting, storing and processing large volumes of multi-format interdisciplinary data and knowledge based upon the use of artificial intelligence, model-driven approach and machine learning").

Conflict of interest: The authors declares no conflict of interest.

List of figures and tables

References

[1] ***Hogan A, Blomqvist E, Cochez M, d'Amato C, Melo GD, Gutierrez C, Gayo JEL, Kirrane S, Neumaier S, Polleres A, Navigli R, Ngomo A-CN, Rashid SM, Rula A, Schmelzeisen L, Sequeda J, Staab S, Zimmermann A.*** Knowledge Graphs. ACM Computing Surveys. 2021; 54(4): 1-37.

[2] ***Villazon-Terrazas B, Garcia-Santa N, Ren Y, Srinivas K, Rodriguez-Muro M, Alexopoulos P, Pan JZ.*** Construction of Enterprise Knowledge Graphs (I). Exploiting Linked Data and Knowledge Graphs in Large Organisations. Springer, Cham. 2017.

[3] ***Lehmberg O, Ritze D, Meusel R, Bizer C.*** A large public corpus of web tables containing time and context metadata // In: Proc. of the 25th Int. Conf. Companion on World Wide Web, 2016. P.75-76.

[4] ***Vidia AV, Dorodnykh NO, Yurin AYu.*** An approach to ontology engineering based on transformation of arbitrary spreadsheets [In Russian]. *Ontology of designing*. 2021; 11(2): 212-226. DOI: 10.18287/2223-9537-2021-11-2-212-226.

[5] ***Ré C, Sadeghian AA, Shan Z, Shin J, Wang F, Wu S, Zhang C.*** Feature engineering for knowledge base construction. IEEE Data Engineering Bulletin. 2014; 37: 26-40.

[6] ***Balog K.*** Populating knowledge bases. In: Entity-Oriented Search. The Information Retrieval Series. Springer, Cham. 2018; 39: 189-222.

[7] ***Zhang S, Balog K.*** Web table extraction, retrieval, and augmentation: A survey // ACM Transactions on Intelligent Systems and Technology. 2020; 11(2): 1-35.

[8] ***De Vos M, Wielemaker J, Rijgersberg H, Schreiber G, Wielinga B, Top J.*** Combining information on structure and content to automatically annotate natural science spreadsheets. *International Journal of Human-Computer Studies*. 2017; 103: 63-76.

[9] ***Maguire E, González-Beltrán A, Whetzel PL, Sansone SA, Rocca-Serra P.*** On-toMaton: A bioportal powered ontology widget for Google Spreadsheets. *Bioinformatics*. 2013; 29(4): 525-527.

[10] ***Ritze D, Bizer C.*** Matching web tables to DBpedia - A feature utility study. In: Proc. of the 20th Int. Conf. on Extending Database Technology (EDBT'17). 2017. P.210-221.

[11] ***Zhang Z.*** Effective and efficient semantic table interpretation using TableMiner+. *Semantic Web*. 2017; 8(6): 921-957.

[12] ***Efthymiou V, Hassanzadeh O, Rodriguez-Muro M, Christophides V.*** Matching web tables with knowledge base entities: From entity lookups to entity embeddings. In: Proc. of the 16th Int. Semantic Web Conf. (ISWC'2017). 2017. P.260-277.

[13] ***Takeoka K, Oyamada M, Nakadai S, Okadome T.*** Meimei: An efficient probabilistic approach for semantically annotating tables. Proc. of the AAAI Conf. on Artificial Intelligence. 2019; 33(1): 281-288.

[14] ***Chen J, Jimenez-Ruiz E, Horrocks I, Sutton C.*** ColNet: Embedding the semantics of web tables for column type prediction. Proc. of the AAAI Conf. on Artificial Intelligence. 2019; 33(1): 29-36.
[15] ***Hulsebos M, Hu K, Bakker M, Zgraggen E, Satyanarayan A, Kraska T, Demiralp Ç, Hidalgo C.*** Sherlock: A Deep Learning Approach to Semantic Data Type Detection. In: Proc. of the 25th ACM SIGKDD Int. Conf. on Knowledge Discovery & Data Mining. 2019. P.1500-1508.
[16] ***Kruit B, Boncz P, Urbani J.*** Extracting Novel Facts from Tables for Knowledge Graph Completion. Proc. of the 18th Int. Semantic Web Conf. (ISWC'2019). Lecture Notes in Computer Science. 2019; 11778: 364-381.
[17] ***Cremaschi M, Paoli FD, Rula A, Spahiu B.*** A fully automated approach to a complete Semantic Table Interpretation. *Future Generation Computer Systems*. 2020; 112: 478-500.
[18] ***Deng X, Sun H, Lees A, Wu Y, Yu C.*** TURL: Table Understanding through Representation Learning. Proc. of the VLDB Endowment. 2020; 14(3): 307-319.
[19] ***Xie J, Lu Y, Cao C, Li Z, Guan Y, Liu Y.*** Joint Entity Linking for Web Tables with Hybrid Semantic Matching. Proc. of the Int. Conf. on Computational Science. Lecture Notes in Computer Science. 2020; 12138: 618-631.
[20] ***Huynh V-P, Liu J, Chabot Y, Deuzé F, Labbé T, Monnin P, Troncy R.*** DAGOBAH: Table and Graph Contexts for Efficient Semantic Annotation of Tabular Data. Proc. of the 20th Int. Semantic Web Conf. (ISWC'2021). SemTab. 2021. P.19-31.
[21] ***Nguyen P, Yamada I, Kertkeidkachorn N, Ichise R, Takeda H.*** SemTab 2021: Tabular Data Annotation with MTab Tool. Proc. of the 20th Int. Semantic Web Conf. (ISWC'2021). SemTab. 2021. P.92-101.
[22] ***Vu B, Knoblock CA, Szekely P, Pham M, Pujara J.*** A Graph-Based Approach for Inferring Semantic Descriptions of Wikipedia Tables. Proc. of the 20th Int. Semantic Web Conf. (ISWC'2021). Lecture Notes in Computer Science. 2021; 12922: 304-320.
[23] ***Berman AF, Kuznetsov KA, Nikolaychuk OA, Pavlov AI, Yurin AYu.*** Information and analytical support for the examination of industrial safety of chemical, petrochemical and oil refining facilities [In Russian]. *Chemical and Petroleum Engineering*. 2018; 8: 30-36.

About the authors

Nikita Olegovych Dorodnykh (b. 1990) graduated from INRTU in 2012, PhD (2018). He is a senior associate researcher at ISDCT SB RAS. Co-author of about 70 publications in the field of computer-aided development of intelligent systems and knowledge bases, knowledge acquisition based on the transformation of conceptual models and tables. ORCID: 0000-0001-7794-4462; Author ID (RSCI): 979843; Author ID (Scopus): 57202323578; Researcher ID (WoS): E-8870-2014. *tualatin32@mail.ru.* ✉.

Alexander Yurievich Yurin (b.1980) graduated from the INRTU in 2002, PhD (2005). He is the Head of the "Information and telecommunication technologies for investigation of natural and technogenic safety" laboratory at ISDCT SB RAS and associate professor of the Institute of information technologies and data analysis of INRTU. He is a member of the Russian Association of Artificial Intelligence (RAAI) and Association for Computing Machinery (ACM). He is a member of the Editorial Board of the international scientific journal "Computer, Communication & Collaboration". The list of scientific works includes more than 100 scientific papers in the field of development of decision support systems, expert systems and knowledge bases, application of the case-based reasoning and semantic technologies in the design of diagnostic intelligent systems, maintenance of reliability and safety of complex technical systems. ORCID: 0000-0001-9089-5730; Author ID (RSCI): 174845; Author ID (Scopus): 16311168300; Researcher ID (WoS): A-4355-2014. *iskander@icc.ru.*

Received June 2, 2022. Revised July 24, 2022. Accepted August 11, 2022.