ИНЖИНИРИНГ ОНТОЛОГИЙ

УДК 004.89 *Научная статья* DOI: 10.18287/2223-9537-2024-14-4-555-568

Разработка предметных графов знаний на основе семантического аннотирования табличных данных

© 2024, Н.О. Дородных ✉, А.Ю. Юрин

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

Аннотация

В статье описывается подход и программное средство для автоматизированного пополнения предметно-ориентированных графов знаний новыми фактами, извлечёнными из семантически аннотированных табличных данных. Для семантического аннотирования столбцов таблиц предлагается использовать комбинацию из трёх эвристических методов, использующих результаты распознавания именованных сущностей в ячейках, лексическое сопоставление и группировку характеристик. Предлагаемый подход реализован в виде специального обработчика, входящего в состав программной платформы *Talisman*. Представлен пример и экспериментальная оценка предлагаемого подхода на этапе семантического аннотирования столбцов с использованием тестового набора табличных данных, который включает шесть тематических категорий: «сотрудники организации», «открытые вакансии», «рынок автомоделей», «известные учёные», «продажа книг», «рейтинг теннисистов». В качестве метрик оценки использовались точность, полнота и *F*-мера. Итоговая оценка по всем шести категориям составила: точности – 79%, полноты – 63%, *F*-меры – 70%. Полученные результаты показывают перспективность использования разработанного подхода для пополнения предметно-ориентированных графов знаний новыми фактами, извлечёнными из семантически аннотированных табличных данных. Приведены ограничения предлагаемого подхода.

Ключевые слова: *граф знаний, семантическая интерпретация таблиц, аннотирование таблиц, извлечение сущностей, пополнение графа знаний, табличные данные.*

Цитирование: *Дородных Н.О., Юрин А.Ю.* Разработка предметных графов знаний на основе семантического аннотирования табличных данных. *Онтология проектирования*. 2024. Т.14, №4(54). С.555-568. DOI:10.18287/2223-9537-2024-14-4-555-568.

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» (№ госрегистрации: 121030500071-2).

Конфликт интересов: авторы заявляют об отсутствии конфликта интересов.

Введение

Интеллектуальные информационно-аналитические системы активно применяются в сфере корпоративного поиска информации (например, *Microsoft SharePoint*[1], *Oracle Secure Enterprise Search*[2], *Elasticsearch*[3] и др.), ведения баз знаний и анализа текстов (*Palantir*

[1] https://www.microsoft.com/ru-ru/microsoft-365/sharepoint/collaboration.
[2] https://www.oracle.com/middleware/technologies/oses-downloads.html.
[3] https://github.com/elastic/elasticsearch/releases/tag/v8.15.0.

Gotham[4], *IQPlatform*[5], *Айтеко «X-files 2.0»*[6] и др.), мониторинга СМИ и социальных сетей (*LexisNexis*[7], *Медиалогия*[8], *BrandAnalytics*[9] и др.), конкурентной разведки (*Maltego*[10], *Hensoldt Analytics*[11], *Виток-OSINT*[12] и др.), прогнозирования и аналитики данных (*SAS Analytics*[13], *IBM Watson Studio*[14], *PolyAnalyst Megaputer*[15] и др.).

Для построения подобного рода систем могут быть использованы графы знаний (ГЗ), предназначенные для накопления и передачи знаний о реальном мире, при этом их узлы представляют интересующие объекты, а рёбра – отношения между этими объектами [1, 2]. Базовой структурной единицей ГЗ является триплет: <*субъект*>, <*предикат*>, <*объект*>. Каждая подобная сущность из этого триплета идентифицируется глобальным унифицированным идентификатором ресурса (*Uniform Resource Identifier, URI*) [3]. ГЗ могут быть масштабированы для обработки больших объёмов данных. ГЗ можно разделить на два типа: *глобальные кросс-доменные* ГЗ и *предметно-ориентированные* ГЗ. Первый тип включает такие международные проекты с открытым исходным кодом, как *DBpedia*[16], *Wikidata*[17], *Yago*[18], *BabelNet*[19] и проприетарные решения, такие как *Google Knowledge Graph*[20] и *Probase*[21]. Такие графы, как правило, содержат большое количество объектов из многих областей. Второй тип ориентируется на описание знаний, которые относятся к определённой конкретной области или предприятию. Предметные ГЗ могут поддерживать эффективный поиск знаний и являться основой для различных приложений [2, 3]. Использование ГЗ при построении интеллектуальных систем позволяет эффективно структурировать знания и выявлять скрытые связи и зависимости между различными понятиями, что бывает полезно для принятия решений или прогнозирования [4]. Однако разработка ГЗ является трудоёмкой задачей и может потребовать обработки больших объёмов данных, полученных из различных информационных источников (например, баз данных, электронных документов, веб-ресурсов) [5, 6]. Таким образом, исследования, ориентированные на автоматизацию построения ГЗ и пополнения их новыми фактами при решении практических, слабо формализованных задач в различных предметных областях (ПрО), являются актуальными.

Основной тенденцией здесь является использование различных информационных источников. Одним из таких источников являются таблицы [7]. В общем случае каждая строка таблицы представляет собой запись, а каждый столбец – атрибут или поле. Согласно [8] из таблиц, содержащихся как в веб-пространстве, так и в составе различных электронных документов, можно извлечь множество полезных фактов. Таблицы неоднородны по своей структуре и не сопровождаются явной семантикой, необходимой для автоматической интерпретации своего содержания. Это затрудняет активное практическое использование табличных данных (ТД) в автоматическом и автоматизированных режимах.

[4] https://www.palantir.com/platforms/gotham/.
[5] https://iqmen.ru/iqplatform.
[6] https://www.i-teco.ru/iskusstvennyyintellekt/x-files-2-0/.
[7] https://www.lexisnexis.com/en-us/gateway.page.
[8] https://www.mlg.ru/.
[9] https://brandanalytics.ru/.
[10] https://www.maltego.com/.
[11] https://analytics.hensoldt.net/.
[12] https://norsi-trans.ru/catalog/osint/vitok-m/.
[13] https://www.sas.com/en_in/home.html.
[14] https://www.ibm.com/products/watson-studio.
[15] https://www.megaputer.ru/.
[16] https://www.dbpedia.org/.
[17] https://www.wikidata.org/wiki/Wikidata:Main_Page.
[18] https://yago-knowledge.org/.
[19] https://babelnet.org/.
[20] https://blog.google/products/search/introducing-knowledge-graph-things-not/.
[21] https://www.microsoft.com/en-us/research/project/probase/.

Подход к автоматизированному наполнению ГЗ сущностями на основе анализа таблиц был предложен в [9]. Особенностью этого подхода является возможность автоматического восстановления семантики ТД на основе применения гибридного метода, сочетающего в себе техники машинного обучения, векторных представлений и интуитивно понятных эвристик.

В данной работе предлагается специализировать предложенный общий подход к конкретной практической задаче извлечения фактов из ТД в рамках индустриальной цифровой платформы *Talisman*[22], разработанной Институтом системного программирования имени В.П. Иванникова Российской академии наук (ИСП РАН). Платформа *Talisman* представляет собой набор связанных программных инструментов для автоматизации типовых задач обработки данных (сбор, интеграция, анализ, хранение, визуализация). Платформа обеспечивает быструю разработку аналитических систем, объединяющих информацию из внутренних баз данных и открытых источников сети Интернет.

1 Состояние исследований

Автоматическое создание предметно-ориентированных ГЗ и пополнение их новыми фактами невозможно без автоматического распознавания структуры и содержания ТД. Восстановлением подобного рода семантики занимается такое научное направление как *семантическая интерпретация* (*аннотирование*) таблиц [10]. Первые работы в данной области были направлены на сопоставление отдельных элементов таблиц с понятиями из ГЗ, онтологии или другого внешнего словаря [11, 12]. Семантическая интерпретация таблиц включает в себя четыре основные задачи [10]:

- *аннотирование ячеек* – сопоставление значений ячеек с сущностями (экземплярами классов) из целевого ГЗ (ЦГЗ);
- *аннотирование столбцов* – сопоставление отдельных столбцов таблицы с семантическими типами (классами) из ЦГЗ;
- *аннотирование отношений между столбцами* – сопоставление связей между столбцами со свойствами (предикатами) из ЦГЗ;
- *аннотирование таблицы* – сопоставление всей таблицы целиком с некоторым классом из ЦГЗ (обнаружение темы таблицы).

Развитие исследований в этой области можно разделить на два основных этапа:

Этап 1 (*2010 – 2019 гг.*). На данном этапе осуществлялась общая формулировка проблемы семантической интерпретации таблиц, определялись основные цели и задачи. Этап характеризуется появлением работ, направленных в основном на анализ естественно-языкового содержания и контекста таблиц с использованием методов сопоставления онтологий, поиска сущностей (как в глобальных ГЗ, так и в предметно-ориентированных онтологиях), связывания сущностей с элементами Википедии и представления в векторном пространстве сущностей ГЗ [13-16]. Здесь можно отметить итерационные методы на основе использования вероятностных графовых моделей [17, 18] и подходы на основе методов машинного обучения [17, 19, 20].

Этап 2 (*2019 г. – по настоящее время*) характеризуется ростом числа работ и получением результатов для отдельных задач семантической интерпретации таблиц. Появляются коммерческие решения по определению семантического типа столбцов таблиц, расширяющие функциональность систем подготовки и анализа данных, таких как *Microsoft Power BI*[23], *Trifacta*[24] и *Google Looker Studio*[25]. На данном этапе большую популярность получили подходы, основанные на глубоком машинном обучении (например, *JHSTabEL* [21], *Sato* [22]), в т.ч. с использованием предобученных языковых моделей (например, *TURL* [23], *TaBERT* [24], *TABBIE* [25] и др.). С 2019 года ежегодно проходит соревнование *SemTab*[26], направленное на выявление решений для сопоставления ТД с ГЗ, в рамках которого сформулированы основные метрики и критерии оценки систем аннотирования

[22] http://talisman.ispras.ru.
[23] https://powerbi.microsoft.com.
[24] https://www.trifacta.com.
[25] https://lookerstudio.google.com.
[26] http://www.cs.ox.ac.uk/isg/challenges/sem-tab/.

таблиц. Кроме того, появляется множество открытых наборов данных для тестирования производительности таких систем (например, *WebTables*[27], *WikiTables*[28] и др.).

Таким образом, за последние годы достигнуты значительные успехи в области исследований по автоматическому пониманию ТД. Однако наблюдается разрыв между эффективностью существующих решений на тестах и их применимостью на практике. Это обусловлено отсутствием качественных наборов размеченных обучающих данных и сложностью настройки существующих моделей, подходов и систем для конкретных ПрО и задач. В большинстве подходов отсутствует этап извлечения новых фактов из семантически аннотированных ТД и пополнения ими ЦГЗ. Это обуславливает актуальность разработки методологического и программного обеспечения, направленного на комплексное решение задач семантической интерпретации таблиц и извлечения фактов в конкретных ПрО.

2 Предлагаемый подход

2.1 Существующий задел

В работе [9] предложен подход к автоматическому извлечению конкретных сущностей (фактов) из таблиц и наполнению ими ЦГЗ. Особенностью этого подхода является возможность поддержки автоматизированного восстановления семантики таблиц на основе модели ПрО (онтологии на терминологическом уровне – *TBox*). Благодаря этому возможно задавать явную семантическую аннотацию для отдельных элементов таблицы (столбцов и связей между ними) и извлекать конкретные сущности из ячеек. При этом подход позволяет решить две задачи семантической интерпретации таблиц: аннотирование столбцов и аннотирование отношений между столбцами. Подход имеет ряд ограничений: ориентирован на обработку только реляционных таблиц, представленных в формате *CSV*; использует ГЗ общего назначения *DBpedia* для аннотирования исходных таблиц.

На рисунке 1 представлена схема, иллюстрирующая пример семантического аннотирования таблицы и извлечения конкретных сущностей (фактов) из её строки. В примере использована таблица международного рейтинга Ассоциации теннисистов-профессионалов (*ATP*).

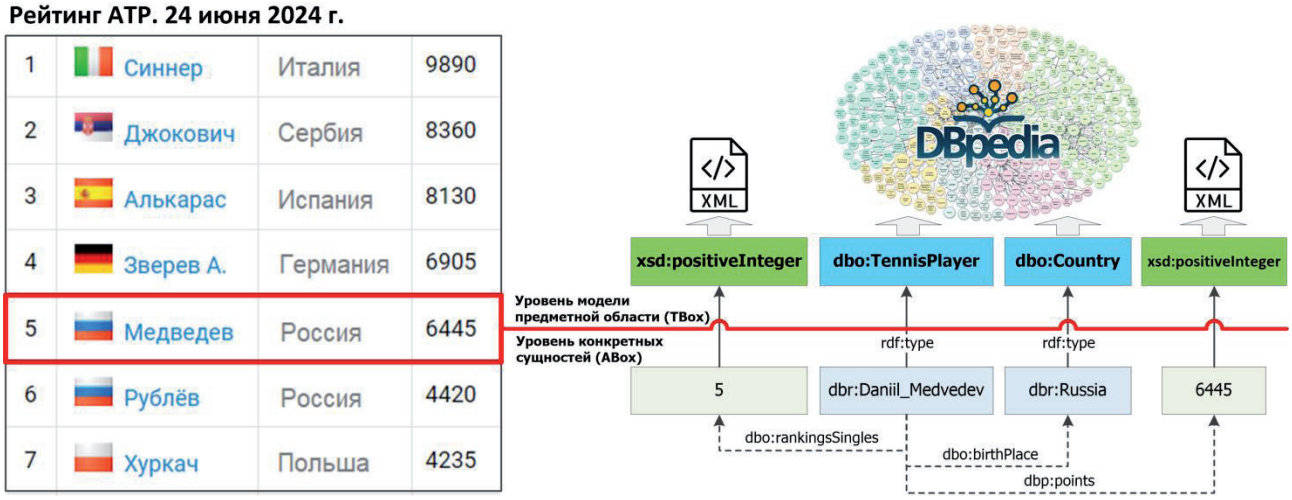


Рисунок 1 – Схема семантического аннотирования таблицы и извлечения фактов на основе подхода из [9]

[27] https://webdatacommons.org/webtables/.
[28] https://paperswithcode.com/dataset/wikitables-turl/.

2.2 Постановка задачи

В качестве входных данных рассматриваются вертикальные таблицы, представляющие собой массив данных, расположенных в форме столбцов (вертикальных колонок). Столбец может содержать заголовок (шапку). В таких таблицах столбцы могут быть двух типов:

- *категориальный столбец* (*столбец именованных сущностей*) содержит названия некоторых именованных сущностей;
- *литеральный столбец* содержит литеральные значения (например, даты, числа, *URL*-адреса).

Вертикальная таблица может быть ненормализованной, однако должна удовлетворять следующим двум предположениям (ограничениям):

Предположение 1. В обрабатываемых таблицах нет объединённых ячеек.

Предположение 2. Исходные таблицы обрабатываются независимо друг от друга.

В качестве ЦГЗ используется ГЗ платформы *Talisman* $KG = \{DM, F\}$, где KG – ГЗ платформы *Talisman*; DM – модель ПрО, задающая онтологическую схему с абстрактным описанием понятий и их отношений; F – набор конкретных сущностей (фактов), которые типизируются на основе модели ПрО. При этом $DM = \{CT, PT, PVT, BVT, RT\}$, где CT – тип концепта (например, *персона*, *организация*, *продукция*); PT – тип характеристики (например, *адрес проживания*, *рабочий телефон*, *дата рождения*); PVT – тип значения характеристики (например, *адрес*, *дата*, *расстояние*); BVT – базовый тип значения характеристики (например, *координаты*, *дата*, *интервал дат*, *строка* и т.д.); RT – тип связи, определённый между двумя типами концептов (например, «*работает в*», «*учится в*», «*является*»). $F = \{C, P, AV, R, M\}$, где C – концепт (например, *определённый человек*, *конкретная организация* или *продукт*); P – характеристика (свойство) концепта, представляющая интерес для конечных пользователей, характеристика может быть идентифицирующей (например, «*название*», которое однозначно характеризует конкретных объект); AV – конкретное атомарное значение характеристики (например, *возраст человека* или *номер мобильного телефона*); R – связь, задающая отношение между двумя концептами; M – упоминание, которое представляет собой фрагмент текста, прямо указывающий на объект, событие или понятие реального / виртуального мира, соответствующее некоторому концепту, характеристике или связи. Пример использования ГЗ *Talisman* приведён на рисунке 2.

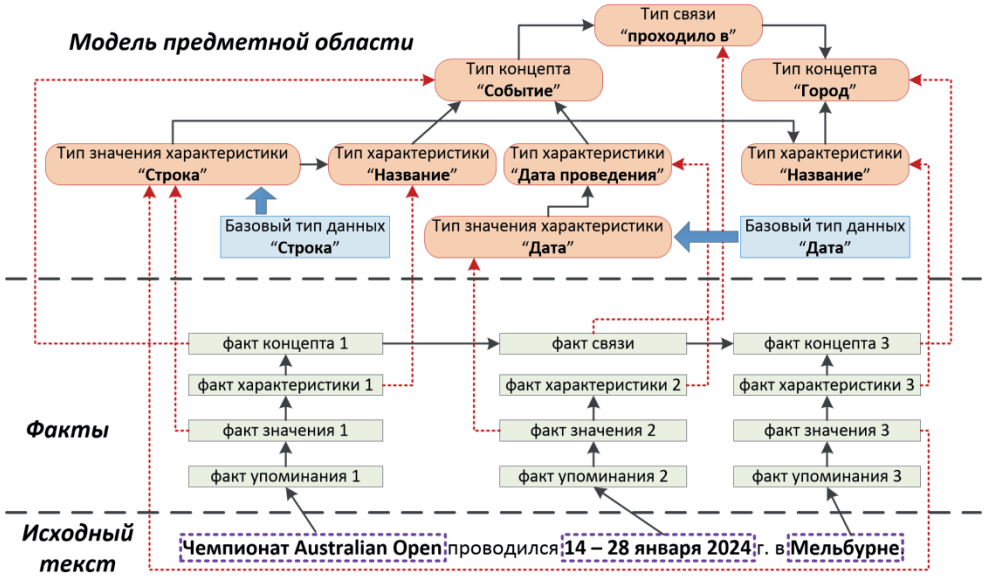


Рисунок 2 – Пример использования графа знаний платформы *Talisman*

Предлагаемый подход реализует семантическое аннотирование столбцов и отношений между ними, которое заключается в сопоставлении столбцам определённых *типов характеристик*, нахождении наиболее подходящего типа концепта на их основе, а также выявление *типов связей* между определёнными типами концептов.

2.3 Этапы подхода

Разработанный подход направлен на обработку *Talisman*-документов в формате *TDM* (*Talisman Document Model*) версии 1.0, которые могут содержать набор вертикальных таблиц. *TDM* используется в *Talisman* для унифицированного представления данных, извлечённых из файлов различных форматов (*PDF*, *DOCX*, *CSV*, *HTML*). Основные этапы предлагаемого подхода представлены на рисунке 3.

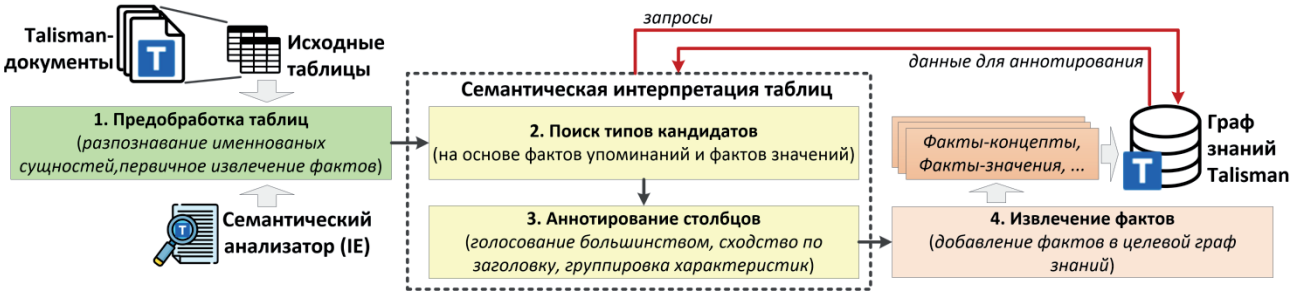


Рисунок 3 – Основные этапы предлагаемого подхода

Этап 1: *Предобработка таблиц***.** На данном этапе осуществляется распознавание именованных сущностей (*Named Entity Recognition – NER*) для каждой ячейки в исходной таблице. Для этой цели используется дообученная модель *XLM-RoBERTa* [26], которая распознаёт в тексте вхождение некоторых именованных сущностей (персон, компаний, местоположений и др.). Модель дообучалась на наборах данных: *CoNLL 2003* (*English*), *OntoNotes* (*English*), *OntoNotes* (*Chineese*) и *DocRED* (*English*). Определённые *NER*-метки именованных сущностей присваиваются каждой ячейке в исходной таблице, характеризуя содержащиеся в ней данные. В зависимости от присвоенной *NER*-метки из ячеек автоматически извлекаются *факты-упоминания* и *факты-значения*, соответствующие типу значения характеристик, определённому в модели ПрО. На данном шаге из ячеек могут быть извлечены предварительные *факты-характеристики* и *факты-концепты*. Данный этап выполняется средствами *семантического анализатора* (*IE*), входящего в состав платформы *Talisman*.

Этап 2: *Поиск типов кандидатов*. Для каждого столбца формируется набор кандидатных типов характеристик, полученных из модели ПрО на основе определённых *фактов-упоминаний* и *фактов-значений*. Столбцы, для которых факты не были извлечены на предыдущем шаге, исключаются из последующей обработки таблицы.

Этап 3: *Семантическое аннотирование столбцов*. На данном этапе осуществляется выбор наиболее подходящего типа характеристики из множества кандидатов для присвоения его столбцу. Для этого используется специальный агрегированный метод состоящей из комбинации следующих эвристик.

- *Голосование большинством*. Данная эвристика является простым базовым решением, которое заключается в том, что наиболее подходящий тип из набора кандидатов назначается столбцу на основе прямого вывода из тех *фактов-характеристик*, которые уже были извлечены для ячеек столбца средствами семантического анализатора. Т.е. для каждого определённого *факта-характеристики* находится набор возможных типов, которым он соответствует. Далее подсчитывается количество (частота появления) каждого

типа-кандидата. Для приведения данного значения к диапазону от 0 до 1 применяется метод нормализации [27].

- *Сходство по заголовку*. Осуществляется лексическое сопоставление заголовка столбца с названиями типов характеристик из множества кандидатов на основе расстояния Левенштейна и в зависимости от этого расстояния даётся оценка каждому типу кандидата. Если в столбце выделены *факты-концепты* (на этапе предобработки), то название заголовка сравнивается не с названиями типов характеристик из множества кандидатов, а с названиями типов концептов, которые связаны с данными типами характеристик.
- *Группировка характеристик*. Данная эвристика основана на предположении, что в таблице может быть один или несколько категориальных столбцов, в которых семантический анализатор уже извлёк некоторые *факты-концепты* с идентифицирующими *фактами-характеристиками* (например, характеристика «*название*» для некоторого концепта *организации*). Для каждого категориального столбца подсчитывается количество возможных характеристик, которые располагаются в других не категориальных (литеральных) столбцах и относятся к данному концепту. Далее определяется, какому категориальному столбцу соответствует максимальное количество характеристик. Такому столбцу и столбцам с характеристиками соответствует оценка равная единице.

На основе этих эвристик определяется общая (агрегированная) оценка того, что определённый тип характеристики из набора кандидатов является наиболее подходящим для аннотирования столбца таблицы.

Этап 4. *Извлечение фактов*. На основе установленных аннотаций столбцов из таблицы извлекаются новые *факты-концепты*, *факты-значения*, *факты-упоминания*, *факты-характеристики* концептов. При этом извлечённые *факты-упоминания* включают значение всей ячейки целиком. Извлечение фактов осуществляется построчно слева направо. *Факты-характеристики* создаются только для самого левого категориального столбца в таблице. Если в таблице в качестве аннотации для нескольких категориальных столбцов определён один и тот же тип характеристики (например, если в таблице есть два столбца с персонами, а все остальные столбцы определены как некоторые характеристики персоны, то только для *фактов-концептов* из первого столбца создаются соответствующие характеристики). При этом идентифицирующие характеристики (названия) извлекаются всегда. На основе извлечённых *фактов-концептов* из таблицы также построчно извлекаются все возможные *факты-связи*, определяющие отношения между двумя концептами. Все извлечённые таким образом факты пополняют ЦГЗ *Talisman*.

2.4 Программная реализация

Предлагаемый подход реализован в форме специального обработчика «*tables-annotator*» на языке *Python* 3.10. Данный обработчик входит в состав подсистемы *Talisman Information Extraction* (*Talisman-IE*) и представляет собой программное средство (*REST-сервер*), выполняющее обработку входного *Talisman*-документа. Обработчик также получает на вход *JSON*-объект, задающий правила и/или ограничения (конфигурацию) трансформации входных документов в обработчике.

Конфигурация для обработчика «*tables-annotator*»:

```
{
   "table_indices": "<порядковые номера таблиц>",
   "column_indices": {
      "<порядковые номера таблиц>": "<порядковые номера столбцов>",
      ...
   },
   "header_numbers": [ <номер строки 1>, ..., <номер строки n> ]
}
```

Параметры конфигурации, опциональный блок:

- «*table_indices*» – задаёт номера таблиц, встречающихся в исходном документе, которые необходимо исключить из обработки. Для этого указывается строка, в которой через запятую могут быть указаны как отдельные порядковые номера таблиц, так и их диапазоны, например: «*1, 2, 3, 5-8, 10*». Если в диапазоне указать специальное значение «*end*», то отсчёт таблиц продолжится автоматически до конца документа, например: «*1, 3, 5-end*». Отсчёт таблиц в документе начинается с единицы;
- «*column_indices*» – задаёт номера столбцов, которые необходимо исключить из обработки в заданных таблицах. Для этого указывается словарь, где ключ – это номера таблиц или их диапазон, а значение – это номера столбцов или их диапазон, относящиеся к указанным таблицам. Данные номера таблиц и столбцов являются текстовыми значениями и составляются по такому же принципу, как и параметр «*table_indices*»;
- «*header_numbers*» – задаёт список номеров строк, являющихся заголовком таблицы. По умолчанию первая строка таблицы считается заголовком. Номера строк должны быть числовыми значениями, указываются без кавычек. Отсчёт строк в таблице начинается с единицы.

Если необходимо обработать все таблицы из документа и при возможности извлечь из них факты, то конфигурация по умолчанию не задаётся.

3 Пример

Разработанный обработчик «*tables-annotator*» использован для решения задачи автоматизированного наполнения предметно-ориентированных ГЗ платформы *Talisman* новыми фактами, извлечёнными из ТД. Тестирование разработанного обработчика производилось при анализе тестовых таблиц, собранных по категориям: «*сотрудники организации*», «*открытые вакансии*», «*рынок автомоделей*», «*известные учёные*», «*продажа книг*», «*рейтинг теннисистов*». Для формирования тестового набора ТД использовались следующие веб-ресурсы:

- сайты научных и образовательных учреждений (например, *ИДСТУ СО РАН*[29], *Иркутский национальный исследовательский технический университет*[30]);
- *банк вакансий Иркутской области*[31] и веб-сервис *hh* (*Иркутск*)[32];
- веб-сервис «*Авито*»[33];
- русскоязычная часть Википедии[34];
- веб-магазин «*Лабиринт*»[35];
- веб-сервис подсчёта рейтинга теннисистов по версии *ATP*[36].

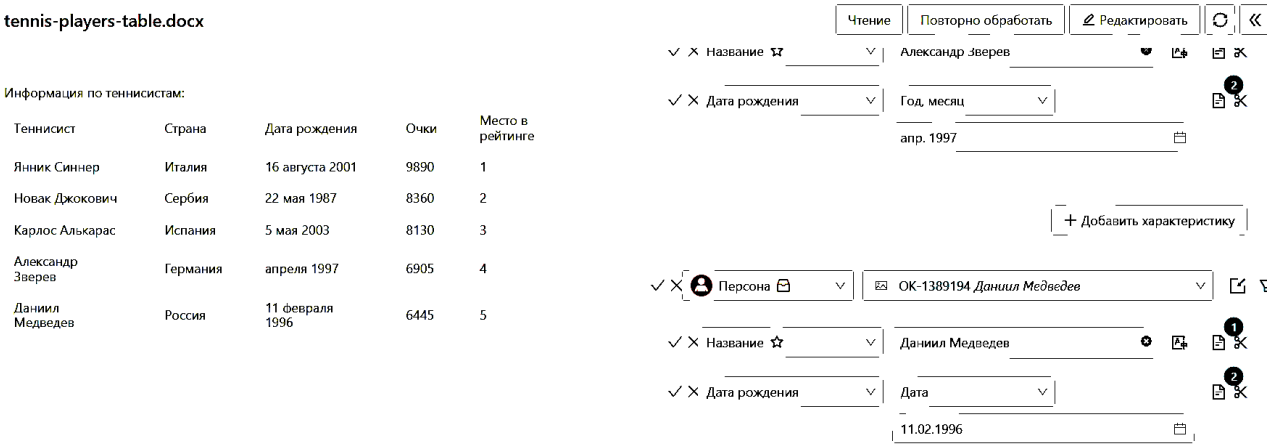
Данные собирались из веб-таблиц и сохранялись в форме *DOCX*-документов. Среднее количество столбцов в собранных таблицах – 5, а среднее количество строк – 12.

Фрагмент модели ПрО, использованной в процессе семантического аннотирования таблиц и на этапе пополнения ТД, показан на рисунке 4. Данный ГЗ представлен в виде ориентированного графа, доступ к которому осуществляется с помощью интерфейса *GraphQL*[37]. В модели описаны основные типы концептов, такие как «*Персона*» (*NER*-метки: *PERSON*, *PER*), «*Организация*» (*NER*-метки: *ORGANIZATION*, *ORG*), «*Вакансия*» (нет соответствующей *NER*-метки), «*Автомобиль*» (*NER*-метки: *PRODUCT*) и «*Книга*» (*NER*-метки: *WORK_OF_ART*).

На рисунке 5 показан пример обработанной исходной таблицы из категории «*рейтинг теннисистов*» (см. рисунок 1).

[29] http://idstu.irk.ru.
[30] https://www.istu.edu.
[31] https://trudvsem.ru/vacancy/search?_regionIds=3800000000000.
[32] https://irkutsk.hh.ru.
[33] https://www.avito.ru.
[34] https://ru.wikipedia.org.
[35] https://www.labirint.ru.
[36] https://www.labirint.ru.
[37] https://live-tennis.eu/ru/atp-live-ranking.

Рисунок 4 – Фрагмент модели предметной области из целевого графа знаний *Talisman*



tennis-players-table.docx

Информация по теннисистам:

Теннисист	Страна	Дата рождения	Очки	Место в рейтинге
Янник Синнер	Италия	16 августа 2001	9890	1
Новак Джокович	Сербия	22 мая 1987	8360	2
Карлос Алькарас	Испания	5 мая 2003	8130	3
Александр Зверев	Германия	апреля 1997	6905	4
Даниил Медведев	Россия	11 февраля 1996	6445	5

Рисунок 5 – Фрагмент обработанной исходной таблицы из категории «*рейтинг теннисистов*» на платформе *Talisman*

Для получения экспериментальной оценки семантического аннотирования столбцов таблиц с помощью обработчика «*tables-annotator*» использовались: *точность* (*precision*), *полнота* (*recall*) и *F-мера* (*F*1):

$$precision = \frac{P}{C},\ recall = \frac{P}{CN},\ F1 = \frac{2 \times precision \times recall}{precision + recall},$$

где P – количество правильно аннотированных столбцов (идеальных аннотаций); C – количество аннотированных столбцов; CN – общее количество столбцов в исходной таблице.

Полученная оценка представлена в таблице 1. Анализ показал определяющее значение этапа распознавания именованных сущностей (Этап 1), в частности, исключение из процесса обработки столбцов, для которых не были определены *NER*-метки (например, для столбца с названием *открытой вакансии* для таблиц из категории «*вакансии*»), что приводит к низким оценкам.

Таблица 1 – Экспериментальная оценка для таблиц из разных категорий

Категория таблиц	Точность	Полнота	F-мера
Сотрудники организации	1.00	0.80	0.89
Открытые вакансии	0.20	0.16	0.18
Рынок автомобилей	1.00	0.83	0.91
Известные ученые	0.75	0.75	0.75
Продажа книг	0.80	0.67	0.73
Рейтинг теннисистов	1.00	0.60	0.75
Итоговая оценка	**0.79**	**0.63**	**0.70**

Другими ограничениями разработанного подхода являются:

- обработка только вертикальных таблиц;

- из ячеек таблицы извлекаются значения (упоминания) целиком (например, не извлекается отдельно «*Имя*», «*Фамилия*» и «*Отчество*» из ячейки с «*ФИО*»);
- не формируется одно значение из значений нескольких ячеек;
- не рассматриваются сложные составные значения характеристик;
- не извлекаются характеристики связей.

Заключение

В статье представлен подход к автоматизированной разработке предметно-ориентированных ГЗ на основе семантического аннотирования ТД. Предлагаемый подход включает комбинацию эвристических решений для аннотирования столбцов таблиц и аннотирования отношений между столбцами. В качестве входных данных используются *Talisman*-документы, а в качестве ЦГЗ – ГЗ платформы *Talisman*. Подход реализован в форме специального модуля-обработчика «*tables-annotator*» для *Talisman-IE*.

Список источников

[1] ***Ji S., Pan S., Cambria E., Marttinen P., Yu P.S.*** A Survey on Knowledge Graphs: Representation, Acquisition and Applications // IEEE Transcations on Neural Networks and Learning Systems. 2021. Vol.33(2). P.494-514. DOI: 10.1109/TNNLS.2021.3070843.

[2] ***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. DOI: 10.1145/3447772.

[3] ***Баклавски К.*** Онтологический Саммит 2020. Коммюнике: Графы знаний / К. Баклавски, М. Беннет, Г. Берг-Кросс, Т. Шнайдер, Р. Шарма, Д. Сингер. Перевод с англ. Д. Боргест // Онтология проектирования. 2020. Т.10, №4(38). С.540-555. DOI: 10.18287/2223-9537-2020- 10-4-540-555.

[4] ***Гаврилова Т.А., Страхович Э.В.*** Визуально-аналитическое мышление и интеллект-карты в онтологическом инжиниринге // Онтология проектирования. 2020. Т.10, №1(35). С.87-99. DOI: 10.18287/2223-9537-2020-10-1-87-99.

[5] ***Martinez-Rodriguez J.L., Hogan A., Lopez-Arevalo I.*** Information Extraction meets the Semantic Web: A Survey // Semantic Web. 2020. Vol.11. P.255-335. DOI: 10.3233/SW-180333.

[6] ***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. DOI: 10.1145/3372117.

[7] ***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(4). e1407. DOI: 10.1002/widm.1407.

[8] ***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. DOI: 10.1145/2872518.2889386.

[9] ***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.

[10] ***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. DOI: 10.1016/j.websem.2022.100761.

[11] ***Limaye G., Sarawagi S., Chakrabarti S.*** Annotating and Searching Web Tables Using Entities, Types and Relationships. In: Proc. VLDB Endowment. 2010. Vol.3. P.1338-1347. DOI: 10.14778/1920841.1921005.

[12] ***Mulwad V., Finin T., Syed Z., Joshi A.*** Using linked data to interpret tables. In: Proc. the First International Conference on Consuming Linked Data (COLD'10). 2010. Vol.665. P.109-120. DOI: 10.5555/2878947.2878957.

[13] ***Bhagavatula C.S., Noraset T., Downey D.*** TabEL: Entity Linking in Web Tables. In: Proc. the 14th International Semantic Web Conference (ISWC'2015). 2015. P.425-441. DOI: 10.1007/978-3-319-25007-6_25.

[14] ***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. DOI: 10.1007/978-3-319-68288-4_16.

[15] ***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. DOI: 10.5441/002/EDBT.2017.20.

[16] ***Zhang Z.*** Effective and efficient semantic table interpretation using TableMiner+. *Semantic Web*. 2017. Vol.8(6). P.921-957. DOI: 10.3233/SW-160242.
[17] ***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. DOI: 10.1609/aaai.v33i01.3301281.
[18] ***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. DOI: 10.1007/978-3-030-30793-6_21.
[19] ***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. DOI: 10.1609/aaai.v33i01.330129.
[20] ***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. DOI: 10.1145/3292500.3330993.
[21] ***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. DOI: 10.1007/978-3-030-50417-5_46.
[22] ***Zhang D., Suhara Y., Li J., Hulsebos M., Demiralp C., Tan W.-C.*** Sato: Contextual semantic type detection in tables. In: Proc. the VLDB Endowment. 2020. Vol.13(11). P.1835-1848. DOI: 10.14778/3407790.3407793.
[23] ***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. DOI: 10.14778/3430915.3430921.
[24] ***Yin P., Neubig G., Yih W.*** TaBERT: Pretraining for Joint Understanding of Textual and Tabular Data. In: Proc. the 58th Annual Meeting of the Association for Computational Linguistics. 2020. P.8413-8426. DOI: 10.18653/v1/2020.acl-main.745.
[25] ***Iida H., Thai D., Manjunatha V., Iyyer M.*** TABBIE: Pretrained Representations of Tabular Data. In: Proc.the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. 2021. P.3446-3456. DOI: 10.18653/v1/2021.naacl-main.270.
[26] ***Conneau A., Khandelwal K., Goyal N., Chaudhary V., Wenzek G., Guzmán F., Grave E., Ott M., Zettlemoyer L., Stoyanov V.*** Unsupervised Cross-lingual Representation Learning at Scale // In: Proc. the 58th Annual Meeting of the Association for Computational Linguistics. 2020. P.8440-8451. DOI: 10.18653/v1/2020.acl-main.747.
[27] ***Dorodnykh N.O., Yurin A.Yu.*** Knowledge Graph Engineering Based on Semantic Annotation of Tables. *Computation*. 2023. Vol. 11(9). 175. DOI: 10.3390/computation11090175.

Сведения об авторах

Дородных Никита Олегович, 1990 г. рождения. Окончил Иркутский национальный исследовательский технический университет (ИрНИТУ) (2012), к.т.н. (2018). Старший научный сотрудник ИДСТУ СО РАН. В списке научных трудов около 80 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. *nikidorny@icc.ru.* ✉.

Юрин Александр Юрьевич, 1980 г. рождения. Окончил Иркутский государственный технический университет (2002), д.т.н. (2022). Заведующий лабораторией Информационно-телекоммуникационных технологий исследования техногенной безопасности ИДСТУ СО РАН, профессор Института информационных технологий и анализа данных ИрНИТУ. Член Российской ассоциации искусственного интеллекта. 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 22.07.2024, после рецензирования 29.10.2024. Принята к публикации 1.11.2024.

Scientific article DOI: 10.18287/2223-9537-2024-14-4-555-568

Development of domain knowledge graph based on semantic annotation of tabular data

© 2024, 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 article outlines an approach and software tool for the automated enrichment of domain-oriented knowledge graphs with new facts derived from semantically annotated tabular data. For semantic annotation of table columns, a combination of three heuristic methods is proposed, leveraging named entity recognition in cells, lexical matching, and feature grouping. This approach is implemented as a dedicated handler within the Talisman software platform. An example and experimental evaluation of the approach during the semantic annotation of columns are provided using a test set of tabular data across six thematic categories: "organization employees," "open vacancies," "car model market," "famous scientists," "book sales," and "tennis player rankings." Evaluation metrics included precision, recall, and F-measure, with final results across all six categories as follows: precision - 79%, recall- 63%, F-measure - 70%. These results highlight the potential of the developed approach for enriching domain-oriented knowledge graphs with new facts from semantically annotated tabular data. The limitations of the proposed approach are also discussed from semantically annotated tabular data. The paper also provides a number of limitations of the proposed approach.

Keywords: *semantic web, knowledge graph, semantic table interpretation, table annotation, entity extraction, knowledge enrichment, tabular data.*

For citation: Dorodnykh NO, Yurin AYu. Development of domain knowledge graph based on semantic annotation of tabular data [In Russian]. *Ontology of designing*. 2024; 14(4): 555-568. DOI:10.18287/2223-9537-2024-14-4-555-568.

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 declare no conflict of interest.

List of figures and tables

References

[1] ***Ji S, Pan S, Cambria E, Marttinen P, Yu PS.*** A Survey on Knowledge Graphs: Representation, Acquisition and Applications. IEEE Transcations on Neural Networks and Learning Systems. 2021; 33(2): 494-514. DOI: 10.1109/TNNLS.2021.3070843.

[2] ***Hogan A, Blomqvist E, Cochez M, d’Amato C, Melo GD, Gutierrez C, Gayo JEL, Kirrane S, Neumaier S, Polleres A, Navigli R, Ngomo ACN, Rashid SM, Rula A, Schmelzeisen L, Sequeda J, Staab S, Zimmermann A.*** Knowledge Graphs. ACM Computing Surveys. 2021; 54(4): 1-37. DOI: 10.1145/3447772.

[3] ***Baklawski K. Bennett M, Berg-Cross G, Schneider T, Sharma R, Singer D.*** Ontology Summit 2020: Knowledge Graphs. Translation from English D. Borgest [In Russian]. Ontology of designing. 2020; 4(38): 540-555. DOI: 10.18287/2223-9537-2020- 10-4-540-555.

[4] ***Gavrilova TA, Strakhovich EV.*** Visual analytical thinking and mind maps for ontology engineering [In Russian]. Ontology of designing. 2020; 10(1): 87-99. DOI: 10.18287/2223-9537-2020-10-1-87-99.

[5] ***Martinez-Rodriguez JL, Hogan A, Lopez-Arevalo I.*** Information Extraction meets the Semantic Web: A Survey. Semantic Web. 2020; 11: 255-335. DOI: 10.3233/SW-180333.

[6] ***Zhang S, Balog K.*** Web table extraction, retrieval, and augmentation: A survey. ACM Transactions on Intelligent Systems and Technology. 2020; 11(2): 1-35. DOI: 10.1145/3372117.

[7] ***Bonfitto S, Casiraghi E, Mesiti M.*** Table understanding approaches for extracting knowledge from hetero-geneous tables. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery. 2021; 11(4): e1407. DOI: 10.1002/widm.1407.

[8] ***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 (Montréal, Québec, Canada, April 11-15, 2016). International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, Switzerland, 2016: 75-76. DOI: 10.1145/2872518.2889386.

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

[10] ***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; 76: 100761. DOI: 10.1016/j.websem.2022.100761.

[11] ***Limaye G, Sarawagi S, Chakrabarti S.*** Annotating and Searching Web Tables Using Entities, Types and Relationships. Proc. the VLDB Endowment. 2010; 3: 1338-1347. DOI: 10.14778/1920841.1921005.

[12] ***Mulwad V, Finin T, Syed Z, Joshi A.*** Using linked data to interpret tables. In: Proc. the First International Conference on Consuming Linked Data (Shanghai, China, November 8, 2010). CEUR-WS, 2010: 109-120. DOI: 10.5555/2878947.2878957.

[13] ***Bhagavatula CS, Noraset T, Downey D.*** TabEL: Entity Linking in Web Tables. In: Proc. the 14th International Semantic Web Conference (Bethlehem, PA, USA, October 11-15, 2015). Lecture Notes in Computer Science, vol. 9366. Springer, Cham, 2015: 425-441. DOI: 10.1007/978-3-319-25007-6_25.

[14] ***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. (Vienna, Austria, October 21-25, 2017). Lecture Notes in Computer Science, vol. 10587. Springer, Cham, 2017: 260-277. DOI: 10.1007/978-3-319-68288-4_16.

[15] ***Ritze D, Bizer C.*** Matching web tables to DBpedia - A feature utility study. In: Proc. of the 20th Int. Conf. on Extending Database Technology (Venice, Italy, March 21-24, 2017). OpenProceedings, 2017: 210-221. DOI: 10.5441/002/EDBT.2017.20.

[16] ***Zhang Z.*** Effective and efficient semantic table interpretation using TableMiner+. Semantic Web. 2017; 8(6): 921-957. DOI: 10.3233/SW-160242.

[17] ***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 (Honolulu, Hawaii, USA, January 27, 2019), vol. 33(1). AAAI Press, 2019: 281-288. DOI: 10.1609/aaai.v33i01.3301281.

[18] ***Kruit B, Boncz P, Urbani J.*** Extracting Novel Facts from Tables for Knowledge Graph Completion. Proc. of the 18th Int. Semantic Web Conf. (Auckland, New Zealand, October 26-30, 2019). Lecture Notes in Computer Science, vol. 11778. Springer, Cham, 2019: 364-381. DOI: 10.1007/978-3-030-30793-6_21.

[19] ***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 (Honolulu, Hawaii, USA, January 27, 2019), vol. 33(1). AAAI Press, 2019: 29-36. DOI: 10.1609/aaai.v33i01.330129.

[20] ***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 (Anchorage, AK, USA, August 4-8, 2019). Association for Computing Machinery, New York, NY, United States, 2019: 1500-1508. DOI: 10.1145/3292500.3330993.

[21] ***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 (Amsterdam, The Netherlands, June 3-5, 2020). Lecture Notes in Computer Science, vol. 12138. Springer Cham, 2020: 618-631. DOI: 10.1007/978-3-030-50417-5_46.

[22] ***Zhang D, Suhara Y, Li J, Hulsebos M, Demiralp C, Tan WC.*** Sato: Contextual semantic type detection in tables. Proc. the VLDB Endowment. 2020; 13(11): 1835-1848. DOI: 10.14778/3407790.3407793.

[23] ***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. DOI: 10.14778/3430915.3430921.

[24] ***Yin P, Neubig G, Yih W.*** TaBERT: Pretraining for Joint Understanding of Textual and Tabular Data. In: Proc. the 58th Annual Meeting of the Association for Computational Linguistics. Online, 2020: 8413-8426. DOI: 10.18653/v1/2020.acl-main.745.

[25] ***Iida H, Thai D, Manjunatha V, Iyyer M.*** TABBIE: Pretrained Representations of Tabular Data. In: Proc.the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Online, 2021: 3446-3456. DOI: 10.18653/v1/2021.naacl-main.270.

[26] ***Conneau A, Khandelwal K, Goyal N, Chaudhary V, Wenzek G, Guzmán F, Grave E, Ott M, Zettlemoyer L, Stoyanov V.*** Unsupervised Cross-lingual Representation Learning at Scale // In: Proc. the 58th Annual Meeting of the Association for Computational Linguistics. Online, 2020: 8440-8451. DOI: 10.18653/v1/2020.acl-main.747.

[27] ***Dorodnykh NO, Yurin AYu.*** Knowledge Graph Engineering Based on Semantic Annotation of Tables. Computation. 2023; 11(9): 175. DOI: 10.3390/computation11090175.

About the authors

Nikita Olegovych Dorodnykh (b. 1990) graduated from the INRTU in 2012, PhD (2018). He is a senior associate researcher at ISDCT SB RAS. Co-author of about 80 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. *nikidorny@icc.ru.* ✉.

Alexander Yurievich Yurin (b.1980) graduated from the INRTU in 2002, Doctor of Science (2022). Head of the laboratory of Information and Telecommunication technologies for Research of Technogenic Security at the Institute of Technical University of the Siberian Branch of the Russian Academy of Sciences, Professor at the Institute of Information Technologies and Data Analysis at the Irkutsk National Research Technical University. He is a member of the Russian Association of Artificial Intelligence (RAAI), 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. ORCID: 0000-0001-9089-5730; Author ID (RSCI): 174845; Author ID (Scopus): 16311168300; Researcher ID (WoS): A-4355-2014. *iskander@icc.ru.*

Received July 22, 2024. Revised October 29, 2024. Accepted November 1, 2024.