**ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ БЮДЖЕТНОЕ ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ «МОСКОВСКИЙ АВИАЦИОННЫЙ ИНСТИТУТ (НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ)»**

**Журнал практики**

Студента \_\_Урубкова Владислава Станиславовича\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_(ф. и. о.)

##### Институт №3 «Системы управления, информатика и электроэнергетика»\_\_\_

###### Кафедра № 319 «Системы интеллектуального мониторинга»\_

Учебная группа\_\_М3О-435Б-18 \_\_\_\_\_

Направление подготовки 09.03.01

*(шифр)*

***«***Информатика и вычислительная техника»

*(название направления)*

Вид практики *\_\_\_\_\_\_производственная (распределенная, преддипломная)*

*(учебной, производственной, преддипломной или другой вид практики)*

Руководитель практики от МАИ

Фомичев Владимир Александрович \_\_\_\_\_\_\_\_\_\_\_\_\_

*(фамилия, имя, отчество) (подпись)*

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_/\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_/ “10” мая 2022 г.

*(подпись студента)* *(дата)*

Москва 2022

**1.Место и сроки проведения практики**

*Сроки проведения практики:*

*-дата начала практики\_\_\_\_\_09 февраля 2022 г.\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_*

*-дата окончания практики\_\_\_10 мая 2022 г.\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_*

*Наименование предприятия кафедра 319 «Системы интеллектуального мониторинга» \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_*

*Название структурного подразделения (отдел, лаборатория) \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ \_\_\_\_\_\_\_\_\_\_\_\_\_*\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

1. **Инструктаж по технике безопасности**

**Желательно написать разные варианты, например, так**

*Первичный инструктаж на рабочем месте.*

Инструктаж по охране труда для работников, занятых эксплуатацией ПЭВМ;

Инструктаж по охране при работе на копировально-множительном аппарате;

???????????

*Противопожарный инструктаж*

*????*

*Инструктаж по технике электробезопасности*.

Инструктаж по электробезопасности на 1-ую группу, предназначенная для персонала, выполняющего работы, при которых может возникнуть опасность поражения электрическим током;

Под электробезопасностью понимается система организационных и технических мероприятий по защите человека от действия электрического тока, электрической дуги, статического электричества, электромагнитного поля.

Инструктаж направлен на понимание студентов, что при пользование любыми электрическими приборами или аппаратами необходимо всегда помнить о том, что некорректное обращение с ними, неисправное состояние электропроводки или самого электроприбора, несоблюдение определенных мер предосторожности могут привести к поражению электрическим током.

Неисправность электропроводки может стать причиной возгорания проводов и возникновения пожаров.

Шевченко Д.А. \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_/ “09”февраля 2022 г.

*(подпись проводившего)* *(дата проведения)*

1. **Индивидуальное задание студенту**

ТЕМА: Семантически-ориентированный естественно-языковой интерфейс для взаимодействия с Системой взаимосвязанных открытых данных (Linked Open Data)

1. Сбор сведений о семантической информационной системе LOD и ее применениях.
2. Изучение основных конструкций языка SPARQL.
3. Исследование основных подходов к формальному описанию семантической структуры текстов на естественном языке.
4. Исследование основных подходов к разработке семантически-ориентированных ЕЯ-интерфейсов для взаимодействия с прикладными интеллектуальными системами.
5. Исследование основных подходов к разработке интеллектуальных интерфейсов для преобразования запроса к LOD на ЕЯ в запросы на языке SPARQL.
6. Разработка логической структуры лингвистической базы данных.
7. Разработка алгоритмов для реализации преобразования вида «ЕЯ-запрос → SPARQL-запрос».
8. Составление и сдача отчета по научно-исследовательской практике
9. **План выполнения индивидуального задания**
10. С 09 февраля 2022 г. по 17 февраля 2022 г. – Сбор сведений о семантической информационной системе LOD и ее применениях.
11. С 18 февраля 2022 г. по 1 марта 2022 г. – Изучение основных конструкций языка SPARQL.
12. С 02 марта 2022 г. по 16 марта 2022 г. – Исследование основных подходов к формальному описанию семантической структуры текстов на естественном языке.
13. С 17 марта 2022 г. по 28 марта 2022 г. – Исследование основных подходов к разработке семантически-ориентированных ЕЯ-интерфейсов для взаимодействия с прикладными интеллектуальными системами.
14. С 29 марта 2022 г. по 07 апреля 2022 г. – Исследование основных подходов к разработке интеллектуальных интерфейсов для преобразования запроса к LOD на ЕЯ в запросы на языке SPARQL.
15. С 08 апреля 2022 г. по 13 апреля 2022 г. – Разработка логической структуры лингвистической базы данных.
16. С 14 апреля 2022 г. по 30 апреля 2022 г. – Разработка алгоритмов для реализации преобразования вида «ЕЯ-запрос → SPARQL-запрос».
17. С 1 мая 2022 г. по 10 мая 2022 г. – Составление и сдача отчета по научно-исследовательской практике.

*Руководитель практики от каф.319*: \_\_\_Фомичев В.А.\_\_/\_\_\_\_\_\_\_\_\_\_\_/

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_/\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_/ “\_09\_”\_февраля\_ 2022 г.

*(подпись студента)* *(дата)*

1. **Отзыв руководителя практики от МАИ**

За время прохождения практики Урубков В.С. зарекомендовал себя как ответственный и исполнительный практикант, показал высокий уровень теоретической подготовки, хорошее умение применить и использовать полученные знания для решения поставленных перед ним задач.

Программа практики выполнена полностью.

В целом работа практиканта Урубкова В.С. заслуживает оценки «отлично».

Материалы, изложенные в отчете студента, полностью соответствуют индивидуальному заданию.

Целью освоения практики Преддипломная практика является достижение следующих результатов освоения(РО):

|  |  |  |
| --- | --- | --- |
| **N** | **Шифр** | **Результат освоения** |
| 1 | У-1 (ПК-1) | Уметь применять на практике модели данных, используемые при описании проектов задач информационных систем; теорию и методологию проектирования документов, диалогов пользователя и технологических процессов обработки информации; теорию и методологию проектирования структуры файлов, их взаимосвязей |
| 2 | В-1 (ПК-1) | Владеть моделями данных, используемыми при описании проектов задач информационных систем; теорией и методологией проектирования документов, диалогов пользователя и технологических процессов обработки информации; теорией и методологией проектирования структуры файлов, их взаимосвязей |
| 3 | З-2 (ПК-1) | Знать теоретические основы баз данных: реляционную модель баз данных, ER- моделирование, нормализацию отношений, язык запросов SQL и технологии проектирования реляционных баз данных |
| 4 | У-2 (ПК-1) | Уметь разрабатывать схемы баз данных |
| 5 | В-2 (ПК-1) | Владеть навыками работы с инструментальными средствами проектирования баз данных и знаний, управления проектами ИС и защиты информации |
| 6 | У-1 (ПК-3) | Уметь обосновывать реализуемые проектные решения, осуществлять постановку и выполнение экспериментов по проверке корректности и эффективности этих решений |
| 7 | В-1 (ПК-3) | Владеть навыками обоснования реализуемых проектных решений, осуществления постановки и выполнения экспериментов по проверке корректности и эффективности этих решений |
| 8 | З-1 (ПК-4) | Знать основные программно-методические комплексы ИВТ, используемые на предприятии |
| 9 | У-1 (ПК-4) | Уметь создавать конспекты лекций, практикумов, методические указания и презентационные материалы для проведения лабораторных, семинарских и практических занятий по применению программно-методических комплексов ИВТ, используемых на предприятии |
| 10 | В-1 (ПК-4) | Владеть методами создания конспектов лекций, практикумов, методических указаний и презентационных материалов для проведения лабораторных, семинарских и практических занятий по применению программно-методических комплексов ИВТ, используемых на предприятии |
| 11 | З-2 (ПК-4) | Знать методы создания конспектов лекций, практикумов, методических указаний и презентационных материалов для проведения лабораторных, семинарских и практических занятий по применению программно-методических комплексов ИВТ, используемых на предприятии |
| 12 | У-2 (ПК-4) | Уметь проводить лекции, практикумы, лабораторные, семинарские и практические занятия по применению программно-методических комплексов ИВТ, используемых на предприятии |
| 13 | В-2 (ПК-4) | Владеть методами проведения лекций, практикумов, лабораторных, семинарских и практических занятий по применению программно-методических комплексов ИВТ, используемых на предприятии |
| 14 | В-3 (ПК-4) | Владеть методами подготовки презентационных и учебно-методических материалов с использованием современных средств подготовки текстовой, графической и мультимедийной информации |

Перечисленные РО являются основой для формирования следующих компетенций:

|  |  |  |
| --- | --- | --- |
| **N** | **Шифр** | **Компетенция** |
| 1 | ПК-1 | Способность разрабатывать модели компонентов информационных систем, включая модели баз данных и модели интерфейсов "человек - электронно-вычислительная машина" |
| 2 | ПК-3 | Способность обосновывать принимаемые проектные решения, осуществлять постановку и выполнять эксперименты по проверке их корректности и эффективности |
| 3 | ПК-4 | Способность готовить конспекты и проводить занятия по обучению работников применению программно-методических комплексов, используемых на предприятии |

.

*Руководитель практики от каф.319*: Фомичев В.А. \_\_\_\_\_\_/\_\_\_\_\_\_\_\_\_/

*(фамилия, имя, отчество) подпись)*

«\_10\_»\_мая\_2022 г.

**6.Отчет студента о практике**

**Семантически-ориентированный естественно-языковой интерфейс для взаимодействия с Системой взаимосвязанных открытых данных (Linked Open Data)**

В современном мире всё большее распространение получают связные данные в связи с внедрением соответствующих принципов в Интернете. Связность данных обеспечивается в системе взаимосвязанных открытых данных (Linked Open Data). В качестве внутреннего формата хранения и представления используется RDF.

RDF (Resource Description Framework) – формат описания ресурсов и их взаимосвязей. Записи в формате RDF представляют собой тройки вида {Субъект, Отношение, Объект}. Субъект и объект являются элементами множества ресурсов, а отношение – множества свойств (подмножества ресурсов, описывающих свойства и отношения между ресурсами). Для обращения к взаимосвязанным данным в формате RDF используется язык SPARQL.

SPARQL – язык запросов к взаимосвязанным данным в формате RDF. Имеет SQL-подобную структуру запроса, а именно «SELECT FROM WHERE». Иными словами, запрос на языке SPARQL определяет какой ресурс является искомым, в каком источнике проводить поиск и каким параметрам должен отвечать искомый ресурс.

Учитывая тот факт, что для обращения к LOD необходимо владение языком SPARQL и хотя бы иметь представление о формате RDF, использование системы становится невозможным или сильно затруднённым для людей, не имеющих опыта использования специальных языков запросов, таких как SQL или SPARQL. В свою очередь LOD была бы полезна в разнообразных отраслях, таких как медицина, здравоохранение, научные исследования и бизнес, т.к. обеспечивает семантическую связь между разными ресурсами, в качестве которых могут выступать научные публикации, разнообразные отчеты компаний и законодательные акты. В данной ситуации очевидным решение является разработка интеллектуального естественно-языкового интерфейса, обеспечивающего обращение к LOD без специальных технических навыков, что позволит расширить использование LOD в отраслях, где система уже используется, и внедрить, где не используется. Но таких естественно-языковых интерфейсов, использующих в качестве языка для запросов русский нет, поэтому темой данной работы ставится разработка такого интерфейса.

Для использования естественно-языкового интерфейса при обращении к LOD необходимо реализовать перевод запроса с естественного языка на язык запросов SPARQL. Поскольку данные в LOD связываются семантически, то и при переводе необходимо учитывать семантику запроса. Следовательно, при переводе требуется промежуточный язык, позволяющий описать семантическое представление запроса на естественном языке, которое далее будет переведено в конструкции языка SPARQL.

Существует несколько подходов описания семантического представления, а именно: AMR, грамматика Монтегю и К-представление Фомичева В.А. (концептуальное представление). AMR и грамматика Монтегю специализированы на семантическом представлении английского языка и с их помощью возможно описать только семантическое представление простых предложений. В свою очередь К-представление представляет формализованный аппарат для описания семантического представления текстов (дискурсов), а не только отдельных предложения, на русском языке. Учитывая все выше написанное, для описания семантического представления запроса на естественном языке выбран подход К-представление.

Владимир Александрович Фомичев в своей научной монографии «Формализация проектирования лингвистических процессоров» в рамках своей теории К-представлений формулирует новый метод преобразования ЕЯ-текста в СП текста [13] (Фомичев 2005). Данный метод предназначается для разработки ЕЯ-интерфейсов для взаимодействия с прикладными интеллектуальными системами. Указанный выше метод состоит из трех частей (этапов) преобразования.

На первом этапе осуществляется компонентно-морфологический анализ входного текста. По входному тексту строится не менее одного компонентно-морфологических представления (КМП), каждое из которых представляет из себя пару (набор) классифицирующего и морфологического представлений текста. Последнее представляет возможные значений морфологических признаков для лексических единиц входной текста. Чаще всего отдельным фразам соответствует единственное КМП. В ситуации неоднозначного разбиения на значащие единицы или невозможности однозначного определения части речи какой-либо единицы пользователю задаются уточняющие вопросы, ответы на которые позволяют снять неоднозначность.

Второй этап предполагает построение матричного семантико-синтаксического представления (МССП). На этом этапе с каждым словом связывается одно из возможных значений и устанавливаются смысловые отношения между единицами текста.

МССП сначала недоопределенно и неопределенность снимется шаг за шагом. Для снятия неопределенности используется информация из лингвистической базы данных (ЛДБ) о возможных способах соединения единиц текста в лингвистически допустимые сочетания.

На заключительном, третьем этапе производится сборка семантического представления текста, которое является его   
К-представлением, по его МССП.

С помощью описанного выше метода возможно устанавливать смысловые соотношения в сочетаниях типа: «Глагол + Предлог + Существительное», «Глагол + Существительное», «Существительное1 + Предлог + Существительное2», «Число + Существительное», «Прилагательное + Существительное», «Существительное1 + Существительное2», «Причастие + Существительное», «Причастие + Предлог + Существительное», «Вопросительно-относительное местоимение или местоименное наречие + Глагол», «Предлог + Вопросительно-относительное местоимение + Глагол».

Довольно важно отметить, что данный метод учитывает многозначность слов, что чрезвычайно важно при проектировании ЕЯ-интерфейсов для взаимодействия с прикладными интеллектуальными системами. А также не предполагает использование синтаксического уровня представления текста (как следствие и синтаксического анализа), хоть синтаксический уровень и используется в течение долго времени как отечественными, так и зарубежными исследователями.

Для выполнения поставленных условий ЛБД должна состоять из 4 компонентов: морфологической базы данных, лексико-семантического словаря, словаря глагольно-предложных семантико-синтаксических фреймов и словаря предложных семантико-синтаксических фреймов

В ходе разработки алгоритма преобразования запроса на естественном языке в запрос на языке SPARQL пришлось столкнуться с рядом проблем, потребовавших дополнительного исследования.

В рамках одной онтологии возможно одновременное наличие различных именований одного и того же предиката, что не позволяет делать унифицированные по структуре запросы на языке SPARQL для запросов на естественном языке даже c одинаковой структурой. Например, в онтологии DBpedia в информации о городе предиката связывающий данный город с количеством жителей может иметь одно из следующих именований: «population», «populationTotal», «p», «pop2010census» (пример представлен в таблице 1). Причем предикат «p» при описании города может использоваться как в значении «количество населения», так и – в значении «название района города». Выше описанная проблема делает затруднительным программное построение запроса на языке SPARQL.

Таблица 1 – Предикаты «Население» у разных городов

|  |  |
| --- | --- |
| **Город** | **Предикат** |
| Оттава | population |
| Москва | populationTotal |
| Ульяновск | p |
| Северодвинск | pop2010census |

Проанализировав работы других исследователей можно выделить два применяемых способа решения описанной выше проблемы. Первый заключается в создании собственной онтологии, зачастую на основе данных других онтологий, но со своими классами объектов и предикатами. Поскольку такая онтология создается непосредственно для целевой системы, в нее закладывается необходимая семантическая структура и система имен, что позволяет унифицировать генерацию запроса на языке SPARQL. Но данный подход требует больших затрат времени на разработку собственной онтологии, поэтому в рамках данной работы применятся не будет.

Также в проектах, целью которых является обращение к системе LOD на естественном языке, применяются онтологии, использующие строго описанный набор классов и предикатов. Примером такой онтологии является YAGO, система классов и типов которой основана на наборе классов и предикатов Schema (специальная онтология, описывающая логическую структуру семантической связанности объектов реального мира). Данная онтология предоставляет возможность генерации унифицированных по структуре запросов на языке SPARQL. Но онтология YAGO содержит большое количество слабо связанных данных, т.е. некоторая часть данных об объектах реального мира просто не указана или указана не корректно. Например, в описании городов связь со страной, в которой находится данных город, указывается с помощью предиката комментария, значение которого является строка приблизительно следующего содержания: «город в России» или «Казахстанский город». Хотя, придерживаясь принципов LOD, необходимо было сделать предикат, например, с название «страна», значением которого являлась бы ссылка (URI) на описание необходимой страны.

Для преодоления проблемы неоднозначности именования параметров (отношений и их значений) в онтологии, к которой предназначается запрос, в рамках данной работы использовался описанный далее подход.

Суть упомянутого подхода заключается в предварительном связывании параметров, использующихся в исходном запросе, с соответствующими параметрами, используемыми в онтологии. Для этого в рамках базы данных организуются таблица связи, которая будет содержать информацию о связи параметров и отношений К-представления и онтологии (пример таблицы связи представлен в таблице 2). При этом, одному параметру исходного запроса может быть поставлено в соответствие несколько параметров, использующихся в онтологии. Например, отношению «Колич-Жителей», обозначающему количество жителей конкретного города (страны и т.п.) в соответствие может быть поставлены отношения «population», «populationTotal», «p» и «pop2010census» онтологии DBpedia. А значению «Россия» – «dbr:Russia».

Таблица 2 – Связь параметров в К-представлении в соответствующими параметрами в онтологии

|  |  |
| --- | --- |
| **Параметр в К-представлении** | **Параметр в онтологии** |
| Колич-Жителей | population |
| populationTotal |
| p |
| pop2010census |
| Россия | dbr: Russia |

При построении SPARQL-запроса по К-представлению запроса на естественном языке для перевода параметров запроса в термины онтологии организуется запрос к базе данных, содержащей таблицу связи параметров, возвращающий все возможные параметры, используемые в онтологии, соответствующие данному параметру исходного запроса. Далее, с помощью специальной конструкции языка SPARQL из них образуется множество, выступающее в запросе самостоятельной смысловой единицей. Например, отношения «population», «populationTotal», «p» и «pop2010census» в запросе будут выступать не в качестве отдельных отношений, а в качестве одного отношения со значением «количество жителей».