УДК 004.89
DOI:10.25729/ESI.2023.30.2.014

Программное средство извлечения сущностей из семантически аннотированных табличных данных

Амирасланов Ильгар Вугарович[1], Дородных Никита Олегович[2], Юрин Александр Юрьевич[1,2]

[1]Иркутский национальный исследовательский технический университет, Россия, Иркутск, *iskander@icc.ru*

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

Аннотация. В настоящее время графы знаний широко применяются в различных предметных областях, например, в промышленности, торговле, финансах и социальных сетях. Граф знаний представляет собой мощное средство объединения и представления информации с использованием стандартизированных методов моделирования знаний. Однако, разработка графов знаний и, в частности, пополнение их новыми конкретными сущностями (фактами) остается сложной задачей. Использование различных информационных источников может облегчить этот процесс. Таким источником могут быть таблицы, которые потенциально содержат богатую семантическую информацию. В статье предлагается подход и его программная реализация для автоматизированного извлечения значимой информации из табличных данных в виде фактов и пополнения ими целевого графа знаний. Основной особенностью предлагаемого подхода является сочетание эвристических методов с моделями глубокого машинного обучения для семантического аннотирования табличных данных. Применимость подхода продемонстрирована на двух примерах: при анализе рынка труда Иркутской области и оценке технического состояния нефтехимического оборудования.

Ключевые слова: семантический веб, приобретение знаний, граф знаний, семантическая интерпретация таблиц, извлечение фактов, пополнение графа знаний, таблица

Цитирование: Амирасланов И.В. Программное средство извлечения сущностей из семантически аннотированных табличных данных / И.В. Амирасланов, Н.О. Дородных, А.Ю. Юрин // Информационные и математические технологии в науке и управлении. – 2023. – № 2(30). – С. 138-151. – DOI:10.25729/ESI.2023.30.2.014.

Введение. В настоящее время разработка предметно-ориентированных интеллектуальных систем, направленных на решение сложных задач (например, диагностирование и оценка технического состояния сложных технических систем, прогнозирование чрезвычайных ситуаций, энергетическая безопасность, медицина и т.д.), остается актуальной задачей. Одной из тенденций в этой области является использование графов знаний, предназначенных для накопления и передачи знаний о реальном мире, узлы которых представляют интересующие объекты, а ребра – отношения между этими объектами [1]. Разработка графов знаний является трудоемкой и сложной задачей, которая до сих пор не решена в полной мере. Поэтому исследования, направленные на создание новых методов обработки информации для конструирования и пополнения графов знаний при решении практических слабоформализуемых задач в различных предметных областях, являются актуальными [2]. В контексте данной проблематики перспективным подходом является использование различных информационных источников (например, баз данных, документов, концептуальных моделей). В качестве такого источника могут быть использованы и таблицы, которые широко используются в различных исследованиях и в анализе данных. Большое количество таблиц размещено в Интернет (например, в формате HTML), а также в электронных документах (например, в форматах DOCX и PDF), в рукописных материалах и в компьютерных программах. Как показали недавние исследования [3], таблицы могут являться ценным источником знаний и содержать миллионы полезных фактов. Однако таблицы весьма неоднородны по своей структуре, содержанию и назначению. В большинстве случаев они не интерпретируются компьютерными программами

без участия человека. Их исходное представление не обеспечивает всей явной семантики, необходимой для их интерпретации в автоматическом режиме. Этот факт препятствует активному практическому использованию таких табличных данных.

Ранее авторами был предложен подход для автоматизированного извлечения конкретных сущностей (фактов) из таблиц и пополнения ими целевого графа знаний [4]. В данной статье предлагается расширить и улучшить этот подход, в частности, используя новый гибридный метод семантического аннотирования столбцов таблицы, основанный на эвристических решениях и методах машинного обучения. Предлагаемый подход реализован в форме веб-ориентированного программного средства – «TabbyLD2» [5]. Также в статье представлены два прикладных примера и полученная экспериментальная оценка.

1. Состояние вопроса. Пополнение графов знаний на основе таблиц требует интерпретации табличных данных. Эта проблема известна как Семантическая Интерпретация Таблиц (СИТ) (Semantic Table Interpretation) [6] и представляет собой распознавание и связывание табличного содержания с внешними понятиями из целевого графа знаний. СИТ включает в себя три основные задачи [7]:

- *семантическое аннотирование ячеек* (*САЯ*) – сопоставление значений ячеек с сущностями (экземплярами классов) из целевого графа знаний;
- *семантическое аннотирование столбцов* (*САС*) – сопоставление столбцов таблицы с классами или типами данных из целевого графа знаний;
- *семантическое аннотирование отношений между столбцами* (*САО*) – сопоставление отношений между столбцами со свойствами (предикатами) из целевого графа знаний.

В последние годы проблеме СИТ уделяется большое внимание. В частности, методы СИТ основаны на использовании: сопоставления онтологий; поиска сущностей, как в глобальных таксономиях (например, DBpedia, Wikidata, Yago, Freebase), так и в предметно-ориентированных графах знаний; викификации и векторном представлении графов знаний. Например, в [8] экспериментально показано, что более эффективным является гибридный подход «FactBase Lookup», который сочетает в себе сервисы поиска сущностей и технику векторного представления графов знаний.

За последние три года наблюдается рост работ, связанных с автоматическим семантическим аннотированием табличных данных, в частности, веб-таблиц. Как правило, они сосредоточены на анализе естественно-языкового содержания и контекста таблиц. Так, в [9] предлагается новая модель гибридного семантического сопоставления под названием «JHSTabEL» для решения задачи САЯ. «Sherlock» [10] и «Sato» [11] определяют семантические типы для столбцов таблицы, используя более тысячи признаков, извлеченных из набора реляционных веб-таблиц для обучения искусственных нейронных сетей. При этом «Sato» дополнительно использует векторное представление таблицы с признаками скрытого размещения Дирихле (Latent Dirichlet Allocation) и попарные зависимости столбцов, моделируемые слоем случайных условных полей (Conditional Random Fields). «ColNet» [12] представляет собой фреймворк, использующий свёрточные нейронные сети (Convolutional Neural Network – CNN), встраивая общую семантику столбцов в векторное пространство для предсказания столбцам таблицы соответствующих семантических типов (классов). Программные решения: «Doduo» [13], «SeLaB» [14], «TURL» [15] и «TaBERT» [16] предоставляют предварительно обученные табличные модели с использованием архитектуры «Transformer» и языковых моделей на основе «BERT» для различных задач СИТ.

Несмотря на повышенный интерес к моделям понимания таблиц, в научной литературе описано очень мало примеров их применения на практике. Существующие подходы представляют ограниченное количество семантических типов (классов) для аннотирования таблиц

(только те типы, для которых обучены классификаторы). Еще одним очень важным недостатком таких подходов является отсутствие этапа генерации семантической разметки таблицы, например, в виде триплетов в формате RDF. Реализация большинства подходов не доводится до конечного пользователя. В частности, программные инструменты подходят только для программистов, сложны в настройке и не имеют графического пользовательского интерфейса.

2. Постановка задачи. Вертикальные таблицы, состоящие из столбцов данных, используются в качестве входных данных для предлагаемого подхода. Каждый такой столбец может описывать только один тип данных, а также может содержать заголовок.

В таких таблицах столбцы могут быть:

- *категориальными* – содержат текстовые упоминания некоторых объектов предметной области;
- *литеральными* – содержат различные литеральные значения, например, даты, числа и т.д.

Вертикальные таблицы могут содержать сущностный (тематический) столбец – это категориальный столбец, который определяет семантическое содержание исходной таблицы и может быть потенциальным первичным ключом. Это делает вертикальную таблицу реляционной.

Предположение 1. В обрабатываемых таблицах нет объединенных ячеек.

Предположение 2. Исходные таблицы обрабатываются независимо друг от друга.

Таким образом, предлагаемый подход направлен на аннотирование вертикальных таблиц (решение задач CAC и CAO) на основе целевого графа знаний. DBpedia [17] используется в качестве целевого графа знаний общего назначения.

3. Предлагаемый подход.

3.1. Основные этапы подхода. Основываясь на результатах [4, 18], предлагаемый подход состоит из следующих основных этапов (рис. 1):

- *предварительная обработка таблицы* – подготовка табличных данных к дальнейшему процессу аннотирования. Данный этап включает в себя четыре основные задачи: (1) преобразование форматов таблиц; (2) очистка данных (например, восстановление не-корректных символов Юникода и тегов HTML, удаление множественных пробелов и «мусорных» символов); (3) атомарная классификация столбцов на категориальные (со-держащие именованные сущности) и литеральные (содержащие различные литеральные значения, например, даты или числа) типы; (4) идентификация сущностного (тематического) столбца, определяющего смысловое содержание таблицы;
- *связывание сущностей* – представляет собой решение задачи САЯ;
- *определение семантических типов для столбцов* – представляет собой решение задачи CAC отдельно для категориальных, включая сущностный столбец, и литеральных столбцов;
- *определение связей между столбцами* – представляет собой решение задачи CAO;
- *извлечение сущностей* – извлечение новых конкретных сущностей (фактов) на основе определенных аннотаций для столбцов и отношений между ними. Извлеченные таким образом сущности могут пополнить целевой граф знаний.

В обновленной версии данного подхода улучшены этапы предварительной обработки таблицы в контексте решения задачи определения атомарных типов столбцов, а также семантического аннотирования категориальных столбцов. Подробное описание разработанного алгоритмического обеспечения, решающего остальные задачи на других этапах, представлено в работе [4]. Далее подробно рассмотрим обновленные этапы (рис.1).

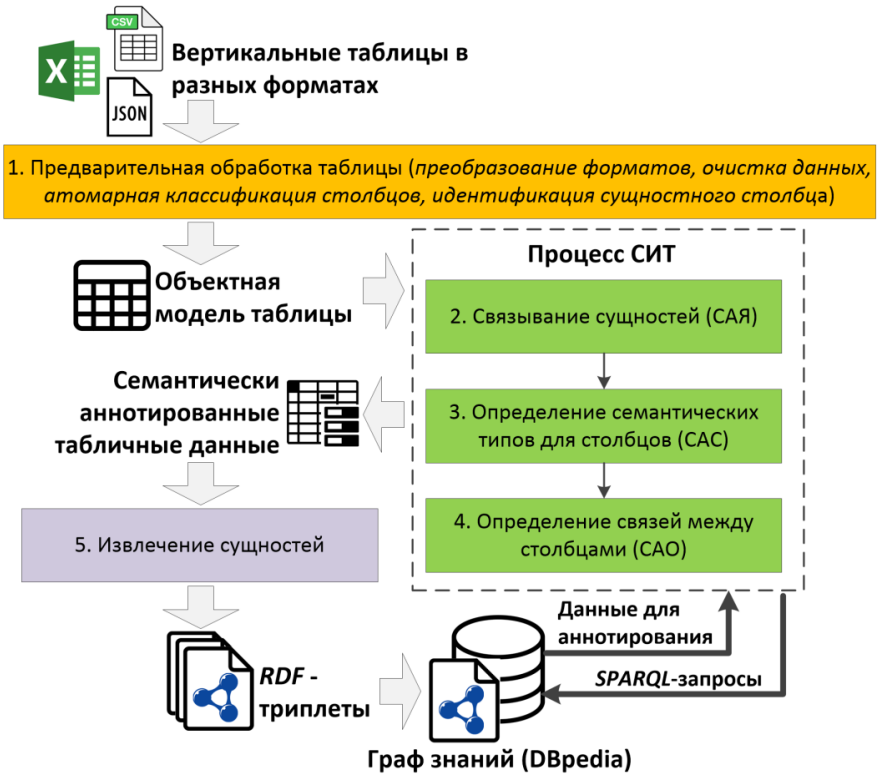

Рис. 1. Основные этапы предлагаемого подхода

3.2. Атомарная классификация столбцов. Атомарная классификация столбцов представляет собой автоматическое определение категориальных и литеральных столбцов в исходной таблице. Для этого используется библиотека обработки естественного языка «Stanford CoreNLP» [19] и, в частности, распознаватель именованных сущностей «Stanford Named-Entity Recognizer (Stanford NER)». «Stanford NER» распознает вхождения различных именованных сущностей в тексте. Для этого распознаватель включает в себя следующий набор так называемых NER-классов: «PERSON», «NORP», «FACILITY», «ORGANIZATION», «GPE», «LOCATION», «PRODUCT», «EVENT», «ART_WORK», «LAW», «DATE», «TIME», «PERCENT», «MONEY», «QUANTITY», «ORDINAL», «CARDINAL». Однако распознаватель плохо справляется с короткими текстами, представленными в ячейках таблицы. По этой причине дополнительно используется библиотека «Duckling» [20]. Данная библиотека написана на языке Haskell и позволяет анализировать входной текст и представлять его в виде структурированных распознанных данных с использованием большого набора именованных классов сущностей, например, «AmountOfMoney», «CreditCardNumber», «Distance» и др. Кроме того, используется библиотека «DateParser» [21] для лучшего определения даты в различных форматах. Таким образом, каждой ячейке в исходной таблице назначается расширенный набор NER-классов. В зависимости от назначенного NER-класса ячейка может быть либо категориальной, либо литеральной. Атомарный тип для столбца определяется на основе общего количества категориальных и литеральных ячеек. Результаты этого этапа используются в дальнейшем процессе СИТ.

3.3. Определение семантических типов для столбцов. В работе [4] использовались следующие методы определения семантического типа (класса) для категориальных столбцов, включая сущностный столбец:

– *голосование методом простого большинства* – это эвристика, в которой используется подсчет частоты встречаемости классов, которым принадлежат определенные релевантные сущности, присвоенные каждой ячейке в столбце. Таким образом, данная эвристика

полностью основана на результатах решения задачи САЯ, поэтому она не гарантирует высокой степени соответствия выбранному классу;

- *сходство по заголовку* – это эвристика, использующая лексическое сопоставление между названием заголовка и названием класса из полученного набора кандидатов. Данная эвристика активно использует заголовки столбцов, которые часто недоступны в реальных табличных данных;
- *предсказание класса* – это метод, основанный на фреймворке «ColNet» [12], который использует настраиваемые бинарные классификаторы (CNN-модели) для предсказания релевантного класса из набора кандидатов соответствующему категориальному столбцу. Однако этот метод, так же, как и эвристика «*голосование методом простого большинства*», зависит от результатов решения задачи САЯ, которые в большинстве практических задач отсутствуют

Для устранения выделенных недостатков предлагается модифицированное решение задачи САС, состоящее из трёх методов:

1. Метод на основе NER-классов. Данный метод основан на информации об уже распознанных именованных сущностях (NER-классах) на этапе атомарной классификации столбцов (см. подраздел 3.2). Определенные для каждой ячейки NER-классы сопоставляются с классами из целевого графа знаний DBpedia (таблица 1).

Таблица 1. Соответствия основных NER-классов и классов из графа знаний DBpedia

NER-классы	Классы из DBpedia	Описание
LOCATION	dbo:Park, dbo:Mine, dbo:Garden, dbo:Cemetery, dbo:WineRegion, dbo:NaturalPlace, dbo:ProtectedArea, dbo:WorldHeritageSite, dbo:SiteOfSpecialScientificInterest	*Места, не относящиеся к GPE, например, горные хребты, водоемы, реки и т.д.*
GPE	dbo:PopulatedPlace	*Страны, города, штаты.*
NORP	dbo:EthnicGroup	*Национальности, религиозные или политические группы.*
PERSON	dbo:Person	*Люди, в том числе вымышленные.*
PRODUCT	dbo:Device, dbo:Food, dbo:MeanOfTransportation	*Транспорт, оружие, еда и т.д. (не услуги).*
FACILITY	dbo:ArchitecturalStructure	*Здания, аэропорты, автомагистрали, мосты и т.д.*
ORG	dbo:Organisation	*Компании, агентства, учреждения и т.д.*
EVENT	dbo:Event	*Названия ураганов, сражений, войн, спортивных событий и т.д.*
WORK_OF_ART	dbo:Work	*Названия книг, песен и т.д.*
LAW	dbo:Law, dbo:LegalCase	*Названия нормативных документов, законов.*
NONE	owl#Thing	*Результат NER пуст.*

Далее подсчитывается количество появлений соответствующего класса DBpedia. Данное количество преобразуется в оценку от 0 до 1 на основе алгоритма «Minimax». Класс с наивысшей нормализованной оценкой определяется как наиболее подходящий и присваивается текущему категориальному столбцу.

2. Метод предсказания класса на основе контекста. Данный метод основан на фреймворке «Sato» [11] и обрабатывает два типа контекста:

- *глобальный контекст* – представляет все значения ячеек исходной таблицы, т.е. он одинаков для каждого столбца исходной таблицы;
- *локальный контекст* – представляет набор независимо предсказанных классов для соседних столбцов.

Указанные зависимости используются в модели глубокого машинного обучения «Sato», которая обучалась на коллекции таблиц «WebTables» (около 80 тысяч реляционных веб-таблиц) из проекта «VizNet». В отличие от фреймворка «ColNet», «Sato» не требует предварительного извлечения конкретных сущностей из графа знаний на основе ячеек таблицы и работает напрямую с табличными данными. Потенциально, этот факт делает подход «Sato» более перспективным, хотя он требует большого количества табличных данных для обучения.

3. Метод связности классов. Данный метод используется для корректировки выбора релевантных классов после применения всех остальных методов аннотирования. Метод основан на идеях вероятностной графовой модели (probabilistic graphical model), в частности, неориентированной модели в виде Марковских случайных полей, которые включают использование переменных узлов для представления текущих состояний аннотаций столбцов таблицы. Каждый такой переменный узел имеет следующие свойства (параметры):

- «*семантический тип*» – это текущий класс из набора кандидатов, который выбирается в качестве целевой аннотации на основе применения всех методов аннотирования;
- «*оценка*» – это итоговая оценка, присвоенная данному классу-кандидату на основе применения всех методов аннотирования;
- «*столбец*» – номер столбца;
- «*связность*» – параметр, показывающий уровень связанности текущего переменного узла с другими (количество отношений между текущими выбранными классами в качестве аннотаций);
- «*изменение*» – параметр, указывающий, следует ли менять данный переменный узел на текущей итерации.

Перед началом работы метода каждый класс из набора кандидатов сортируется по убыванию агрегированной оценки, которая считается путем суммирования всех оценок, полученных в результате применения всех остальных методов аннотирования. Далее осуществляется процедура инициализации переменных узлов. В частности, первый класс из отсортированного набора кандидатов присваивается параметру «семантический тип», а его агрегированная оценка присваивается параметру «оценка» в переменной узла. Параметр «связность» по умолчанию равен нулю, а параметр «изменение» выставлен в состояние «True» для всех переменных узлов. Далее делается попытка найти такую комбинацию переменных узлов, где их «связность» по возможности будет ненулевой, но при этом также будет учитываться и их «оценка». Для этого переменные узлы итеративно обновляются следующими классами из набора кандидатов. В конечном итоге класс-кандидат с наивысшим параметром «связности» и «оценкой» определяется как наиболее подходящий (релевантный) класс и назначается текущему столбцу в качестве итоговой аннотации.

4. Программная реализация. Модифицированный подход реализован в виде программного средства «TabbyLD2» [5], написанного на языке Python и включающего в себя следующие основные программные модули:

- *модель данных* (*datamodel*) – содержит объектные модели Python для описания табличных данных и элементов целевого графа знаний;
- *модуль предобработки* (*preprocessing*) – реализует этап очистки табличных данных, атомарную классификацию столбцов и идентификацию сущностного столбца;
- *семантический аннотатор* (*table_annotation*) – реализует этапы САЯ, САС и САО, а также содержит генератор RDF-триплетов и модуль для взаимодействия с целевыми графами знаний, включая сервисы «DBpedia SPARQL Endpoint» и «DBpedia Lookup»;
- *вспомогательный модуль* (*helpers*) – содержит различные полезные функции для работы с файлами, данными и т.д.;
- *модуль экспериментальной оценки* (*experimental_evaluation*) – содержит сценарии для запуска экспериментов на различных тестовых наборах данных.

Разработанное средство имеет два режима работы:

- *консольный режим* – запускается через командную строку и является основным режимом для обработки таблиц;
- *веб-режим* – запускается в виде веб-сервера и использует программный интерфейс взаимодействия через REST API для доступа к основным функциям семантического аннотатора.

В дополнение к основному программному средству разработано веб-ориентированное клиентское приложение «TabbyLD2-Client» [22], предоставляющее пользовательский графический интерфейс доступа к функциям СИТ и генерации RDF-триплетов, реализованных в «TabbyLD2». Клиентское веб-приложение разработано с использованием языка PHP и фреймворка Yii2 на основе паттерна проектирования «Model-View-Controller» и ориентированно на непрограммирующих пользователей (например, предметных экспертов, аналитиков данных, инженеров по знаниям).

Клиентское веб-приложение реализует следующие основные функции:

- импорт табличных данных в формате CSV;
- автоматическая атомарная классификация столбцов на категориальный и литеральный типы, а также идентификация сущностного столбца (средствами «TabbyLD2»);
- автоматическое аннотирование ячеек, столбцов и отношений между столбцами (средствами «TabbyLD2»);
- просмотр загруженной таблицы и результатов её предобработки и аннотирования;
- изменение (модификация) автоматического составления аннотаций (пользователь может выбрать из списка найденных кандидатов необходимую сущность, класс или свойство);
- генерация предметных графов знаний в формате RDF на основе аннотированных табличных данных (средствами «TabbyLD2»).

5. Примеры практического применения. Разработанное программное обеспечение было использовано при решении следующих прикладных задач.

5.1. Проект Talisman. В сотрудничестве с Институтом системного программирования имени В.П. Иванникова Российской академии наук (ИСП РАН) решалась задача формирования различных предметно-ориентированных графов знаний и их пополнения конкретными сущностями (фактами), извлеченными из таблиц. Данные графы знаний представляют собой семантические ориентированные графы, доступ к которым предоставляется через интерфейс

«GraphQL». Графы знаний разрабатывались в рамках платформы «Talisman» [23], которая объединяет различные программные компоненты «Big Data» и использует в качестве основных сервисов такие технологии ИСП РАН, как «Dedoc» (система извлечения структуры документов) и «Texterra» (система извлечения семантики из текстов). Таким образом, платформа «Talisman» представляет собой интеллектуальную систему для структурирования и формализации знаний предметной области, извлеченных из разных неструктурированных информационных источников (например, Веб или документы), на основе которых в дальнейшем могут решаться задачи аналитики.

Демонстрационный стенд с платформой «Talisman» был развернут на сервере в Институте динамики систем и теории управления имени В.М. Матросова СО РАН (https://isdct.talisman.ispras.ru/). В рамках данного стенда была разработана онтологическая схема (рис. 2), представляющая собой основу для предметного графа знаний и содержащая только основные понятия и отношения из области анализа рынка труда Иркутской области (например, вакансии, персоны, компании, учебные заведения, навыки). Созданная онтологическая схема наполнялась конкретными сущностями (фактами) извлеченными из веб-ресурсов, документов и, в частности, таблиц. Всего было извлечено больше 100 тысяч конкретных сущностей.

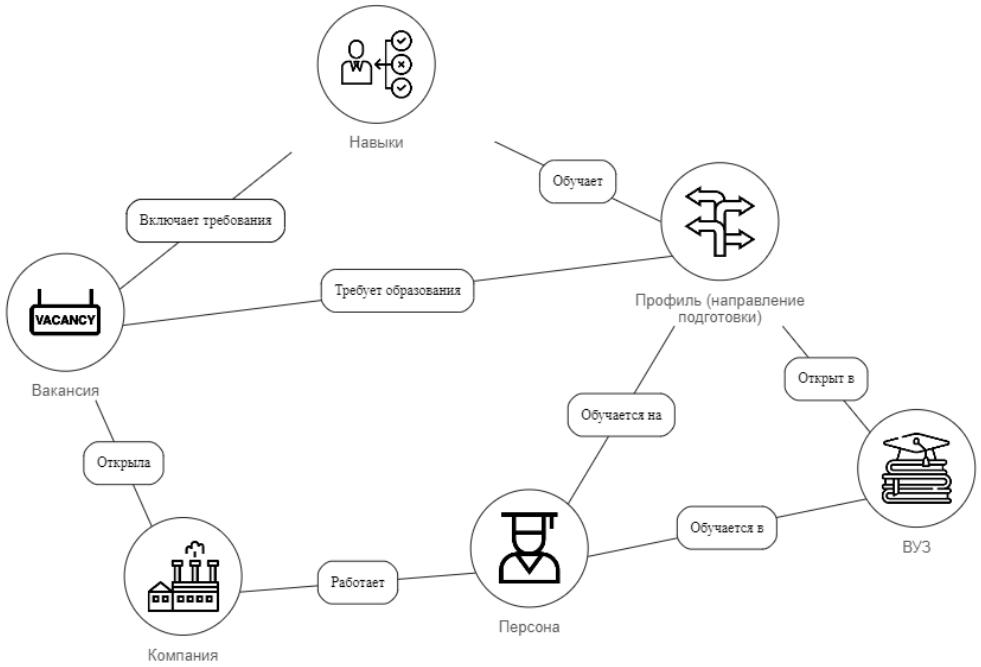

Рис. 2. Онтологическая схема, описывающая область анализа рынка труда Иркутской области на стенде платформы «Talisman»

Отдельно проводился эксперимент, целью которого было показать принципиальную возможность использования предлагаемого подхода и разработанного средства для СИТ, в частности, для решения задачи САС. Для этого в качестве тестового набора данных был сформирован документ, содержащий 8 таблиц. Каждая таблица описывает вакансии, открытые в Иркутской области по категории «информационные технологии». Табличные данные собирались вручную на различных веб-ресурсах: hh.ru, superjob.ru, rabota.ru, avito.ru, zarplata.ru, irk.rosrabota.ru, а также банков вакансий Иркутской области, ГАУ «Иркутский областной многофункциональный центр предоставления государственных и муниципальных услуг (МФЦ)». Для получения экспериментальной оценки использовалась хорошо известная мера точности (*accuracy*):

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

где *CA* – это количество корректно аннотированных столбцов, т.е. столбцы, которым верно были присвоены соответствующие классы (*типы концептов*) или типы данных (*типы значений характеристик*); *C* – общее количество столбцов в таблице.

Результаты оценки точности для двух базовых методов («*голосование методом простого большинства*» и «*сходство по заголовку*»), а также после уточнения аннотаций на основе метода связности классов представлены в таблице 2. Полученные результаты показывают перспективность использования разработанного подхода и средства для поддержки процесса инженерии предметно-ориентированных графов знаний.

Таблица 2. Экспериментальная оценка задачи CAC для тестовых таблиц

Название таблицы	Accuracy (два базовых метода)	Accuracy (два базовых метода + метод связности классов)
Банк вакансий Иркутской Обл.	0,833	0,833
Работа на hh.ru	1,000	1,000
Работа на superjob.ru	0,750	0,750
Работа на rabota.ru	0,857	0,857
Работа на www.avito.ru	1,000	1,000
Работа на zarplata.ru	0,667	0.778
Работа на Росработа	1,000	1,000
Гос. авто. уч. Иркутский МФЦ	0,778	0,778
Итоговая оценка	0,860	0.874

5.2. Экспертиза промышленной безопасности. В рамках пилотного проекта создания системы поддержки принятия решений при Экспертизе Промышленной Безопасности (ЭПБ) в нефтехимии для Иркутского научно-исследовательского и проектного института химического и нефтехимического машиностроения (АО ИркутскНИИхиммаш) также решалась задача создания и пополнения графов знаний. ЭПБ – это процедура оценки технического состояния различного рода оборудования. База знаний разработанной системы формировалась в течение нескольких лет, при этом использовались знания экспертов, концептуальные модели, а также автоматический анализ отчетов по ЭПБ, которые содержат разнородную информацию в виде текстов, диаграмм, графиков и таблиц. В большинстве случаев именно таблицы являются наиболее перспективным источником для автоматического извлечения информации и наполнения баз знаний.

В предыдущих работах [4, 18] рассматривалась задача заполнения предметно-ориентированного графа знаний на основе табличных данных, извлеченных из отчетов по ЭПБ. Разработанный граф знаний содержит информацию о различном нефтехимическом оборудовании, исследуемом в процессе ЭПБ. В этой статье представлено расширение созданного графа знаний на основе данных различных измерений. Для решения текущей задачи из отчетов по ЭПБ были извлечены 70 таблиц, представляющих результаты измерений технического состояния нефтехимического оборудования. В таблице 3 представлен фрагмент, описывающий результаты измерений резервуара. Все исходные таблицы были семантически аннотированы на основе разработанной онтологической схемы (рис. 3).

Процесс семантического аннотирования для этих таблиц осуществлялся автоматически с использованием предлагаемых подхода и средства. В частности, для решения задачи CAC использовались все методы аннотирования, кроме «*голосования методом простого большинства*» и метода предсказания класса на основе фреймворка «ColNet», поскольку в разработанной онтологической схеме изначально отсутствовали конкретные сущности, описывающие

рассматриваемую предметную область. Аннотации для отношений между столбцами таблицы (например, между сущностным столбцом и другими категориальными или литеральными столбцами) были выведены автоматически с использованием полученных точных аннотаций столбцов. Следует отметить, что не все аннотации для столбцов и отношений между ними были установлены таким образом. Поэтому данные аннотации были уточнены вручную экспертами предметной области.

Таблица 3. Фрагмент, описывающий результаты измерений резервуара

Наименование определяемой величины	Единицы измерения	Расчетная формула или обозначение	Числовое значение
Остаточный ресурс	лет	Тк	20
Фактическая толщина стенки обечайки с учетом минусового допуска	мм	Sф	7,3
Минимально допустимая толщина стенки элемента	мм	So	4
Исполнительная толщина стенки обечайки	мм	Sи	8
Время эксплуатации	лет	Т1	27
Расчетный срок службы	лет	Тир	10
Коэффициент, учитывающий отличие средней ожидаемой скорости коррозии с доверительной вероятностью 0,75-0,9		К1	0,5
Коэффициент, учитывающий погрешность определения скорости коррозии по линейному закону, от скорости коррозии, рассчитанной по более точным (нелинейным) законам изменения контролируемого параметра		К2	0,75
Скорость коррозии	мм/год	а	0,1

Рис. 3. Фрагмент онтологической схемы, описывающий процесс измерения

Таким образом, 636 уникальных конкретных сущностей были извлечены из строк всего набора таблиц с использованием определенных аннотаций для столбцов и их взаимосвязей. Извлеченные сущности пополнили целевую онтологическую схему, в результате чего был получен предметно-ориентированный граф знаний для процедуры ЭПБ, описывающий результаты измерений технического состояния нефтехимического оборудования. Пример экранной

формы отображения таблицы с определенными атомарными типами и семантическими аннотациями для столбцов представлен на рисунке 4.

Заключение. Таблицы являются распространенным способом представления информации и повсеместно используются в различных предметных областях. Эффективное использование таблиц требует разработки специализированных средств для автоматической семантической интерпретации их содержимого. Это особенно актуально при решении реальных практических задач.

☐ Сущностный (тематический) столбец ☐ Категориальный столбец ☐ Литеральный столбец

Показаны записи **1-9** из **9**.

#	Наименование определяемой величины [Измерения]	Единицы измерения | [ЕдиницыИзмерения]	Расчетная формула или обозначение | [ФормулаРасчета]	Числовое значение | [xsd:positiveInteger]
1	Остаточный ресурс	лет	Тк	20
2	Фактическая толщина стенки обечайки с учетом минусового допуска	мм.	Sф	7,3
3	Минимально допустимая толщина стенки элмента	мм.	So	4
4	Исполнительная толщина стенки обечайки	мм.	Si	8
5	Время эксплуатации	лет	t1	27
6	Расчетный срок службы	лет	Тир	10
7	Коэффициент, учитывающий отличие средней ожидаемой скорости коррозии от гарантированной скорости коррозии с доверительной вероятностью 0.75-0.9	None	K1	0,5
8	Коэффициент, учтитывающий погрешность определения скорости коррозии по линейному закону, от скорости коррозии, расчитанной по более точным (нелинейным) законам изменения контролируемого параметра	None	K2	0,75
9	Скорость коррозии	мм./год	a	0,1

Рис. 4. Пример экранной формы отображения таблицы с определенными атомарными типами и семантическими аннотациями для столбцов

В статье был предложен подход для автоматизированного извлечения конкретных сущностей (фактов) из семантически аннотированных таблиц и пополнения ими целевого графа знаний. Данный подход был реализован в виде двух взаимодействующих между собой программных средств: «TabbyLD2» и «TabbyLD2-Client». Разработанные программные средства были успешно применены при решении практических задач пополнения предметно-ориентированных графов знаний. Полученная экспериментальная оценка показала перспективность использования предлагаемого подхода и программного обеспечения. В будущем планируется изучить возможность использования предварительно обученных языковых моделей для более улучшенной семантической аннотации таблиц.

Благодарности. 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» (№ гос. регистрации: 121030500071-2).

Список источников

1. Hogan A., Blomqvist E., Cochez M. Knowledge Graphs, 2021.
2. Villazon-Terrazas B., Garcia-Santa N., Ren Y. 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. A large public corpus of web tables containing time and context metadata. Proceedings of the 25th International conference companion on World Wide Web, 2016, pp. 75-76.

4. Дородных Н.О. Подход к автоматизированному наполнению графов знаний сущностями на основе анализа таблиц / Н.О. Дородных, А.Ю. Юрин // Онтология проектирования, 2022. – Т.12. – № 3. – С. 336-352.
5. TabbyLD2. Available at: https://github.com/tabbydoc/tabbyld2 (accessed: 04/11/2023).
6. Bonfitto S., Casiraghi E., Mesiti M. Table understanding approaches for extracting knowledge from heterogeneous tables. Wiley interdisciplinary reviews: Data mining and knowledge discovery, 2021, vol.11, no. 4, e1407.
7. Liu J., Chabot Y., Troncy R. From tabular data to knowledge graphs: A survey of semantic table interpretation tasks and methods. Journal of Web Semantics, 2023, vol. 76, 100761.
8. Efthymiou V., Hassanzadeh O., Rodriguez-Muro M. Matching web tables with knowledge base entities: From entity lookups to entity embeddings. Proceedings of 16th International semantic Web Conference (ISWC'2017), Lecture notes in computer science, 2017, vol. 3316, pp. 260-277.
9. Xie J., Lu Y., Cao C. Joint Entity Linking for Web tables with hybrid semantic matching. Proceedings of the International Conference on Computational Science, Lecture notes in computer science, 2020, vol. 12138, pp. 618-631.
10. Hulsebos M., Hu K., Bakker M. Sherlock: A Deep learning approach to semantic data type detection. KDD'19: Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, 2019, pp. 1500-1508.
11. Zhang D., Suhara Y., Li J. Sato: Contextual semantic type detection in tables. Proceedings of the VLDB Endowment, 2020, vol. 13, no.11, pp. 1835-1848.
12. Chen J., Jimenez-Ruiz E., Horrocks I. ColNet: Embedding the semantics of web tables for column type prediction. Proceedings of the AAAI Conference on Artificial Intelligence, 2019, vol. 33, no.1, pp. 29-36.
13. Suhara Y., Li J., Li Y. Annotating columns with pre-trained language models. SIGMOD'22: Proceedings of the 2022 International Conference on Management of Data, 2022, pp. 1493-1503.
14. Trabelsi M., Cao J., Heflin J. SeLaB: Semantic labeling with BERT. Proceedings of the International Joint Conference on Neural Networks (IJCNN), 2021, pp. 1-8.
15. Deng X., Sun H., Lees A. TURL: Table understanding through representation learning. Proceedings of the VLDB Endowment, 2020, vol. 14, no. 3, pp. 307-319.
16. Yin P., Neubig G., Yih W. TaBERT: Pretraining for joint understanding of textual and tabular Data. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 2020, pp. 8413-8426.
17. Bizer C., Lehmann J., Kobilarov G. DBpedia A Crystallization point for the Web of data. Journal of Web Semantics, 2009, vol. 7, no. 3, pp. 154-165.
18. Dorodnykh N.O., Yurin A.Yu. Knowledge graph augmentation based on tabular data: a case study for industrial safety inspection. Proceedings of the Sixth International Scientific Conference "Intelligent Information Technologies for Industry" (IITI'22), Lecture notes in networks and systems, 2022, vol. 566, pp. 314-324.
19. Stanford CoreNLP. Available at: https://stanfordnlp.github.io/CoreNLP/ (accessed: 04/11/2023).
20. Duckling. Available at: https://github.com/facebook/duckling (accessed: 04/11/2023).
21. Dateparser. Available at: https://dateparser.readthedocs.io/en/latest/ (accessed: 04/11/2023).
22. TabbyLD2-Client. Available at: https://github.com/tabbydoc/tabbyld2_client (accessed: 04/11/2023).
23. Talisman. Available at: https://talisman.ispras.ru/ (accessed: 04/11/2023).

Амирасланов Ильгар Вугарович. *Студент Института информационных технологий и анализа данных Иркутского национального исследовательского технического университета (ИрНИТУ). Основные направления исследований: семантическая интерпретация таблиц. ilgar-amiraslanov@mail.ru.*

Дородных Никита Олегович. *К.т.н., старший научный сотрудник Института динамики систем и теории управления им. В.М. Матросова СО РАН (ИДСТУ СО РАН). Основные направления исследований: автоматизация создания интеллектуальных систем и баз знаний, получение знаний из документов, таблиц, концептуальных моделей, семантическая интерпретация 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. *Д.т.н., заведующий лабораторией Информационных технологий исследования природной и техногенной безопасности ИДСТУ СО РАН, доцент Института информационных технологий и анализа данных ИрНИТУ. Основные направления исследований: разработка систем интеллектуальных систем и баз знаний, использование 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.*

UDC 004.89
DOI:10.25729/ESI.2023.30.2.014

A tool for extraction of entities from semantically annotated tabular data

Ilgar V. Amiraslanov[1], Nikita O. Dorodnykh[2], Alexander Yu. Yurin[1,2]

[1]Irkutsk National Research Technical University, Russia, Irkutsk, *iskander@icc.ru*
[2]Matrosov Institute for System Dynamics and Control Theory SB RAS, Russia, Irkutsk

Abstract. Today, knowledge graphs are widely used in various domains, for example, in industry, commerce, finance, and social networks. A knowledge graph is a powerful means of information combination and representation using standardized knowledge modelling methods. However, the development of knowledge graphs and, in particular, their population with new specific entities (facts) remains a difficult task. The use of various information sources can facilitate this process. Such a source can be tables that potentially contain rich semantic information. In this paper proposes an approach and its software implementation for automated extraction of significant information from tabular data in the form of facts and population of a target knowledge graph with them. The main feature of the proposed approach is the combination of heuristic methods with deep machine learning models for semantic table annotation. The applicability of the proposed approach is demonstrated by two examples: labor market analyzing for the Irkutsk region and assessing the technical state of petrochemical equipment.

Keywords: semantic web, knowledge acquisition, knowledge graph, semantic table interpretation, fact extraction, knowledge graph population, table

Acknowledgements: 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").

References

1. Hogan A., Blomqvist E., Cochez M. Knowledge Graphs, 2021.
2. Villazon-Terrazas B., Garcia-Santa N., Ren Y. 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. A large public corpus of web tables containing time and context metadata. Proceedings of the 25th International Conference Companion on World Wide Web, 2016, pp. 75-76.
4. Dorodnykh N.O., Yurin A.Yu. Podhod k avtomatizirovannomu napolneniyu grafov znanij sushchnostyami na osnove analiza tablic [An approach for automated knowledge graph population with entities based on table analysis]. Онтология проектирования [Ontology of designing], 2022, vol.12, no.3, pp. 336-352.
5. TabbyLD2. Available at: https://github.com/tabbydoc/tabbyld2 (accessed: 04/11/2023).
6. Bonfitto S., Casiraghi E., Mesiti M. Table understanding approaches for extracting knowledge from heterogeneous tables. Wiley interdisciplinary reviews: Data mining and knowledge discovery, 2021, vol.11, no.4, e1407.
7. Liu J., Chabot Y., Troncy R. From tabular data to knowledge graphs: A survey of semantic table interpretation tasks and methods. Journal of Web Semantics, 2023, vol.76, 100761.
8. Efthymiou V., Hassanzadeh O., Rodriguez-Muro M. Matching web tables with knowledge base entities: From entity lookups to entity embeddings. Proceedings of 16th International semantic Web Conference (ISWC'2017), Lecture notes in computer science, 2017, vol.3316, pp. 260-277.
9. Xie J., Lu Y., Cao C. Joint Entity Linking for Web tables with hybrid semantic matching. Proceedings of the International Conference on Computational Science, Lecture notes in computer science, 2020, vol.12138, pp. 618-631.
10. Hulsebos M., Hu K., Bakker M. Sherlock: A Deep learning approach to semantic data type detection. KDD'19: Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, 2019, pp. 1500-1508.
11. Zhang D., Suhara Y., Li J. Sato: Contextual semantic type detection in tables. Proceedings of the VLDB Endowment, 2020, vol.13, no.11, pp. 1835-1848.
12. Chen J., Jimenez-Ruiz E., Horrocks I. ColNet: Embedding the semantics of web tables for column type prediction. Proceedings of the AAAI Conference on Artificial Intelligence, 2019, vol.33, no.1, pp. 29-36.
13. Suhara Y., Li J., Li Y. Annotating columns with pre-trained language models. SIGMOD'22: Proceedings of the 2022 Inter-national Conference on Management of Data, 2022, pp. 1493-1503.

14. Trabelsi M., Cao J., Heflin J. SeLaB: Semantic labeling with BERT. Proceedings of the International Joint Conference on Neural Networks (IJCNN), 2021, pp. 1-8.
15. Deng X., Sun H., Lees A. TURL: Table understanding through representation learning. Proceedings of the VLDB Endowment, 2020, vol.14, no.3, pp. 307-319.
16. Yin P., Neubig G., Yih W. TaBERT: Pretraining for joint understanding of textual and tabular Data. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 2020, pp. 8413–8426.
17. Bizer C., Lehmann J., Kobilarov G. DBpedia A Crystallization point for the Web of data. Journal of Web Semantics, 2009, vol.7, no.3, pp. 154-165.
18. Dorodnykh N.O., Yurin A.Yu. Knowledge graph augmentation based on tabular data: a case study for industrial safety inspection. Proceedings of the Sixth International Scientific Conference “Intelligent Information Technologies for Industry” (IITI’22), Lecture notes in networks and systems, 2022, vol.566, pp. 314-324.
19. Stanford CoreNLP. Available at: https://stanfordnlp.github.io/CoreNLP/ (accessed: 04/11/2023).
20. Duckling. Available at: https://github.com/facebook/duckling (accessed: 04/11/2023).
21. Dateparser. Available at: https://dateparser.readthedocs.io/en/latest/ (accessed: 04/11/2023).
22. TabbyLD2-Client. Available at: https://github.com/tabbydoc/tabbyld2_client (accessed: 04/11/2023).
23. Talisman. Available at: https://talisman.ispras.ru/ (accessed: 04/11/2023).

Amiraslanov Ilgar Vugarovich. *Student of the Institute of Information Technologies and Data Analysis, Irkutsk National Research Technical University (INRTU). Main research domains: semantic table interpretation. ilgar-amiraslanov@mail.ru.*

Dorodnykh Nikita Olegovych. *Ph.D., senior associate researcher at Matrosov Institute for System Dynamics and Control Theory SB RAS (ISDCT SB RAS). Main research domains: computer-aided development of intelligent systems and knowledge bases, knowledge acquisition based on documents, tables and conceptual models, semantic table interpretation. ORCID: 0000-0001-7794-4462; Author ID (RSCI): 979843; Author ID (Scopus): 57202323578; Researcher ID (WoS): E-8870-2014, tualatin32@mail.ru.*

Yurin Alexander Yurievich. *Ph.D., head of a laboratory "Information and telecommunication technologies for investigation of natural and technogenic safety" at ISDCT SB RAS and associate professor of INRTU. Main research domains: development of intelligent 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.*

Статья поступила в редакцию 12.04.2023; одобрена после рецензирования 28.04.2023; принята к публикации 10.05.2023.
The article was submitted 04/12/2023; approved after reviewing 04/28/2023; accepted for publication 05/10/2023.