УЧРЕЖДЕНИЕ ОБРАЗОВАНИЯ
«МИНСКИЙ ГОСУДАРСТВЕННЫЙ ЛИНГВИСТИЧЕСКИЙ УНИВЕРСИТЕТ»

УДК 811. 111’322.2 (043.3)

ГОЛЯК
Юлия Дмитриевна

ПРЕДИКТИВНОЕ АВТОДОПОЛНЕНИЕ ЗАПРОСОВ ПОЛЬЗОВАТЕЛЕЙ В СИСТЕМАХ ИНФОРМАЦИОННОГО ПОИСКА
(на материале русского языка)

Автореферат диссертации на соискание ученой степени
кандидата филологических наук

по специальности 10.02.21 – прикладная и математическая лингвистика

Минск, 2023

Научная работа выполнена в Белорусском государственном университете

Научный руководитель: **Совпель Игорь Васильевич**, доктор технических наук, профессор

Официальные оппоненты: **Беляева Лариса Николаевна**, доктор филологических наук, профессор, академик РАЕН, заслуженный деятель науки Российской Федерации, ГОУ ВПО «Российский государственный педагогический университет им. А. И. Герцена», кафедра образовательных технологий в филологии

Детскина Раиса Владимировна, кандидат филологических наук, доцент, УО «Минский государственный лингвистический университет», кафедра информатики и прикладной лингвистики

Оппонирующая организация: УО «Гродненский государственный университет имени Янки Купалы»

Защита состоится 13 марта 2023 года в 10:00 на заседании совета по защите диссертаций К 02.22.02 при Минском государственном лингвистическом университете по адресу: 220034, г. Минск, ул. Захарова, 21, ауд. Б-303, e-mail: info@mslu.by. Телефон ученого секретаря: 289-46-43.

С диссертацией можно ознакомиться в библиотеке учреждения образования «Минский государственный лингвистический университет».

Автореферат разослан «8» февраля 2023 года.

Ученый секретарь
совета по защите диссертаций,
кандидат филологических наук, доцент Н.В. Михалькова

ВВЕДЕНИЕ

Возможности развитого автоматического лингвистического анализа текста обеспечили в последнее время широкое распространение такого класса систем информационного поиска (СИП), как вопросно-ответные системы с естественно-языковым (ЕЯ) интерфейсом пользователя, ориентированные на поисковое пространство в виде полнотекстовых баз данных (ПБД) большого объема, прежде всего корпоративных ПБД. Преимущества такого интерфейса вполне очевидны: необходимость минимальной подготовки пользователя для работы с СИП, высокая скорость формулирования ПЗ, в том числе за счет его автоматизации, фактически неограниченные возможности в формулировании на естественном языке информационной потребности пользователя, а в конечном счете – получение качественного решения поисковой задачи. Именно поэтому разработка и использование ЕЯ-интерфейса пользователя стали устойчивой тенденцией в построении современных СИП, а одним из главных компонентов в его составе является автодополнение (автоматическое завершение, предиктивный ввод) ПЗ, в дальнейшем QTA (query type-ahead).

Существующие решения задачи QTA в подавляющем большинстве ориентированы на историю поиска – персональную или общую статистику поисковых запросов в СИП, что, как оказалось, не всегда является оптимальным, а иногда и просто невозможным. Также предлагаемые системами подсказки для автодополнения ПЗ чаще всего исходят из принципа последовательного завершения строки вводимого запроса, рассматриваемой в качестве его префикса, а не погружения ее в более общий контекст, что существенно снижает релевантность реакции системы по отношению к информационной потребности пользователя. Практически все существующие решения задачи QTA не ориентированы на использование лингвистических ресурсов, прежде всего средств автоматического лингвистического анализа текста, при том что в ряде случаев сами алгоритмы индексирования ПЗ и текстовых документов в СИП основываются на таковых ресурсах. Не учитываются типы ПЗ и их синтаксических структур, а также само наполнение ПБД с целью автоматического построения базы подсказок, что существенно сдерживает качественное решение задачи QTA.

Настоящая диссертационная работа посвящена исследованию и решению, с учетом указанных выше недостатков, задачи автодополнения русскоязычных ПЗ в СИП, работающих с ПБД, и его реализации в виде промышленного прототипа.

ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

Связь работы с крупными научными программами и темами. Тема диссертации соответствует приоритетным направлениям «Информационно-управляющие системы» и «Искусственный интеллект и робототехника» согласно пункту 1 приоритетных направлений научной, научно-технической и инновационной деятельности на 2021-2025 годы (Указ Президента Республики Беларусь от 7 мая 2020 г. № 156).

Диссертационное исследование выполнялось на кафедре прикладной лингвистики БГУ в рамках госбюджетной НИР «Комплексное научно-методическое обеспечение преподавания русского языка как иностранного в контексте межкультурной коммуникации» в соответствии с пунктом 11 «Общество и экономика» приоритетных научных исследований БГУ на 2016-2020 годы, а также в соответствии с программой научных исследований отдела разработки средств интеллектуализации информационных систем ООО «АйЭйчЭс Глобал».

Цель и задачи исследования. *Целью* диссертационной работы является разработка метода, алгоритмов и лингвистического обеспечения решения задачи предиктивного автодополнения русскоязычных запросов пользователей в системах информационного поиска и их реализации в виде промышленного прототипа.

Для достижения поставленной цели необходимо решить следующие основные *задачи*:

– сформулировать концепцию решения задачи QTA, свободного от недостатков уже существующих решений и ориентированного на промышленный характер его приложения и высокие показатели эффективности;

– определить необходимый для исследования текстовый материал и, опираясь на его анализ, дать классификацию основных типов ПЗ и их базовых синтаксических структур;

– определить метод решения задачи QTA, сформулировать основные требования к ее базовому лингвистическому обеспечению и разработать структурно-функциональную схему системы QTA;

– разработать, в виде расширения базового, собственное лингвистическое обеспечение и алгоритмы автоматического распознавания в ПБД подсказок, соответствующих базовым синтаксическим структурам основных типов ПЗ;

– построить прототип системы QTA, осуществить ее тестирование, определить качественные характеристики и внедрить в промышленную эксплуатацию.

Объектом исследования является естественно-языковой интерфейс пользователя в системах информационного поиска.

Предметом исследования является автоматическое дополнение русскоязычных пользовательских запросов в корпоративных системах информационного поиска.

Материалом исследования являются два корпуса:

1) корпус ПЗ, включающий (1) набор данных RuBQ 2.0, содержащий 3000 русскоязычных поисковых запросов в форме вопросов; (2) набор из 2500 запросов на английском языке Query-wellformedness Dataset; (3) набор из 800 англоязычных ПЗ к базе Topic Explorer; (4) список англоязычных ПЗ в количестве 1200 единиц, доступных на платформе Zenodo; (5) список из 2300 англоязычных запросов, доступный на платформе Kaggle; (6) набор русскоязычных ПЗ, полученных в системе Wordstat.Yandex в количестве 3200 ПЗ;

2) корпус русскоязычных текстовых документов, являющийся информационной моделью корпоративной ПБД и служащий источником для формирования базы подсказок, а также для разработки и тестирования алгоритмов автоматического распознавания подсказок в ПБД, включающий в себя выборку документов различных областей знаний корпоративной ПБД компании IHS Markit и ее дополнение документами из открытых источников (энциклопедия «Википедия», научные статьи из научной электронной библиотеки «КиберЛенинка», новостные статьи с сайта Reuters и др.), общим объемом более 180 миллионов словоупотреблений, около 7,5 миллионов предложений и более 0,4 гигабайт текста в заархивированном виде.

Всего для исследования были образованы корпус ПЗ, включающий 6200 русскоязычных и 6800 англоязычных ПЗ, и корпус русскоязычных текстовых документов общим объемом около 7,5 миллионов предложений.

Научная новизна

1. Сформулирована и обоснована концепция решения задачи QTA, новизна которой заключается в ее ориентации на: промышленный характер приложения и высокие показатели эффективности; вопросно-ответную СИП с ЕЯ-интерфейсом пользователя и русскоязычную ПБД большого объема, являющуюся поисковым пространством СИП; использование «истории» не уже осуществленного, а предполагаемого поиска в виде множества подсказок, заранее автоматически распознаваемых в текстовых документах (ТД) из ПБД; не только последовательное завершение уже набранной части ПЗ, а погружение ее в контекст, а именно дополнение в начале, в конце и даже внутри; глубокий, вплоть до семантического уровня ЕЯ, лингвистический анализ ТД.

2. Впервые на основе анализа используемого для исследования

текстового материала дана классификация основных типов ПЗ (одно или несколько несогласованных ключевых слов; грамматически согласованные словосочетания; вопросительные предложения; утвердительные предложения; комбинация двух последних типов, нередко без знака препинания между ними и чаще без вопросительного знака в конце ПЗ) и их базовых синтаксических структур (именные группы (простые именные группы, расширенные именные группы с предложно-падежными зависимыми конструкциями), глагольные группы (инфинитив глагола с прямым дополнением, с косвенным дополнением с предлогом, с прямым и косвенным дополнениями, с прямым дополнением и причастным оборотом; форма 3 лица единственного или множественного числа глагола с прямым дополнением, с косвенным дополнением с предлогом, с прямым и косвенным дополнениями, с прямым дополнением и причастным оборотом), грамматическая основа предложения (предикативный центр)).

3. Определен метод решения задачи QTA, который, в отличие от существующих, основан на классификации основных типов ПЗ и их базовых синтаксических структур, сводится к автоматическому распознаванию этих структур в ПБД и выбору их лексических наполнений, которые в совокупности составляют базу подсказок для автодополнения ПЗ.

4. Разработана структурно-функциональная схема системы QTA/R автодополнения русскоязычных ПЗ в СИП и определен ее базовый лингвистический процессор. Новизна заключается в самой функциональности этой системы (базового лингвистического анализа ПБД с целью построения лингвистического индекса ТД в виде БД САО-отношений и автоматического распознавания в ней подсказок с целью формирования базы подсказок), ориентированной на сформулированную концепцию и предложенный метод решения задачи QTA и основанной на лингвистическом процессоре задачи, включающем БЛП и его расширение.

5. Разработано, в виде расширения базового, собственное лингвистическое обеспечение и алгоритмы автоматического распознавания в ПБД подсказок, соответствующих базовым синтаксическим структурам основных типов ПЗ, с целью их автодополнения. Новизна заключается в том, что такого рода ресурсы впервые используются для решения задачи QTA.

6. Разработан прототип оригинальной системы QTA/R предиктивного автодополнения русскоязычных ПЗ в СИП. Ее отличительными чертами являются: предоставление пользователю возможности более точно формулировать свою информационную потребность, существенно минимизировать общее время решения поисковой задачи, решать задачу QTA, не ориентируясь на историю информационного поиска, получать гарантированно релевантную реакцию поисковой системы. Осуществлено

тестирование и внедрение системы в промышленную эксплуатацию, которое показало, что она, в силу использования разработанных концептуальных и алгоритмических решений, а также лингвистических ресурсов, обладает высокими качественными характеристиками.

Положения, выносимые на защиту

1. Концепция решения задачи QTA, ориентированная на вопросно-ответную СИП с ЕЯ-интерфейсом пользователя и русскоязычную ПБД большого объема, являющуюся поисковым пространством СИП, на промышленный характер приложения и высокие показатели эффективности.

2. Классификация основных типов пользовательских запросов в СИП и их базовых синтаксических структур, полученная в результате анализа используемого для исследования текстового материала.

3. Метод решения задачи QTA, который, в отличие от существующих, основан на классификации основных типов ПЗ и их базовых синтаксических структур.

4. Структурно-функциональная схема системы QTA/R автодополнения русскоязычных пользовательских запросов в СИП, основанная на сформулированной концепции и предложенном методе решения задачи QTA.

5. Расширение базового лингвистического обеспечения задачи и алгоритмы автоматического распознавания в ПБД подсказок, соответствующих базовым синтаксическим структурам основных типов ПЗ.

6. Прототип оригинальной системы QTA/R автодополнения русскоязычных ПЗ в СИП, его качественные характеристики и результаты внедрения в промышленную эксплуатацию.

Личный вклад соискателя ученой степени в результат диссертации. Все основные результаты и положения, выносимые на защиту, получены автором самостоятельно и составляют его личный вклад в исследование темы диссертации. Научный руководитель принимал участие в выборе направления исследования, постановке задач, обсуждении теоретических и практических результатов, полученных автором.

Апробация диссертации и информация об использовании ее результатов. Основные результаты диссертационной работы докладывались и обсуждались на: заседаниях кафедры прикладной лингвистики БГУ в 2016–2021 гг.; XI Карповских научных чтениях (г. Минск, Беларусь, 17–18 марта 2017 г.); IX Международной научной конференции «Молодые ученые в инновационном поиске» (г. Минск, Беларусь, 27–28 мая 2020 г.); V Международной научно-практической конференции «Лингвистика, лингводидактика, лингвокультурология: актуальные вопросы и перспективы развития» (г. Минск, Беларусь, 18–19 марта 2021 г.); XI Международной научно-технической конференции «Открытые семантические технологии для

проектирования интеллектуальных систем» (OSTIS, Open Semantic Technology for Intelligent Systems) (г. Минск, Беларусь, 16–18 сентября 2021 г.).

Результаты диссертационной работы внедрены в учебный процесс в Минском государственном лингвистическом университете, а разработанная система QTA/R внедрена в состав известной многоязычной информационно-поисковой платформы IHS Goldfire, используемой для решения инновационных задач многими крупнейшими компаниями мира, что подтверждается соответствующими актами.

Опубликованность результатов диссертации. По материалам выполненных исследований опубликовано 8 научных работ, в том числе 5 статей (две из них в соавторстве) в рецензируемых изданиях (из них четыре – в изданиях по филологическим наукам и одна – по техническим), 2 статьи в сборниках научных статей и 1 публикация в виде материалов научной конференции. Общий объем опубликованных материалов составляет 4,21 авторского листа.

Структура и объем диссертации. Диссертация состоит из введения, общей характеристики работы, трех глав с выводами по каждой из них, заключения, библиографического списка, списка публикаций автора и пяти приложений. Полный объем диссертации составляет 159 страниц машинописного текста, из них 100 страниц занимает основной текст, включая 13 таблиц на 17 страницах и 17 рисунков на 9 страницах, библиографию из 115 источников на русском и английском языках на 11 страницах, включая 8 публикаций соискателя, 5 приложений на 59 страницах.

ОСНОВНАЯ ЧАСТЬ

Во введении обоснована актуальность темы диссертационной работы, обозначены цель и основные задачи исследования.

В первой главе даны общее описание задачи QTA и результаты проведенного анализа существующих ее решений. Отмечено, что разработка и использование ЕЯ-интерфейса пользователя, актуальную роль в котором играет автодополнение ПЗ, стали устойчивой тенденцией в построении современных СИП, доминирующими среди которых становятся вопросно-ответные системы, ориентированные на работу с большими по объему и относительно постоянными по составу корпоративными полнотекстовыми БД, специальными фондами и т. п. Актуальность задачи QTA обусловлена тем, что ее решение обеспечивает, во-первых, существенное сокращение общего времени, затрачиваемого пользователем на ввод ПЗ, во-вторых,

повышение качества формируемого запроса с точки зрения информационной потребности пользователя. Существующие решения задачи QTA в подавляющем большинстве обладают следующими ограничениями:

– ориентированы на историю поиска – персональную или общую статистику поисковых запросов, что иногда в принципе невозможно, поскольку такая история представляет собой коммерческую тайну клиента;

– чаще всего исходят из принципа последовательного завершения строки вводимого запроса, рассматриваемой в качестве его префикса, а не погружения ее в более общий контекст;

– не ориентированы на использование лингвистических ресурсов, прежде всего средств автоматического лингвистического анализа текста;

– не учитывают типы ПЗ и их синтаксических структур, а также само наполнение ПБД с целью автоматического построения базы подсказок.

Очевидно, что указанные ограничения не способствуют построению качественного ПЗ. В силу изложенного, в данной работе имеет место следующая постановка задачи.

Пусть вводимый пользователем запрос в определенный момент времени представляет собой цепочку (формула 1):

$$W_1 W_2 \ldots W_n \qquad (1)$$

где каждое W_i, $i = 1, \ldots, n-1$, $n > 1$, является словом естественного языка, а W_n, $n \geq 1$ – словом либо префиксом слова.

Задача автодополнения запроса вида (формула 1) заключается в автоматическом формировании списка таких грамматически и семантически корректных цепочек слов естественного языка, которые включают в себя члены цепочки (формула 1), то есть речь идет не только о последовательном завершении строки вводимого запроса, рассматриваемого в качестве его префикса, но о куда более сложной процедуре, предполагающей, в общем случае, погружение уже набранной пользователем части запроса в контекст, а именно дополнение ее в начале, в конце и даже внутри.

Для достижения поставленной цели, с учетом того, что пользователи в большинстве своем следуют уже сложившейся практике формирования поисковых запросов и что они традиционно в той или иной степени ориентированы на ключевой поиск, необходимо решить следующие основные задачи:

– сформулировать общую концепцию решения задачи QTA, ориентированную на ее промышленный характер и высокие показатели эффективности, и определить необходимый для исследования текстовый материал;

– осуществить экспертный анализ наиболее частотных поисковых ЕЯ-запросов и дать классификацию их типов;

– дать классификацию синтаксических структур ПЗ для каждого их типа;

– сформулировать основные требования к лингвистическому обеспечению решаемой задачи QTA, определить ее базовый лингвистический процессор и разработать метод ее решения;

– разработать алгоритмы и, в дополнение к базовому, собственное лингвистическое обеспечение для автоматического распознавания в ПБД подсказок для каждого типа ПЗ с целью построения их множества Р, что и составляет решение целевой задачи;

– осуществить внедрение полученного решения в состав промышленногопрототипа.

В целом, обоснованно предложенная концепция решения задачи QTA состоит в его ориентировании на промышленный характер приложения и высокие показатели эффективности; вопросно-ответную СИП с ЕЯ-интерфейсом пользователя и русскоязычную ПБД большого объема, являющуюся поисковым пространством СИП; использование «истории» не уже осуществленного, а предполагаемого поиска в виде множества подсказок, заранее автоматически распознаваемых в ТД из ПБД; не только на последовательное завершение уже набранной части ПЗ, а на погружение ее в контекст, а именно дополнение в начале, в конце и даже внутри; глубокий, вплоть до семантического уровня ЕЯ, лингвистический анализ ТД.

Во второй главе рассмотрен круг вопросов, касающихся формирования необходимого для исследования текстового материала, классификации основных типов ПЗ и их базовых синтаксических структур, метода решения задачи, а также функциональности базового лингвистического процессора, предоставляющего необходимый для решения целевой задачи уровень автоматического лингвистического анализа текстовых документов.

Исходя из поставленной задачи, необходимый для исследования текстовый материал включил, во-первых, корпус ПЗ, обеспечивающий формирование представления о лексической, синтаксической и семантической природе наиболее частотных ПЗ, во-вторых, корпус текстовых документов, который фактически является информационной моделью корпоративной ПБД и служит информационным источником для формирования базы подсказок как основы решения задачи автодополнения ПЗ, а также для разработки и тестирования алгоритмов автоматического распознавания подсказок в ПБД для различных типов ПЗ. В основу формирования корпуса ПЗ был положен целый ряд «прямых» открытых

источников, из которых, в общей сложности, получено 3000 русскоязычных ПЗ и 6800 англоязычных ПЗ. Качественным «косвенным» источником для формирования корпуса ПЗ стала общедоступная система Wordstat.Yandex, которая представляет собой не просто открытый список ПЗ, а базу данных русскоязычных ПЗ поисковой системы Yandex с интерактивным интерфейсом и, таким образом, позволяет формировать любого объема массивы русскоязычных ПЗ (в нашем случае 3200 ПЗ) с информацией об их частотности для различных синтаксических структур и синонимической вариативности. В основу формирования корпуса текстовых документов была положена доступная нам корпоративная ПБД транснациональной IT-компании IHS Markit, которая содержит ТД из самых разнообразных областей знаний (стандарты, патенты, научные статьи и книги и т. д.) на английском, немецком, французском, японском, китайском и русском языках. В целях нашего исследования из данной ПБД была сделана выборка, несколько дополненная из других доступных источников (открытая энциклопедия «Википедия», научные статьи из рецензируемых журналов, новостные статьи с сайта Reuters и др.) в виде корпуса ТД общим объемом более 75 миллионов предложений.

На основе экспертного анализа построенного корпуса поисковых запросов определены основные их типы: одно или несколько несогласованных ключевых слов; грамматически согласованные словосочетания; вопросительные предложения; утвердительные предложения; комбинация двух последних типов, нередко без знака препинания между ними и чаще без вопросительного знака в конце ПЗ. Последующий экспертный анализ показал, что запросам каждого из приведенных типов соответствуют определенные синтаксические структуры, базовыми составляющими которых являются: именные группы (простые именные группы, расширенные именные группы с предложно-падежными зависимыми конструкциями), глагольные группы (инфинитив глагола с прямым дополнением, с косвенным дополнением с предлогом, с прямым и косвенным дополнениями, с прямым дополнением и причастным оборотом; форма 3 лица единственного или множественного числа глагола с прямым дополнением, с косвенным дополнением с предлогом, с прямым и косвенным дополнениями, с прямым дополнением и причастным оборотом), грамматическая основа предложения (предикативный центр). Полученная классификация основных типов ПЗ и их базовых синтаксических структур фактически предопределяет метод решения целевой задачи, который сводится к автоматическому распознаванию этих структур на основе лингвистического анализа в текстовых документах из ПБД как предполагаемого поискового пространства и выбору их лексических

наполнений, которые и составляют базу подсказок для автодополнения ПЗ.

В качестве требуемого, в соответствии с представленным методом решения автодополнения ПЗ, лингвистического процессора определен многоязычный базовый лингвистический процессор (БЛП) IHS Goldfire, поскольку показано, что компоненты распознаваемых им так называемых САО-отношений в результате лексического, лексико-грамматического, синтаксического и семантического анализа входного текста в значительной степени коррелируют с базовыми синтаксическими структурами ПЗ. Имеет место следующая структура САО-отношений (приводится на примере предложения: *Антикоррозийное покрытие аэрозольного нанесения, создающее защитный слой, предотвращает повреждения металла, вызываемые коррозией*) (таблица 1).

Таблица 1. – Структура САО-отношений

Поле	САО 1	САО 2	САО 3
ОС	–	повреждения металла	антикоррозийное покрытие аэрозольного нанесения
С	–	–	–
П	антикоррозийное покрытие аэрозольного нанесения	коррозия	антикоррозийное покрытие аэрозольного нанесения
Ск	предотвращать	вызывать	создавать
ИчС	–	–	–
ПД	повреждения металла	повреждения металла	защитный слой
ДДП	–	–	–
ДТП	–	–	–
Пр	–	–	–
КДП	–	–	–
О	–	–	–
ВК	–	–	–
ОПС	предотвращает	вызываемые	создающее

Здесь ОС – Опорное слово (компонент, через который одно САО-отношение предложения связывается с другим, если таковое имеет место), С – Союз, П – Подлежащее, Ск – Сказуемое (в форме инфинитива), ИчС – Именная часть Сказуемого, ПД – Прямое дополнение, ДДП – Дополнение в дательном падеже, ДТП – Дополнение в творительном падеже, Пр – Предлог, КДП – Косвенное дополнение с Предлогом, О – Обстоятельство, ВК – Вводная конструкция, ОПС – Оригинальное представление сказуемого в тексте.

Именно конкретное наполнение полей САО-структуры является исходным материалом, на основе анализа которого и учета приведенной классификации базовых синтаксических структур ПЗ осуществляется их автоматическое построение для наиболее частотных ПЗ.

В целом, БЛП и лингвистические ресурсы последующего расширения его функциональности (определенной фильтрации САО-отношений и преобразования их отдельных компонентов, а также автоматического синтеза на их основе базы подсказок для автодополнения ПЗ наиболее актуальных типов) составляют лингвистическое обеспечение решения целевой задачи.

В третьей главе результаты, представленные в предыдущих главах диссертации, были положены в основу практической реализации системы QTA/R автодополнения русскоязычных ПЗ.

Разработана структурно-функциональная схема системы QTA/R (рисунок 1), согласно которой автодополнение ПЗ осуществляется на этапе взаимодействия пользователя с базой подсказок, создающейся предварительно и автоматически. С этой целью используется ПБД, с которой, предпочтительно будет работать СИП с ЕЯ-интерфейсом пользователя. Лингвистический процессор (ЛП) задачи, включающий функциональность БЛП и ее расширение, осуществляет базовый лингвистический анализ текстов из русскоязычной ПБД, результатом которого является БД SAO-отношений с конкретным лексическим наполнением из ПБД. Далее, тот же ЛП задачи осуществляет в соответствии с разработанными алгоритмами и классификацией основных типов ПЗ и их базовых синтаксических структур автоматическое распознавание подсказок в БД SAO-отношений и формирует базу подсказок системы QTA/R. Собственно, автодополнение ПЗ инициируется системой на этапе его ввода в соответствии со следующей схемой. Контролируя процесс ввода запроса, система в какой-то момент времени t_1 (он определяется ее состоянием «уже готова что-то предложить») предлагает Пользователю продолжение части запроса q_1, введенной им к этому моменту. Пользователь принимает определенное решение (принять подсказку или нет) и продолжает ввод далее. В момент времени t_2 цикл повторяется относительно части запроса q_2 и т. д., вплоть до момента времени t_n, когда окончательно формируется запрос Пользователя q_n, который затем поступает в основной блок СИП. Обосновано и построено расширение функциональности БЛП в виде совокупности правил (базовых процедур БП1, БП2, БП3, БП4, БП5) и соответствующих лингвистических ресурсов в виде специальных списков и словарей, а также множества паттернов, обеспечившее эффективность алгоритмов автоматического построения базы подсказок и ее высокое качество:

– БП1. Исключение определенных САО-отношений и их полей из

списка кандидатов для формирования подсказок;

– БП2. Фильтрация именных групп САО-отношений;

– БП3. Обработка именных групп с зависимой именной частью в косвенных падежах;

– БП4. Обработка именных групп с согласованными атрибутами;

– БП5. Обработка главных слов именных групп.

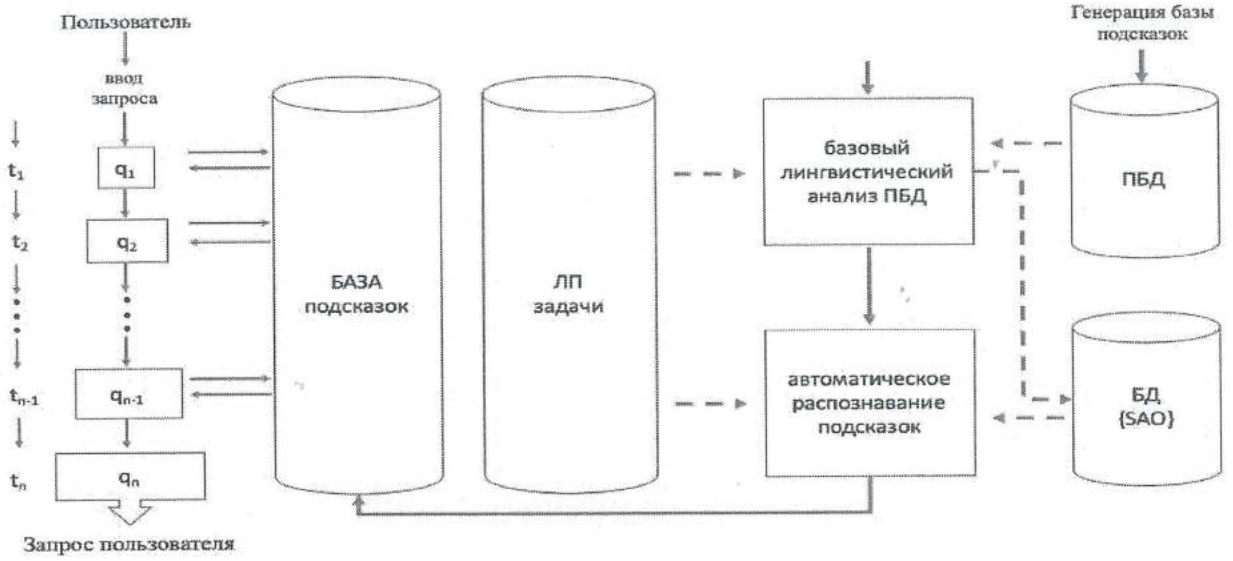


Рисунок 1. – Структурно-функциональная схема системы QTA/R

Представленные БП, как расширение функциональности БЛП, являются важными составляющими алгоритмов автоматического построения списков подсказок для автодополнения ПЗ, формулируемых с использованием базовых синтаксических структур.

Разработан алгоритм автоматического построения списка подсказок типа «простая именная группа». Источником для формирования списка подсказок в данном случае являются наполнения таких полей САО-отношений, как: *Опорное слово, Подлежащее, Прямое дополнение, Дополнение в дательном падеже, Дополнение в творительном падеже, Косвенное дополнение с предлогом.* Сам алгоритм реализуется последовательным выполнением базовых процедур БП1, БП2, БП3, БП4, БП5.

Разработаны алгоритмы автоматического построения списков подсказок типа «расширенная именная группа» («расширенная именная группа с предложно-падежной зависимой конструкцией» и «расширенная именная группа с причастным оборотом»). Здесь ИГ являются двусоставными. В первом случае их главная именная группа формируется из наполнения таких полей САО-отношений, как *Подлежащее, Прямое дополнение, Дополнение в дательном падеже, Дополнение в творительном падеже*, а предложно-

падежная группа – полей *Предлог* и *Косвенное дополнение*. В основу реализации алгоритма положены БП2 и БП5, словарь многозначных составных предлогов, а также паттерны, описывающие характерную лексическую сочетаемость с целью разрешения семантической многозначности. Во втором случае главная ИГ подсказки формируется, как и в первом случае, а источником для формирования ее причастного оборота являются такие поля САО-отношений, как *Оригинальное представление сказуемого в тексте, Прямое дополнение, Дополнение в дательном падеже, Дополнение в творительном падеже, Косвенное дополнение с предлогом.* В основу реализации алгоритма, наряду с обычными операциями с наполнением указанных полей, также положены БП2 и БП5.

Разработан алгоритм автоматического построения списка подсказок типа «глагольная группа». Как и выше, здесь также речь идет о двусоставных конструкциях. Первая их часть представлена собственно сказуемым, а вторая – дополнениями разных видов. Источником формирования первой части являются наполнения таких полей САО-отношений, как *Сказуемое* и *Оригинальное представление сказуемого в тексте,* а второй – *Опорное слово, Прямое дополнение, Дополнение в дательном падеже, Дополнение в творительном падеже, Косвенное дополнение с предлогом.* Имеет место следующая постадийная схема алгоритма:

– Стадия А1. Фильтрация наполнения глагольных групп в зависимости от сказуемого;

– Стадия А2. Преобразование многословного сказуемого;

– Стадия А3. Формирование глагольных групп с прямым и косвенными дополнениями, а также их комбинациями;

– Стадия А4. Формирование глагольных групп с прямым дополнением и причастным оборотом.

В основу реализации алгоритма положены операция объединения наполнения указанных ранее полей САО-отношений, а также разработанные экспертные правила и список неинформативных глаголов, обеспечившие выполнение стадии А1 алгоритма.

Разработан алгоритм автоматического построения подсказок типа «грамматическая основа предложения», под которым понимается объединение полей *Подлежащее* и *Оригинальное представление сказуемого в тексте* САО-отношений. В данном случае также имеет место постадийная схема алгоритма:

– Стадия В1. Фильтрация описанных САО-отношений с неинформативными и несогласованными подлежащим и сказуемым;

– Стадия В2. Объединение наполнения полей *Подлежащее* и *Оригинальное представление сказуемого в тексте*;

– Стадия В3. Формирование подсказок с усеченным наполнением поля *Подлежащее*.

Разработан алгоритм автоматического построения подсказок типа «лексикон». К ним относятся однословные подсказки таких частей речи, как прилагательные, определительные местоимения, наречия, числительные, глаголы, предлоги. Для их получения все поля САО-отношений, прошедшие упомянутые ранее виды фильтрации, проходят через ряд однотипных стадий, целью которых является распознавание в этих полях слов с определенными лексико-грамматическими категориями, задаваемыми в виде тегов, приведение далее этих слов к начальной форме и добавление в базу подсказок.

Представленные метод решения задачи QTA, структурно-функциональная схема системы QTA/R, лингвистические ресурсы в виде различных словарей, базовых процедур, лингвистических паттернов и алгоритмов автоматического построения списков подсказок для автодополнения русскоязычных ПЗ, формулируемых с использованием базовых синтаксических структур, были внедрены в состав известной многоязычной информационно-поисковой платформы IHS Goldfire, что подтверждается соответствующим актом о внедрении. Разработанная система автодополнения русскоязычных ПЗ QTA/R является оригинальной. Ее отличительными чертами являются: предоставление пользователю возможности более точно формулировать свою информационную потребность, существенно минимизировать общее время решения поисковой задачи, решать задачу QTA, не ориентируясь на историю информационного поиска, получать гарантированно релевантную реакцию поисковой системы. Проведенное экспертное тестирование показало, что качество получаемых системой QTA/R решений с позиции таких классических для практики СИП и компьютерной лингвистики показателей, как точность и полнота, близко к их абсолютным значениям по отношению к БД САО-отношений и составляет 84–85 % по отношению к ПБД, из которой генерируется указанная БД. Ниже представлены фрагменты результатов, полученных системой QTA/R при автодополнении конкретных ПЗ (рисунки 2–4).

Рисунок 2. – Автоматическое дополнение запроса с набранной частью «химический род»

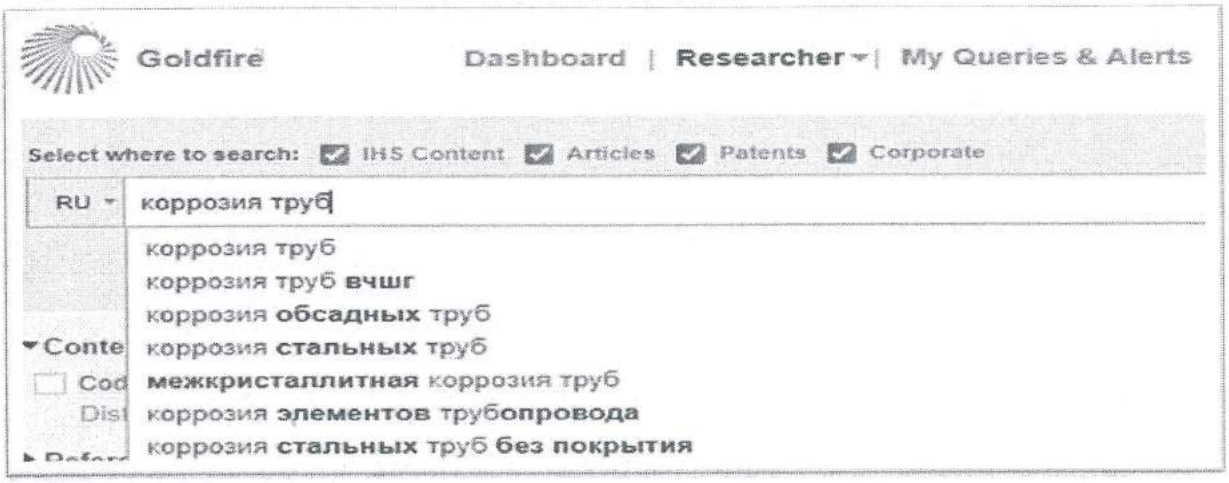


Рисунок 3. – Автоматическое дополнение запроса с набранной частью «коррозия труб»

Рисунок 4. – Автоматическое дополнение запроса с набранной частью «неотключения»

ЗАКЛЮЧЕНИЕ

Основные научные результаты диссертации

1. Сформулирована и обоснована концепция решения задачи QTA, ориентированная на промышленный характер приложения и высокие показатели эффективности; вопросно-ответную СИП с ЕЯ-интерфейсом пользователя и русскоязычную ПБД большого объема, являющуюся поисковым пространством СИП; использование «истории» не уже осуществленного, а предполагаемого поиска в виде множества подсказок, заранее автоматически распознаваемых в ТД из ПБД; не только на последовательное завершение уже набранной части ПЗ, а погружение ее в контекст, а именно дополнение в начале, в конце и даже внутри; глубокий, вплоть до семантического уровня ЕЯ, лингвистический анализ ТД [1–А; 2–А; 5–А; 7–А].

2. Построена классификация основных типов ПЗ: одно или несколько несогласованных ключевых слов; грамматически согласованные словосочетания; вопросительные предложения; утвердительные предложения; комбинация двух последних типов, нередко без знака препинания между ними и чаще без вопросительного знака в конце ПЗ. Дана классификация базовых синтаксических структур для основных типов ПЗ: именные группы (простые именные группы, расширенные именные группы с предложно-падежными зависимыми конструкциями), глагольные группы (инфинитив глагола с прямым дополнением, с косвенным дополнением с предлогом, с прямым и косвенным дополнениями, с прямым дополнением и причастным оборотом; форма 3 лица единственного или множественного числа глагола с прямым дополнением, с косвенным дополнением с предлогом, с прямым и косвенным дополнениями, с прямым дополнением и причастным оборотом), грамматическая основа предложения (предикативный центр) [1–А; 2–А].

3. Предложен новый метод решения задачи QTA, который, в отличие от существующих, основан на классификации основных типов ПЗ и их базовых синтаксических структур, сводится к автоматическому распознаванию, на основе глубокого лингвистического анализа текстовых документов, этих структур в ПБД и выбору их лексических наполнений, которые в совокупности составляют базу подсказок для автодополнения ПЗ [1–А; 2–А; 3–А; 4–А; 5–А; 7–А; 8–А].

4. Разработана структурно-функциональная схема системы QTA/R автодополнения русскоязычных ПЗ в СИП, функциональность которой включает базовый лингвистический анализ ПБД с целью построения лингвистического индекса ТД в виде БД САО-отношений и автоматическое

распознавание в ней подсказок для формирования базы подсказок. Указанная схема ориентирована на сформулированную концепцию и предложенный метод решения задачи QTA и основана на лингвистическом процессоре задачи, включающем наряду с доступным БЛП необходимое для решения целевой задачи его расширение [2–A; 5–A; 7–A].

5. Разработано в виде расширения функциональности БЛП собственное лингвистическое обеспечение и алгоритмы автоматического распознавания в ПБД подсказок, соответствующих базовым синтаксическим структурам основных типов ПЗ, с целью их автодополнения. Такого рода ресурсы впервые используются для решения задачи QTA [3–A; 4–A; 6–A; 8–A].

6. Разработан прототип оригинальной системы QTA/R. Осуществлено ее тестирование, которое показало, что качество получаемых решений по показателям точности и полноты близко к их абсолютным значениям по отношению к БД САО-отношений и составляет 84–85 % по отношению к ПБД. Осуществлено внедрение системы в промышленную эксплуатацию, которое показало, что она, в силу использования разработанных концептуальных и алгоритмических решений, а также лингвистических ресурсов, обладает наряду с уже отмеченными высокими показателями эффективности такими актуальными особенностями, как: предоставление пользователю возможности более точно формулировать свою информационную потребность, существенно минимизировать общее время решения поисковой задачи, решать задачу QTA, не ориентируясь на историю информационного поиска, получать гарантированно релевантную реакцию поисковой системы [2–A; 5–A].

Рекомендации по практическому использованию результатов

Разработанные метод, алгоритмы и лингвистические ресурсы рекомендуются к использованию при проектировании и реализации различных систем информационного поиска с ЕЯ-интерфейсом пользователя, а также в учебном процессе в высших учебных заведениях, осуществляющих подготовку специалистов в области интеллектуальных информационных систем и компьютерной лингвистики.

Построенная система QTA/R автодополнения русскоязычных ПЗ внедрена в СИП в состав известной многоязычной информационно-поисковой платформы IHS Goldfire, используемой для решения задач автоматизации инженерии и управления знаниями крупнейшими компаниями мира.

СПИСОК ПУБЛИКАЦИЙ СОИСКАТЕЛЯ УЧЕНОЙ СТЕПЕНИ

Статьи в рецензируемых научных изданиях

1–А. Голяк, Ю. Д. Автодополнение поискового запроса на основе автоматического извлечения подсказок из проиндексированных документов предметной области / Ю. Д. Голяк // Вес. БДПУ. Сер. 1, Педагогіка. Псіхалогія. Філалогія. – 2018. – № 3. – С. 91–95.

2–А. Голяк, Ю. Д. Принципиальная схема решения задачи автодополнения пользовательских поисковых запросов на русском языке и её анализ / Ю. Д. Голяк // Учен. зап. ВГУ им. П. М. Машерова. – 2020. – Т. 31. – С. 142–147.

3–А. Голяк, Ю. Д. Автоматическое дополнение русскоязычных пользовательских запросов, формулируемых в виде именных групп / Ю. Д. Голяк, И. В. Совпель // Вестн. МГЛУ. Сер. 1, Филология. – 2021. – № 3. – С. 101–109.

4–А. Голяк, Ю. Д. Автоматическое дополнение русскоязычных пользовательских запросов на основе подсказок типа «глагольная группа», «грамматическая основа предложения» и «лексикон» / Ю. Д. Голяк, И. В. Совпель // Вестн. МГЛУ. Сер. 1, Филология. – 2021. – № 4. – С. 84–93.

5–A. Haliak, J. The system for automatic suggestions generation for the purpose of autocompletion of russian-language user search queries / J. Haliak // Открытые семантические технологии проектирования интеллектуальных систем (OSTIS-2021) : сб. науч. тр. / Белорус. гос. ун-т информатики и радиоэлектроники. – Минск, 2021. – Вып. 5. – С. 289–292.

Статьи в сборниках научных статей

6–А. Совпель, И. В. Прикладная лингвистика в свете современных естественно-языковых приложений: общие принципы и особенности разработки базовых лингвистических процессоров на примере русского языка / И. В. Совпель, Ю. Д. Голяк // Карповские научные чтения : сб. науч. ст. по материалам XI Карповских научных чтений, Минск, 17–18 марта 2017 г. : Вып. 11. – в 2 ч. / Белорус. гос. ун-т ; редкол.: А. И. Головня (отв. ред.) [и др.]. – Минск, 2017. – Ч. 1. – С. 45–53.

7–А. Голяк, Ю. Д. Лингвистический процессор для автоматического дополнения пользовательских запросов на русском языке / Ю. Д. Голяк // Молодые ученые в инновационном поиске : сб. науч. ст. по материалам IX Междунар. науч. конф., Минск, 27–28 мая 2020 г. : в 2 ч. / Мин. гос. лингв. ун-т ; редкол.: Т. П. Карпилович (отв. ред.) [и др.]. – Минск, 2021. – Ч. 1. – С. 221–227.

Материалы научных и научно-практических конференций

8–А. Голяк, Ю. Д. Автоматическое дополнение пользовательского запроса типа «именная группа» / Ю. Д. Голяк // Лингвистика, лингводидактика, лингвокультурология: актуальные вопросы и перспективы развития : материалы V Междунар. науч.-практ. конф., Минск, 18–19 марта 2021 г. / Белорус. гос. ун-т ; редкол.: О. Г. Прохоренко (гл. ред.) [и др.]. – Минск, 2021. – С. 309–314.

РЭЗЮМЭ

Галяк Юлія Дзмітрыеўна

Прэдыктыўнае аўтадапаўненне запытаў карыстальнікаў у сістэмах інфармацыйнага пошуку (на матэрыяле рускай мовы)

Ключавыя словы: тэкставы дакумент, карыстальніцкі запыт, лінгвістычны працэсар, паўнатэкставая база дадзеных, сістэма інфармацыйнага пошуку, лінгвістычны аналіз, алгарытм, аўтадапаўненне

Мэта даследавання: распрацоўка метаду, алгарытмаў і лінгвістычнага забеспячэння для вырашэння задачы прэдыктыўнага аўтадапаўнення рускамоўных запытаў карыстальнікаў у сістэмах інфармацыйнага пошуку і іх рэалізацыя ў выглядзе прамысловага прататыпа.

Метады даследавання: апісальны метад, кантэкстуальны і супастаўляльны аналіз, статыстычны метад, метад кампанентнага аналізу, алгарытмізацыя лінгвістычнага аналізу тэксту, экспертны кантроль якасці.

Атрыманыя вынікі і іх навізна. Сфармуляваная і абгрунтаваная канцэпцыя вырашэння задачы QTA, арыентаваная, у адрозненне ад існых, на пытанне-адказную сістэму інфармацыйнага пошуку з натуральна-моўным інтэрфейсам карыстальніка і рускамоўную паўнатэкставую базу дадзеных, выкарыстанне яе ў якасці гісторыі прадугледжанага пошуку, змяшчэнне запыту, які ўводзіць карыстальнік, у шырокі кантэкст; прапанаваны метад вырашэння задачы QTA, які ўпершыню быў заснаваны на класіфікацыі асноўных тыпаў запытаў карыстальніка і іх базавых сінтаксічных структур; распрацаваная структурна-функцыянальная схема сістэмы QTA/R аўтадапаўнення рускамоўных запытаў карыстальніка ў сістэмах інфармацыйнага пошуку, яе лінгвістычнае забеспячэнне і алгарытмы аўтаматычнага распазнання ў паўнатэкставай базе дадзеных падказак, якія адпавядаюць базавым сінтаксічным структурам асноўных тыпаў запытаў карыстальніка, з мэтай іх аўтадапаўнення; распрацаваны прататып арыгінальнай сістэмы QTA/R з высокімі паказчыкамі эфектыўнасці, здейснена яго ўбудаванне ў прамысловую сістэму.

Рэкамендацыі па выкарыстанні і галіны ўжывання. Распрацаваныя метад, алгарытмы і лінгвістычныя рэсурсы рэкамендуюцца да выкарыстання пры пабудове інтэлектуальных інфармацыйных сістэм рознага прызначэння, а таксама ў навучальным працэсе ў вышэйшых навучальных установах, якія ажыццяўляюць падрыхтоўку спецыялістаў у галіне інфарматыкі і камп'ютарнай лінгвістыкі.

Атрыманыя вынікі былі ўбудаваныя ў склад вядомай шматмоўнай інфармацыйна-пошукавай платформы IHS Goldfire, якая выкарыстоўваецца для вырашэння інавацыйных задач буйнымі кампаніямі свету.

РЕЗЮМЕ

Голяк Юлия Дмитриевна

Предиктивное автодополнение запросов пользователей в системах информационного поиска (на материале русского языка)

Ключевые слова: текстовый документ, пользовательский запрос, лингвистический процессор, полнотекстовая база данных, система информационного поиска, лингвистический анализ, алгоритм, автодополнение

Цель исследования: разработка метода, алгоритмов и лингвистического обеспечения решения задачи предиктивного автодополнения русскоязычных запросов пользователей в системах информационного поиска и их реализация в виде промышленного прототипа.

Методы исследования: описательный метод, контекстуальный и сопоставительный анализ, метод компонентного анализа, статистический метод, алгоритмизация лингвистического анализа текста, экспертное тестирование.

Полученные результаты и их новизна. Сформулирована и обоснована концепция решения задачи QTA, которая, в отличие от существующих, ориентирована на вопросно-ответную СИП с ЕЯ-интерфейсом пользователя и русскоязычную ПБД, использование ее в качестве истории предполагаемого поиска, погружение вводимого пользователем запроса в широкий контекст; предложен метод решения задачи QTA, который впервые основан на классификации основных типов ПЗ и их базовых синтаксических структур; разработана структурно-функциональная схема системы QTA/R автодополнения русскоязычных ПЗ в СИП, ее лингвистическое обеспечение и алгоритмы автоматического распознавания в ПБД подсказок, соответствующих базовым синтаксическим структурам основных типов ПЗ, с целью их автодополнения; разработан прототип оригинальной системы QTA/R с высокими показателями эффективности, осуществлено его внедрение.

Рекомендации по использованию и область применения. Разработанные метод, алгоритмы и лингвистические ресурсы рекомендуются к использованию припостроении интеллектуальных информационных систем различного назначения, а также в учебном процессе в высших учебных заведениях, осуществляющих подготовку специалистов в области информатики и компьютерной лингвистики.

Полученные результаты внедрены в состав известной многоязычной информационно-поисковой платформы IHS Goldfire, используемой для решения инновационных задач крупнейшими компаниями мира.

SUMMARY

Haliak Julia Dmitrievna

Predictive autocompletion of user queries in information search systems (based on the Russian language)

Key words: text document, user query, linguistic processor, full-text database, information search system, linguistic analysis, algorithm, autocompletion

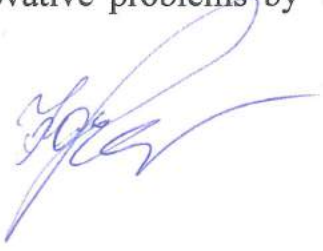
The goal of research: development of the method, algorithms and linguistic support for solving the problem of predictive autocompletion of Russian-language user queries in information search systems and their implementation in the form of the industrial prototype.

The methods of research: descriptive method, contextual and comparative analysis, statistical method, componential analysis, algorithmization of linguistic text analysis, expert testing.

The scientific novelty of the obtained results. The concept of solving the QTA problem is formulated and substantiated in the way that, unlike the existing solutions, is focused on the question-answering systems of informational search with the NL-user interface and the Russian-language full-text databases, using it as history of the intended search, immersing the query entered by the user in a wide context; the suggested method for solving the QTA problem is for the first time based on the classification of the main types of user queries and their basic syntactic structures; the structural and functional schema of the QTA/R system for autocompletion of Russian-language user queries in systems of informational search, its linguistic support and algorithms for automatic autocompletion recognition in full-text databases corresponding to the basic syntactic structures of the main types of user queries, with the aim of their autocompletion, have been developed; the prototype of the original QTA/R system with high efficiency was developed and implemented.

Recommendations on the usage of the results and the sphere of application. The developed method, algorithms and linguistic resources are recommended for use in the development of intelligent information systems for various purposes, as well as in the educational process in higher educational institutions for future specialists in the field of computer science and computational linguistics.

The obtained results are incorporated into the multilingual information retrieval platform IHS Goldfire, which is used to solve innovative problems by large world companies.



Научное издание

Голяк Юлия Дмитриевна

ПРЕДИКТИВНОЕ АВТОДОПОЛНЕНИЕ ЗАПРОСОВ ПОЛЬЗОВАТЕЛЕЙ В СИСТЕМАХ ИНФОРМАЦИОННОГО ПОИСКА
(на материале русского языка)

Автореферат
диссертации на соискание ученой степени
кандидата филологических наук
по специальности 10.02.21 – прикладная и математическая лингвистика

Ответственный за выпуск *Ю. Д. Голяк*

Подписано в печать 07.02.2023. Формат 60×84 $^{1}/_{16}$. Бумага офсетная. Гарнитура Таймс. Ризография. Усл. печ. л. 1,45. Уч.-изд. л. 1,30. Тираж 100 экз. Заказ 6.
Издатель и полиграфическое исполнение: учреждение образования «Минский государственный лингвистический университет». Свидетельство о государственной регистрации издателя, изготовителя, распространителя печатных изданий от 02.06.2017 г. № 3/1499. ЛП № 02330/458 от 10.07.2020 г.
Адрес: ул. Захарова, 21, 220034, г. Минск.