БЕЛОРУССКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ

УДК 002:004+81’32

КРАПИВИН
Юрий Борисович

МЕТОДЫ И АЛГОРИТМЫ АВТОМАТИЧЕСКОГО РАСПОЗНАВАНИЯ ВОСПРОИЗВЕДЕННЫХ ФРАГМЕНТОВ ТЕКСТОВЫХ ДОКУМЕНТОВ

Автореферат
диссертации на соискание ученой степени
кандидата технических наук
по специальности
05.13.17 – Теоретические основы информатики

Минск – 2019

Работа выполнена
в УО «Брестский государственный технический университет».

Научный руководитель – **Головко Владимир Адамович,**
доктор технических наук, профессор,
заведующий кафедрой
интеллектуальных информационных технологий
УО «Брестский государственный
технический университет».

Официальные оппоненты: **Липницкий Станислав Феликсович**,
доктор технических наук,
главный научный сотрудник отдела совместных программ космических и информационных технологий ГНУ «Объединенный институт
проблем информатики НАН Беларуси»;

Шибут Марина Станиславовна,
кандидат технических наук, доцент,
доцент кафедры управления информационными ресурсами Академии управления при Президенте Республики Беларусь.

Оппонирующая организация – **УО «Белорусский государственный университет информатики и радиоэлектроники».**

Защита состоится 22 марта 2019 г. в 10.00 на заседании совета по защите диссертаций Д 02.01.02 при Белорусском государственном университете по адресу: г. Минск, ул. Ленинградская 8 (корпус юридического факультета), ауд. 407. Телефон учёного секретаря: 209-57-09.

Почтовый адрес: пр-т Независимости 4, Минск, 220030.

С диссертацией можно ознакомиться в Фундаментальной библиотеке Белорусского государственного университета.

Автореферат разослан « 11 » февраля 2019 г.

Ученый секретарь совета
по защите диссертаций
кандидат физ.-мат. наук доцент Е.С. Чеб

ВВЕДЕНИЕ

Одной из наиболее характерных черт настоящего времени являются интенсивное развитие и проникновение информационных технологий практически во все сферы жизнедеятельности человека. Существующие информационные системы, оперирующие многочисленными информационными ресурсами, представленными как в собственных базах данных (БД) пользователей, так и в сети Интернет, обеспечивают быстрый поиск и различного рода обработку интересующей их информации. И, естественно, подготовка практически любой квалификационной или исследовательской работы, начиная от школьного реферата и заканчивая диссертацией, так или иначе, основывается на этих возможностях. В связи с этим очень актуальной является проблема автоматического обнаружения в текстовых документах заимствованных фрагментов (ЗФ), т. е. тех их фрагментов, которые заимствованы из других источников, каковыми всё чаще становятся именно Интернет-доступные документы, и последующего анализа этих ЗФ на предмет соблюдения норм цитирования и, возможно, наличия плагиата. При этом особенно трудоёмкой является первая задача, которая, прежде всего, требует разработки эффективных средств автоматизации её решения.

Существующие в настоящее время системы автоматического распознавания ЗФ оперируют алгоритмами только явного, но не всегда точного заимствования фрагментов текста: их соответствия по лексическому составу с учётом простейших морфологических преобразований и отношений синонимии. К тому же, эти решения ориентированы на одноязычные поисковые пространства. Наблюдаемая в последние годы тенденция усиления лингвистической составляющей в системах информационного поиска с целью повышения качественных показателей их работы тем более должна иметь место в системах автоматического распознавания ЗФ, особенно, если речь идёт о неявно, т. е. семантически заимствованных фрагментах и о решении задачи на множестве документов, представленных в поисковом пространстве на различных естественных языках (ЕЯ) (так называемая cross-language функциональность). Должна быть разработана концепция решения задачи, универсальная по отношению к различным ЕЯ, ориентированная на актуальные типы поискового пространства, а также на различные уровни глубины лингвистического анализа текста. Безусловно, такая постановка обостряет проблему эффективности как лингвистического, так и алгоритмического обеспечения задачи автоматического распознавания ЗФ, особенно на этапах распознавания языка текстовых документов, поиска релевантных им текстовых документов в Интернет-доступных источниках, машинного перевода (МП) текстовых документов и их поисковых образов, собственно распознавания лексически и семантически заимствованных фраг-

ментов текстовых документов.

Системы рассматриваемого типа в Республике Беларусь не разрабатывались и не использовались. Тем более, не существует решения задачи в белорусско-русской языковой среде, которое, учитывая что оба указанных языка являются в нашей стране государственными, является очень актуальным. С этой точки зрения предлагаемая в работе модель системы автоматического распознавания заимствованных фрагментов, её лингвистическое, алгоритмическое и программное обеспечение должны соответствовать промышленному характеру решения задачи в белорусско-русской информационной среде.

Именно эти, перечисленные выше, задачи и их решения рассматриваются в данной диссертационной работе.

ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

Связь работы с научными программами (проектами), темами

Тема диссертации соответствует приоритетным направлениям научных исследований в Республике Беларусь на 2011–2015 гг., утвержденным Постановлением Совета Министров Республики Беларусь от 12 августа 2010 г. № 1196, пункт 5.1 «Методы математического и компьютерного моделирования, компьютерные технологии и интеллектуальные системы поддержки принятия решений».

Диссертационное исследование выполнялось на кафедре интеллектуальных информационных технологий факультета электронно-информационных систем учреждения образования «Брестский государственный технический университет» (БрГТУ) в соответствии с: 1) договором на выполнение НИР от 1 марта 2008 г. № 204/35 между Высшей аттестационной комиссией Республики Беларусь и Белорусским государственным университетом: задание «Разработать технологию и инструментально-программный комплекс распознавания в диссертационных работах случаев заимствования без ссылок на авторов», шифр – ИПК «ПлагиатКонтроль», № госрегистрации 20080836; 2) гранта Министерства образования Республики Беларусь № ГБ11/118 «Разработать инструментально-программный комплекс распознавания в дипломных работах случаев заимствования без ссылок на авторов», № госрегистрации 20111348; 3) НИР кафедры интеллектуальных информационных технологий «Модели и алгоритмы обработки информации в интеллектуальных системах: разработка средств автоматизации обучения моделированию с использованием стохастических сетей», № госрегистрации 20131891; 4) государственной программой научных исследований на 2016–2018 гг. «Информатика, космос и безопасность», подпрограмма («Информатика и космические исследования»), задание

1.6.05 «Методы и алгоритмы интеллектуальной обработки и анализа большого объема данных на основе нейронных сетей глубокого доверия», № госрегистрации 20163595.

Цель и задачи исследования

Целью диссертационной работы является разработка принципов, методов, алгоритмов и программных средств решения задачи автоматического распознавания заимствованных фрагментов текстовых документов в многоязычной информационной среде.

Для достижения поставленной цели необходимо решить следующие основные задачи:

1. Разработать концепцию системы автоматического распознавания заимствованных фрагментов текстовых документов, универсальную по отношению к различным ЕЯ, ориентированную на актуальные типы поискового пространства и его многоязычность, а также на различные уровни глубины лингвистического анализа текста.
2. Разработать универсальный по отношению к различным ЕЯ алгоритм распознавания языка текстовых документов.
3. Разработать стратегию использования необходимой функциональности машинного перевода текстов и алгоритм МП в строящихся автоматически поисковых образах текстовых документов.
4. Разработать алгоритмы автоматического распознавания лексически и семантически заимствованных фрагментов текстовых документов.
5. Разработать промышленную систему автоматического распознавания заимствованных фрагментов текстовых документов в белорусско-русской языковой среде, включая её лингвистическое, алгоритмическое и программное обеспечение.

Объектом исследования являются текстовые документы.

Предметом исследования являются методы, алгоритмы и модели систем автоматического распознавания заимствованных фрагментов текстовых документов в одно- и многоязычной среде.

Научная новизна

1. Разработана и обоснована структурно-функциональная схема оригинальной системы автоматического распознавания заимствованных текстовых фрагментов, ориентированной, в общем случае, на поисковое пространство, включающее как текстовые Интернет-документы, так и документы из БД пользователя. Новизна заключается в ориентации системы на распознавание не

только лексически, но и семантически заимствованных фрагментов, на многоязычность информационной среды и cross-language функциональность.

2. Построен эффективный алгоритм автоматического распознавания языка текстовых документов. Новизна заключается в использовании в его основе комбинации предложенных алфавитного метода, метода грамматических слов и алфавитно-триграммного метода, что в совокупности привело к существенному сокращению объёма требуемых языковых данных, обеспечило один из лучших показателей точности решения задачи (99,8 %), причём, с учётом текстов типа одной короткой фразы (7 слов). Последнее особенно важно с точки зрения использования алгоритма для ЕЯ запросов пользователя в вопросно-ответных системах.

3. Исходя из целевой задачи, определена стратегия применения для её решения функциональности машинного перевода и построен общий алгоритм МП строящихся автоматически поисковых образов текстовых документов. Новизна заключается в учёте здесь характеристик поискового пространства и особенностей поисковой службы Google, а также в использовании в основе алгоритма бинарных словарей концептов, акций и их атрибутов. Указанные компоненты являются универсальными по отношению к различным ЕЯ и обеспечивают эффективные решения как задачи МП, так и задачи автоматического распознавания семантически заимствованных текстовых фрагментов. Показано, что поисковые образы текстовых документов могут быть качественно скорректированы с использованием отношений синонимии и поправок на весовые коэффициенты, учитывающих принадлежность лексических единиц к наиболее информативным лексико-грамматическим, синтаксическим и семантическим классам.

4. Построен эффективный алгоритм решения задачи автоматического распознавания лексически заимствованных фрагментов текстовых документов. Новизна заключается в использовании в его основе понятия лексически равных, с точностью до канонических форм слов и отношений синонимии, предложений и процедур построения обратных индексов. Показано, что начисляемые при этом накапливаемые веса для сравниваемых предложений и вводимая эвристика, учитывающие статистические особенности ЕЯ, существенно оптимизируют трудоёмкость алгоритма. Построен общий алгоритм решения оригинальной задачи автоматического распознавания семантически заимствованных текстовых фрагментов. В его основе используется система знаний, ориентированная на распознавание в текстовых документах объектов и фактов, а также их атрибутов.

5. Разработано лингвистическое, алгоритмическое и программное обеспечение системы автоматического распознавания лексически заимствованных фрагментов текстовых документов, обеспечившей впервые решение задачи в

белорусско-русской языковой среде с функциональностью cross-language. Осуществлено внедрение системы, которая в силу использования полученных концептуальных, алгоритмических и технологических решений, обладает высокими качественными и техническими характеристиками.

Положения, выносимые на защиту

1. Структурно-функциональная схема оригинальной системы автоматического распознавания лексически и семантически заимствованных фрагментов, ориентированной, в общем случае, на поисковое пространство, включающее как текстовые Интернет-документы, так и документы из базы данных пользователя, на многоязычность информационной среды и cross-language функциональность.

2. Алгоритм автоматического распознавания языка текстовых документов, основанный на комбинировании предложенных алфавитного метода, метода грамматических слов и алфавитно-триграммного метода и, как следствие, требующий нетрудоёмкого статистического и лингвистического анализа языковых данных.

3. Стратегия применения функциональности машинного перевода, и общий алгоритм машинного перевода строящихся автоматически поисковых образов текстовых документов, основанный на бинарных словарях концептов, акций и их атрибутов, представленных в базе данных, аналогичной известной лексической базе данных WordNet, которые, в совокупности, обеспечивают решение как задачи машинного перевода, так и задачи автоматического распознавания семантически заимствованных фрагментов текстовых документов.

4. Алгоритм решения задачи автоматического распознавания лексически заимствованных фрагментов текстовых документов, в основу которого положены понятие лексически равных с точностью до канонических форм слов и отношений синонимии предложений, и процедуры построения обратных индексов, а также общий алгоритм решения задачи автоматического распознавания семантически заимствованных фрагментов текстовых документов, основанный на системе знаний, ориентированной на распознавание в текстовых документах объектов и фактов и их атрибутов.

5. Промышленная система «ПлагиатКонтроль», включая её лингвистическое и универсальное по отношению к различным естественным языкам алгоритмическое и программное обеспечение, которая впервые обеспечила решение задачи автоматического распознавания лексически заимствованных фрагментов текстовых документов в белорусско-русской языковой среде с функциональностью cross-language.

Личный вклад соискателя ученой степени

Все результаты и положения, выносимые на защиту, получены автором самостоятельно. Научный руководитель принимал участие в выборе направления исследований, обсуждении теоретических и практических результатов, полученных автором.

Апробация диссертации и информация об использовании ее результатов

Основные результаты диссертационной работы докладывались и обсуждались на: Международных научно-технических конференциях «Искусственный интеллект. Интеллектуальные системы» (п. Кацивели, АР Крым, Украина, 2008 г., 2012 г., 2013 г.), 13-м, 14-м, 15-м Международных симпозиумах аспирантов, докторантов и молодых учёных Западной, Центральной и Восточной Европы – International PhD Workshop OWD (Висла, Польша, 2011 г., 2012 г, 2013 г.), Международной научной конференции «Молодые учёные в инновационном поиске» (Минск, 2012 г.), 10-й Европейской конференции молодых исследователей и учёных – «TRANSCOM 2013» (Жилина, Словакия, 2013 г.), Международной научной конференции «Информационные технологии и системы 2013» (Минск, 2013 г.), 4-й Международной научно-практической интернет-конференции «Инновационные технологии обучения физико-математическим дисциплинам» (Мозырь, 2014 г.), Международном конгрессе по информатике Информационные системы и технологии CSIST'2016 (Минск, 2016 г.), III Международной научной конференции «Библиотеки в информационном обществе: сохранение традиций и развитие новых технологий». Тема 2018 года – «Научная библиотека как центр культурно-информационного пространства» (Минск, 2018 г.).

Результаты диссертации внедрены в ряде университетов Республики Беларусь и в ВАК Беларуси, имеется 4 акта о внедрении.

Опубликованность результатов диссертации

Основные результаты диссертации опубликованы в 18 научных работах, из которых: 6 статей в научных изданиях в соответствии с п.18 Положения о присуждении учёных степеней и присвоении учёных званий в Республике Беларусь (общим объемом 3,1 авторского листа), 12 статей в сборниках материалов научных конференций.

Структура и объем диссертации

Диссертационная работа состоит из перечня сокращений и условных обозначений, введения, общей характеристики работы, четырёх глав, заключения, библиографического списка и приложения. Полный объём диссертации составляет 129 страниц, в том числе 19 рисунков занимают 9 страниц, 10 таблиц на 5 страницах, 1 приложение занимает 5 страниц. Библиографический список содержит 229 наименований, включая собственные публикации соискателя ученой степени.

ОСНОВНАЯ ЧАСТЬ

В первой главе исследуется задача информационного поиска и, в частности, поиска релевантных документов, т. е. тех, которые схожи по содержанию с заданным документом-образцом, который и выступает в качестве запроса пользователя. Отмечается, что наблюдаемая в настоящее время тенденция усиления лингвистической составляющей в системах подобного типа с целью повышения качественных показателей их работы, а также необходимость осуществлять поиск во множестве документов, представленных на различных ЕЯ, а не только на языке документа-запроса (так называемая cross-language функциональность) требует в свою очередь решения многих «лингвистически нагруженных» задач, а это приводит к необходимости разработки достаточно развитых лингвистических процессоров.

Отмечается, что актуальным частным случаем рассматриваемой задачи является задача автоматического распознавания заимствованных текстовых фрагментов (плагиата), который может быть явным (лексическим), когда речь идет об одном и том же фрагменте текста, принадлежащем разным текстовым документам (могут допускаться минимальные расхождения, например, за счёт использования вводных слов, синонимов и т. п.), и неявным (семантическим), когда речь идёт о фрагментах различных текстовых документов, имеющих одинаковый, по отношению к заданной системе знаний, но выраженный разными цепочками символов, смысл.

Показано, что существующие решения задачи автоматического распознавания заимствованных фрагментов (ЗФ) получены в рамках следующих наиболее распространенных подходов: строкового соответствия, атрибутно-подсчётного и информационно-поискового. Они оперируют алгоритмами только явного, но не всегда точного заимствования фрагментов текста: их соответствия по лексическому составу и позициям лексических единиц с учётом простейших морфологических преобразований и отношений синонимии. К тому же, эти решения ориентированы на одноязычное поисковое пространство. Рас-

ширение поискового пространства (от полнотекстовых БД пользователя до Интернет-доступных источников) и его многоязычность обостряют проблему эффективности как лингвистического, так и алгоритмического обеспечения задачи автоматического распознавания ЗФ, особенно на этапах распознавания языка текстовых документов, поиска релевантных им текстовых документов в Интернет-доступных источниках, собственно распознавания лексически и семантически заимствованных фрагментов текстовых документов. Отмечается также, что не существует решения указанной задачи для текстовых документов на белорусском языке, которое, безусловно, является актуальным, а учитывая, что в Республике Беларусь два государственных языка, то оно тем более актуально именно в белорусско-русской информационной среде.

В заключение даётся общая постановка задачи и сформулированы основные её подзадачи.

Во второй главе рассмотрен круг вопросов, касающихся принципиальной схемы решения задачи автоматического распознавания ЗФ (рисунок 1) и функциональности необходимого для достижения этой цели лингвистического процессора.

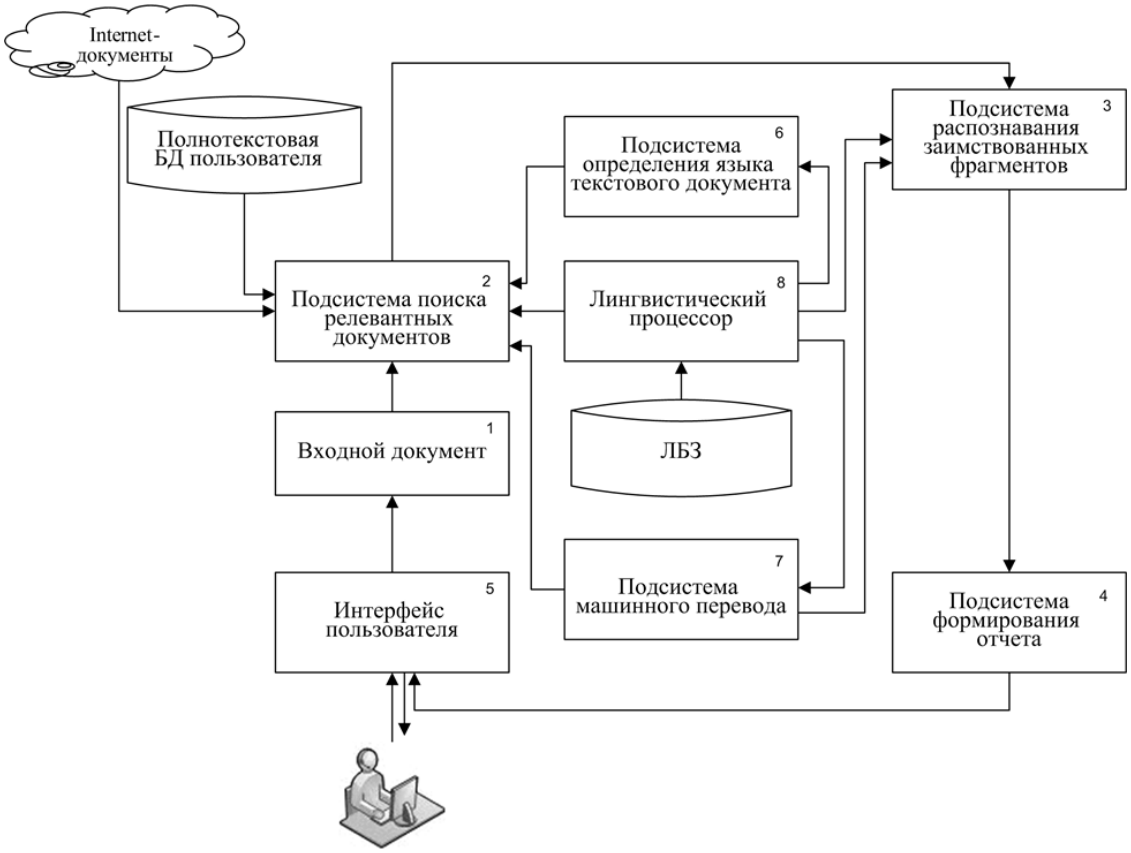

Рисунок 1. – Структурно-функциональная схема системы автоматического распознавания ЗФ

Так, разработана и обоснована структурно-функциональная схема системы автоматического распознавания ЗФ, ориентированная, в общем случае, на по-

исковое пространство, включающее как текстовые Интернет-документы, так и документы из БД пользователя, на многоязычность информационной среды и cross-language функциональность. Показано, что в соответствии с приведенной схемой задача автоматического распознавания ЗФ в целом включает в совокупности следующие задачи: 1) поиска релевантных входному текстовых документов в полнотекстовой БД и в сети Интернет; 2) автоматического распознавания языка текстового документа; 3) машинного перевода текстовых документов и их поисковых образов во множестве заданных ЕЯ; 4) автоматического распознавания лексически и семантически заимствованных фрагментов текстовых документов; 5) автоматического построения отчета; 6) автоматического лингвистического анализа текстовых документов.

Автоматический лингвистический анализ текстовых документов, распознавание их языка и их машинный перевод составляют в совокупности базовую лингвистическую обработку текстовых документов при решении общей целевой задачи. При этом задачи 2), 4) и 5) в полном объёме требуют своего эффективного решения, задача 1) – для случая Интернет-документов, причём с учётом особенностей механизмов поиска, реализованных в современных поисковых машинах, задача 3) – для МП поисковых образов текстовых документов в случае работы системы в режиме cross-language. Сформулированы и обоснованы требования, предъявляемые к результатам автоматического лингвистического анализа текстовых документов (задача 6)) в той степени, в какой это необходимо для решения задач 1), 2) и 4).Показано, что этим требованиям удовлетворяет функциональность известного многоязычного базового лингвистического процессора системы GoldFire.

Что касается задачи автоматического распознавания языка текстового документа, то в результате целого ряда экспериментов показано, что эффективный алгоритм её решения, сочетающий в себе возможности нетрудоёмкого статистического и лингвистического анализа языковых данных, может быть основан на комбинировании алфавитного метода, метода грамматических слов и алфавитно-триграммного метода. Таким образом, имеет место следующая принципиальная схема алгоритма:

1. Начало.

2. Обработать текст T алфавитным методом для всех ЕЯ из их заданного множества. Если на выходе получен единственный язык L_i , то обозначить его как решение задачи и перейти к п. 5.

3. Зафиксировать для T языковую группу и обработать его методом грамматических слов для тех ЕЯ, которые входят в эту группу. Если на выходе получен единственный язык L_i, то обозначить его как решение задачи и перейти к п. 5.

4. Зафиксировать для T образовавшуюся на шаге 3 новую языковую

группу и обработать его алфавитно-триграммным методом для тех ЕЯ, которые входят в эту группу. Обозначить L_i в качестве решения задачи.

5. Конец.

Если L_i на шаге 4 окажется не единственным, то необходим некоторый дополнительный ресурс снятия многозначности, хотя бы, например, диалог с пользователем. Реализация и тестирование на множестве ЕЯ (белорусский, русский, английский, французский, немецкий) показали высокую точность построенного алгоритма (99,8 %) и его эффективность при обработке даже коротких фраз (длиной в 7 слов, что значительно меньше среднестатистического по длине предложения: например, для известного корпуса текстов LOB Corpus значение этого параметра равняется примерно 19 слов).

Исследована функциональность машинного перевода в целях решения общей задачи в многоязычной информационной среде и режиме cross-language. Показано, что для МП самих текстовых документов могут использоваться уже существующие практические решения, а для необходимого МП их поисковых образов обоснованно предложено решение, основанное на автоматически получаемых двуязычных словарях концептов, акций и атрибутов, что обеспечивает эффективное решение задачи автоматического распознавания ЗФ даже на семантическом уровне. Действительно, указанные компоненты словарей, с одной стороны, являются уникальными семантическими понятиями, которые, в принципе, от языка не зависят и, таким образом, выступают в роли интерлингвы, а с другой – они являются «носителями» ключевых слов. Для хранения таких словарей предложена эффективная структура, аналогичная по своей структуре известной лексической БД WordNet, которая описывает концепты внешнего мира в форме пронумерованных понятий (синсетов), выраженных набором синонимичных слов и словосочетаний на всех языках из множества L, а также различные семантические отношения между концептами («общее–частное», «часть–целое», «группа–элемент» и т. д.). При этом сохраняются, очевидно, существовавшие бинарные отношения между лексическими единицами и достигается их минимальная избыточность. Идея использования такой структуры для трансляции ПОЗа входного текстового документа на языке L_i во множество ПОЗов на языках из $\overline{L}$ значительно упрощает эту процедуру, сводя её к поиску для каждого ключевого слова L_i – ПОЗа его синсета и последующему выбору из него представленных там его синонимов, а также эквивалентов на языках из $\overline{L}$.

В третьей главе рассмотрены вопросы, касающиеся уточнения изложенной во второй главе принципиальной схемы решения целевой задачи, решения задачи поиска текстовых документов, релевантных данному, в Интернет-доступных источниках и решения целевой задачи с использованием различных уровней глубины лингвистического анализа текста. Так, уточнение принципиальной схемы решения задачи автоматического распознавания ЗФ касается

только подсистемы поиска релевантных документов, а в рамках её – стратегии применения функциональности машинного перевода. Обосновано, что в общем случае наиболее приемлемым является использование указанной функциональности по отношению к ПОЗу, а не к самому запросу. В этом случае во входном документе сначала распознаются ключевые слова и строится его ПОЗ (одна процедура индексирования). Далее, на основании многоязычной лексической БД осуществляется перевод полученного ПОЗа на оставшиеся $(n-1)$ языков (всего $(n-1)$ процедур МП ПОЗа) и затем – поиск релевантных документов и формирование минимизированного поискового пространства (МПП), с целью быстрого и эффективного, с точки зрения релевантности входящих в ПП документов данному, отбора документов для последующего анализа на предмет наличия заимствований из них.

Обосновано, что при решении задачи с использованием Интернет-доступных источников целесообразно ориентироваться на поисковую службу Google. Определены те её особенности и накладываемые ограничения (количество свободно обрабатываемых запросов, количество результатов в поисковой выдаче, длина ПОЗа), которые являются наиболее существенными с точки зрения решаемой задачи. Таким образом, именно процедура поиска релевантных документов является механизмом минимизации ПП. При этом в основу автоматического построения ПОЗа, подаваемого на вход поисковой службы, обоснованно предложен известный $TF*IDF$ метод. Показано, что получаемый этим методом список ключевых слов может быть качественно скорректирован использованием отношений синонимии и поправок на их весовые коэффициенты, учитывающих принадлежность лексических единиц к наиболее информативным лексико-грамматическим, синтаксическим и семантическим классам.

Что касается решения целевой задачи с использованием различных уровней глубины лингвистического анализа текста, то дана постановка задачи и построен эффективный алгоритм её решения на лексико-грамматическом уровне ЕЯ, в основу которого положены понятие лексически равных предложений (два предложения лексически равны, если равны соответствующие им множества словоупотреблений) с точностью до канонических форм слов и отношений синонимии предложений, процедуры построения обратных индексов I_T и I_{DB} соответственно для входного текста T и текстов из поисковой БД, а также построения обратного индекса $I_C = I_T \cap I_{DB}$ и его сортировки:

1. Начало.

2. Построение обратного индекса I_T входного текста T: выбор из T множества всех попарно различных канонических слов, т. е. построение словаря W_T канонических слов входного текста, с указанием для каждого слова $w_i \in W_T$ множества N_i всех номеров тех предложений из T, в которых это слово содержится:

$$I_T = \{w_i, N_i\},\ i = \overline{1, |W_T|}.$$

3. Построение обратного индекса I_{DB} БД текстов: создание словаря W_{DB} канонических слов корпуса текстов, включающего все тексты БД, с указанием для каждого канонического слова $w_i \in W_{DB}$ множеств $N_i^{(1)}, N_i^{(2)}, ..., N_i^{(k_j)}$ всех номеров тех предложений каждого текста $T_i^{(m)}, m = \overline{1, k_i}$, из БД, в которых это слово содержится:

$$I_{DB} = \{w_j; N_j^{(1)}, N_j^{(2)}, ..., N_j^{(k_j)}\},\ j = \overline{1, |W_{DB}|}.$$

4. Построение обратного индекса I_C путём пересечения обратных индексов I_T и I_{DB} относительно их лексических компонентов с целью получения списка W слов, с точностью до синонимии, общих для I_T и I_{DB}, с сохранением для каждого $w_s \in W$ его веса p_s, равного количеству предложений из БД, в которые входит данное и синонимичные ему слова:

$$I_C = \{w_s; N_s; N_s^{(1)}, N_s^{(2)}, ..., N_s^{(k_s)}; p_s\},$$

где $p_s = \sum \left| N_s^{(m)} \right|,\ s = \overline{1, |W|}$.

5. Сортировка обратного индекса I_C в порядке возрастания весов входящих в него слов.

6. Распознавание во входном текстовом документе T предложений, заимствованных из текстовых документов БД.

6.1 Пошаговый выбор из списка I_C очередного слова w_s и его поиск (фиксирование) в каждом предложении текста T, определяемом по номеру из множества N_s; начисление предложению накапливаемых веса p', равного количеству таких слов в нём, и множества весов, каждый из которых, обозначим его p'', равен количеству всех слов данного предложения, а также им синонимичных слов, входящих в одно и то же предложение БД, определяемое одинаковым значением его номера из множеств $N_s^{(k_s)}$; сохранение только тех весов p'' и соответствующих номеров из множеств $N_s^{(k_s)}$, для которых, начиная с $p' > \mu,\ p' - p'' \le \mu$.

6.2 Как только $p'' = l - \mu$, то данное предложение из T является заимствованным из соответствующего текстового документа БД.

7. Конец.

Показано, что начисляемые при построении обратного индекса I_C для сравниваемых предложений накапливаемые веса, сортировка I_C и использование вводимой эвристики μ (максимально допустимого количества слов предложения из T, не входящих в сравниваемое предложение из БД), учитывающие

статистические особенности ЕЯ, существенно оптимизируют решение задачи. Теоретически трудоёмкость предложенного алгоритма в худшем случае (при $\mu = 0$) составляет $O(|DB|)$, где $|DB|$ – количество слов в БД текстов.

Предложен метод решения задачи на семантическом уровне ЕЯ. В его основу положена система знаний, ориентированная на распознавание в текстовых документах объектов и фактов, а также их атрибутов. Принципиальная схема алгоритма в этом случае аналогична той, которая разработана для лексико-грамматического уровня ЕЯ, только здесь текстовый документ рассматривается не как цепочка слов, а как цепочка фактов.

В четвертой главе представлены наиболее важные результаты применения полученных теоретических решений при разработке инструментально-программного комплекса (ИПК) «ПлагиатКонтроль», который ориентирован на обработку текстовых документов, прежде всего, диссертационных и дипломных работ, представленных как на русском, так и, впервые, на белорусском языках, с целью автоматического распознавания в них ЗФ, причём, также впервые в практике таких систем в режиме cross-language. Функционируя в поисковом пространстве, содержащем как текстовые Интернет-документы, так и документы из поисковой БД пользователя, он обеспечивает решение задачи на лексико-грамматическом уровне в полном соответствии с предложенной структурно-функциональной схемой системы автоматического распознавания ЗФ и её уточнением.

Что касается входящих в состав ИПК подсистем, то его подсистема «Поиска релевантных документов» использует поисковый механизм Google, распознавание ключевых слов основано на методе TF*IDF, наиболее приемлемая длина ПОЗа составляет 3–5 ключевых слов, а количество возвращаемых поисковой машиной релевантных документов – не более 50. В основе подсистем «Определения языка текстового документа» и «Распознавания заимствованных фрагментов» лежат представленные ранее соответствующие алгоритмы, причём, для второго алгоритма было взято пороговое значение $\mu = 4$. Подсистема «Машинного перевода», а также базовая автоматическая лингвистическая обработка текстовых документов реализована с использованием лингвистического процессора известной системы МП с белорусского языка на русский и обратно СМП Б/Р.

ИПК обеспечивает пользователю различные режимы работы в зависимости от типа поискового пространства, способов формирования ПОЗа, а также личных предпочтений, но в любом случае формирует и предоставляет ему отчёт, подсистема «Формирования отчёта», о результатах автоматической экспертизы в виде сравнительной таблицы (рисунок 2) достаточной для принятия пользователем решения на предмет наличия плагиата.

C:\PlgInternet\Волк.rtf	c:\PlgInternet\Saved\13032013 123710\k0wbdmwq.yru.html
|0.| Кроме того товарная политика предполагает определенный набор действий или заранее обдуманных методов и принципов деятельности, благодаря которым обеспечивается преемственность и целенаправленность мер по формированию и управлению ассортиментом товаров.	|1.| Товарная политика предполагает определенный набор действий или заранее обдуманных методов и принципов деятельности, благодаря которым обеспечивается преемственность и целенаправленность мер по формированию и управлению ассортиментом товаров.[1]
|0.| Обеспечить преемственность решений и мер по формированию оптимального ассортимента.	|1.| Цели товарной политики: обеспечить преемственность решений, мер по формированию оптимального ассортимента; поддерживать конкурентоспособность товаров на заданном уровне; целенаправленно адаптировать ассортиментный набор к требованиям рынка и покупателей; находить для товаров перспективные сегменты и ниши; способствовать разработке и осуществлению стратегии товарных знаков, упаковки, сервиса.
|0.| Способствовать разработке и осуществлению стратегии товарных знаков, упаковки, сервиса.	|1.| Цели товарной политики: обеспечить преемственность решений, мер по формированию оптимального ассортимента; поддерживать конкурентоспособность товаров на заданном уровне; целенаправленно адаптировать ассортиментный набор к требованиям рынка и покупателей; находить для товаров перспективные сегменты и ниши; способствовать разработке и осуществлению стратегии товарных знаков, упаковки, сервиса.
|0.| Вследствие снижения интереса со стороны потребителя любой товар рано или поздно уходит с рынка.	|1.| Любой товар рано или поздно уходит с рынка в силу различных причин, но главным образом вследствие снижения интереса к нему со стороны потребителя.
|0.| Жизненный цикл товара может быть представлен в виде определенной последовательности стадий существования его на рынке с определенными временными рамками.	|1.| Жизненный цикл товара может быть представлен как определенная последовательность стадий существования его на рынке, имеющая определенные рамки.

Рисунок 2. – Фрагмент сравнительной таблицы текстов анализируемых документов

Тестирование и практическое использование ИПК подтвердили его высокую актуальность, а также эффективность принятых при разработке комплекса решений и полное его соответствие поставленной цели. В среднем время автоматической экспертизы одного текстового документа, например, дипломного проекта на ЭВМ массового типа составляет от 90 до 180 секунд, а время, необходимое эксперту для обработки полученных результатов с целью (не)признания фактов плагиата, составляет в среднем от 20 до 60 минут.

Инструментально-программный комплекс «ПлагиатКонтроль» внедрён и успешно используется в ряде университетов Республики Беларусь для экспертизы дипломных и курсовых работ и проектов, а также в ВАК Беларуси для экспертизы диссертаций с привлечением информационных ресурсов Национальной библиотеки Беларуси.

ЗАКЛЮЧЕНИЕ

Основные научные результаты диссертации

1. Разработана и обоснована структурно-функциональная схема оригинальной системы автоматического распознавания лексически и семантически ЗФ, ориентированной, в общем случае, на поисковое пространство, включающее как текстовые Интернет-документы, так и документы из БД пользователя, на многоязычность информационной среды и cross-language функциональность.

Сформулированы основные задачи, требующие в совокупности своего решения с целью реализации предложенной схемы: 1) поиска релевантных входному текстовых документов из поискового пространства; 2) автоматического распознавания языка текстовых документов; 3) машинного перевода текстовых документов и их поисковых образов во множестве заданных ЕЯ; 4) автоматического распознавания лексически и семантически ЗФ текстовых документов; 5) автоматического построения отчёта; 6) автоматического лингвистического анализа текстовых документов. Показано, что задачи 2), 4) и 5) в полном объёме требуют своего эффективного решения, задача 1) – для случая Интернет-документов. Для задач 3) и 6) сформулированы требования, предъявляемые к их выходному результату, в соответствии с которым для решения этих задач определены инструментально-программные средства из ряда существующих [1, 5, 17].

2. Построен эффективный алгоритм автоматического распознавания языка текстовых документов, основанный на комбинировании предложенных алфавитного метода, метода грамматических слов и алфавитно-триграммного метода. Его эффективность обосновывается тем, что он требует нетрудоёмкого статистического и лингвистического анализа языковых данных, имеет один из лучших показателей точности (99,8 %), причём, с учётом текстов типа одной короткой фразы (7 слов). Последнее особенно важно с точки зрения решения задачи для естественно-языковых запросов пользователя в вопросно-ответных системах [2, 9, 13].

3. В зависимости от характеристик поискового пространства и особенностей поисковой службы Google определена, с целью решения поставленной задачи, стратегия применения функциональности МП. Разработан общий алгоритм МП строящихся автоматически поисковых образов текстовых документов, основанный на бинарных словарях концептов, акций и их атрибутов, представленных в БД, аналогичной известной лексической БД WordNet. Указанные компоненты являются универсальными по отношению к различным ЕЯ и обеспечивают эффективное решение как задачи МП, так и задачи автоматического распознавания семантически ЗФ текстовых документов. Показано, что поисковые образы текстовых документов, получаемые с использованием известного TF*IDF метода, могут быть качественно скорректированы использованием отношений синонимии и поправок на весовые коэффициенты, учитывающих принадлежность лексических единиц к наиболее информативным лексико-грамматическим, синтаксическим и семантическим классам [4, 5, 11, 12, 17].

4. Разработан эффективный алгоритм решения задачи автоматического распознавания лексически ЗФ текстовых документов. В его основу положены понятия лексически равных с точностью до канонических форм слов и отношений синонимии предложений, процедуры построения обратных индексов I_T и

I_{DB} соответственно для входного текста T и текстов из поисковой БД, а также построения обратного индекса $I_C = I_T \cap I_{DB}$ и его сортировки. Показано, что начисляемые при построении I_C накапливаемые веса для сравниваемых предложений, сортировка I_C и использование вводимой эвристики μ, учитывающие статистические особенности ЕЯ, существенно оптимизируют трудоёмкость алгоритма. Разработан общий алгоритм решения оригинальной задачи автоматического распознавания семантически ЗФ текстовых документов. В его основу положена система знаний, ориентированная на распознавание в текстовых документах объектов и фактов, а также их атрибутов [3, 4, 7, 8, 10–12, 17].

5. Разработана промышленная система «ПлагиатКонтроль», включая её лингвистическое и универсальное по отношению к различным ЕЯ алгоритмическое и программное обеспечение, которая впервые обеспечила решение задачи автоматического распознавания лексически ЗФ текстовых документов в белорусско-русской языковой среде и с функциональностью cross-language. Осуществлено внедрение системы, которая в силу использования полученных концептуальных, алгоритмических и технологических решений, обладает высокими качественными и техническими характеристиками [6, 14–16, 18].

Рекомендации по практическому использованию результатов

Разработанные концепции, технологии, алгоритмы, лингвистическое обеспечение, отдельные модули и система «ПлагиатКонтроль» в целом могут быть использованы при разработке различных информационных систем, включая системы информационного поиска, вопросно-ответные системы, системы распознавания релевантных, вплоть до семантически эквивалентных, текстовых фрагментов, а также для повышения эффективности уже существующих систем указанного типа. Кроме того, они могут быть использованы в учебном процессе в высших учебных заведениях, осуществляющих подготовку специалистов в области современных информационных технологий, интеллектуальных информационных систем и компьютерной лингвистики.

Система «ПлагиатКонтроль» внедрена и успешно используется в ряде университетов Республики Беларусь для экспертизы дипломных и курсовых работ и проектов, а также в ВАК Беларуси для экспертизы диссертационных работ на предмет наличия в них плагиата. Результаты диссертации также внедрены в учебный процесс БрГТУ. Внедрения подтверждены актами о внедрении результатов.

СПИСОК ПУБЛИКАЦИЙ СОИСКАТЕЛЯ УЧЕНОЙ СТЕПЕНИ

Статьи в научных изданиях в соответствии с п. 18 Положения о присуждении ученых степеней и присвоении ученых званий в Республике Беларусь

1. Крапивин, Ю.Б. К задаче автоматического распознавания воспроизведенных фрагментов текстовых документов / Ю.Б. Крапивин // Вестник БрГТУ, серия «Физика, математика, информатика». – 2009. – № 5 (59). – С. 109–112.

2. Крапивин, Ю.Б. Автоматическое определение языка текстового документа для основных европейских языков / Ю.Б. Крапивин // Информатика. – 2011. – № 3 (31). – С. 112–117.

3. Крапивин, Ю.Б. Автоматический поиск заимствованных из Интернет-источников фрагментов / Ю.Б. Крапивин // Искусственный интеллект. – 2012. – № 4. – С. 183–189.

4. Крапивин, Ю.Б. Функциональность cross-language в задаче автоматического распознавания семантически эквивалентных фрагментов текстовых документов / Ю.Б. Крапивин // Искусственный интеллект. – 2013. – №. 4. – С. 187–194.

5. Крапивин, Ю.Б. Лингвистический анализ текста в задаче автоматического распознавания заимствованных фрагментов текстовых документов / Ю.Б. Крапивин // Вестник БрГТУ, серия «Физика, математика, информатика». – 2017. – № 5. – С. 54–58.

6. Крапивин, Ю. Б. Система «ПлагиатКонтроль» как инструмент экспертизы текстовых документов / Ю. Б. Крапивин // Информатика. – 2018. – Т. 15, № 1. – С. 103–109.

Статьи в сборниках материалов научных конференций

7. Крапивин, Ю.Б К задаче автоматического распознавания схожих документов / Ю.Б. Крапивин // Искусственный интеллект. Интеллектуальные системы ИИ-2008: материалы Междунар. науч.-техн. конф., (пос. Кацивели, АР Крым, 22–27 сентября 2008 г). / НАН Украины, Мин. обр-я и науки Украины, Российская акад-я наук, НАН Беларуси, Институт проблем искусственного интеллекта; ред.: С.Б. Иванова. – Донецк: ИПИИ "Наука і освіта", 2008. – С. 399–401.

8. Krapivin, Y. Plagiarism Identification In Multilingual Information Environment / Y. Krapivin // OWD 2011 : proceedings of the XIII International PhD

Workshop, Wisla, Poland, October 22–25, 2011 / Silesian University of Technology; ed. : M. Szczygiel. – Gliwice, 2011. – P. 107–109.

9. Крапивин, Ю. Б. Идентификация языка текстовых документов и запросов в задачах их автоматической обработки / Ю. Б. Крапивин // Молодые ученые в инновационном поиске : материалы Междунар. науч. конф., Минск, 25 мая 2012 г. : в 2 ч. / Мин. гос. лингвист. ун-т ; редкол.: Т. П. Карпилович (отв. ред.) [и др.]. – Минск, 2012. – Ч. 2. – С. 70–75.

10. Крапивин, Ю.Б Автоматическое распознавание фрагментов текстового документа, заимствованных из интернет-доступных источников / Ю.Б. Крапивин // Искусственный интеллект. Интеллектуальные системы ИИ-2012: материалы Междунар. науч.-техн. конф., (пос. Кацивели, АР Крым, 1–5 октября 2012 г). / НАН Украины, Мин. обр-я и науки Украины, Российская акад-я наук, НАН Беларуси, Институт проблем искусственного интеллекта; ред.: С.Б. Иванова. – Донецк: ИПИИ "Наука і освіта", 2012. – С. 48–52.

11. Krapivin, Y. Automatic identification of the semantically equivalent fragments of the text documents / Y. Krapivin // OWD 2012 : proceedings of the XIV International PhD Workshop, Wisla, Poland, October 20–23, 2012 / Silesian University of Technology; ed. : M. Szczygiel. – Gliwice, 2012. – P. 290–292.

12. Крапивин, Ю.Б. Распознавание семантически эквивалентных фрагментов текстовых документов в многоязычной среде / Ю.Б. Крапивин // Искусственный интеллект. Интеллектуальные системы ИИ-2013: материалы Междунар. науч.-техн. конф., (пос. Кацивели, АР Крым, 23–27 сентября 2013 г). / НАН Украины, Мин. обр-я и науки Украины, Российская акад-я наук, НАН Беларуси, Институт проблем искусственного интеллекта; ред.: С.Б. Иванова. – Донецк: ИПИИ "Наука і освіта", 2013. – С. 95–99.

13. Krapivin, Y. Identification of the language in the task of the automatic recognition of reproduced fragments of the text documents / Y. Krapivin // Transcom 2013 : proceedings of the 10th European conference of young researchers and scientists, Žilina, Slovak Republic, June 24–26, 2013 / University of Žilina; ed. : M. Kochláň, A. Lieskovský. – Zilina, 2013. – P. 53–56.

14. Krapivin, Y. System of the automatic recognition of reproduced fragments of the text documents as a tool of improvement the quality control of educational process / Y. Krapivin // OWD 2013 : proceedings of the XV International PhD Workshop, Wisla, Poland, October 19–22, 2013 / Silesian University of Technology; ed. : G. Klapyta. – Gliwice, 2013. – P. 383–385.

15. Крапивин, Ю.Б. Об использовании системы автоматического распознавания воспроизведенных фрагментов текстовых документов в учебном процессе / Ю.Б. Крапивин // Информационные технологии и системы 2013 (ИТС 2013) : материалы международной научной конференции, Минск, Беларусь, 23

октября 2013 г./ редкол. : Л. Ю. Шилин [и др.]. – Минск : БГУИР, 2013. – С. 142–143.

16. Крапивин, Ю.Б. К задаче повышения качества контроля учебного процесса на этапе дипломного проектирования в вузе / Ю.Б. Крапивин // Инновационные технологии обучения физико-математическим дисциплинам : материалы VI Междунар. науч.-практ. интернет-конференции, Мозырь, 25–28 марта 2014 г. / УО МГПУ имени И.П. Шамякина ; редкол. : И.Н. Ковальчук (отв. ред.) [и др.]. – Мозырь, 2014. – С. 35.

17. Крапивин, Ю.Б. Лингвистическая составляющая в задаче автоматического распознавания заимствованных фрагментов / Ю.Б. Крапивин // Международный конгресс по информатике: информационные системы и технологии: материалы междунар. науч. конгресса, Республика Беларусь, Минск, 24–27октяб. 2016 г. / редкол.: С.В. Абламейко (отв. ред.) [и др.]. – Минск: БГУ, 2016. – С. 553–557.

18. Крапивин, Ю.Б. Помощник эксперту – «ПлагиатКонтроль» / Ю.Б. Крапивин // Библиотеки в информационном обществе: сохранение традиций и развитие новых технологий. Тема 2018 года – «Научная библиотека как центр культурно-информационного пространства» : докл. III Междунар. науч. конф., Минск, 6–7 декабря 2018 г. / Белорус. с.-х. б-ка им. И.С. Лупиновича Нац. акад. наук Беларуси ; редкол.: В.Н. Гердий [и др.]. – Минск, 2018. – С. 68–73.

РЕЗЮМЕ

Крапивин Юрий Борисович

«Методы и алгоритмы автоматического распознавания воспроизведённых фрагментов текстовых документов»

Ключевые слова: информационный поиск, естественный язык, анализ текста, лингвистический процессор, машинный перевод, алгоритм, заимствованный фрагмент, плагиат.

Цель работы: разработка принципов, методов, алгоритмов и программных средств решения задачи автоматического распознавания ЗФ текстовых документов в многоязычной информационной среде.

Методы исследования и использованная аппаратура: методы компьютерной лингвистики, теория формальных языков, анализа и синтеза текста, объектно-ориентированное программирование, экспертное тестирование.

Полученные результаты и их новизна: разработана концепция автоматического распознавания лексически и семантически ЗФ, ориентированная, в общем случае, на поисковое пространство, включающее как текстовые Интернет-документы, так и документы из БД пользователя, на многоязычность информационной среды и cross-language функциональность. Разработан алгоритм автоматического распознавания языка текстовых документов, требующий нетрудоёмкого статистического и лингвистического анализа языковых данных. Предложена стратегия применения функциональности машинного перевода и его общий алгоритм для строящихся автоматически поисковых образов текстовых документов. Разработан алгоритм решения задачи автоматического распознавания лексически ЗФ текстовых документов, в основу которого положены понятие лексически равных, с точностью до канонических форм слов и отношений синонимии, предложений и процедуры построения обратных индексов, а также общий алгоритм решения задачи автоматического распознавания семантически ЗФ текстовых документов, основанный на системе знаний, ориентированной на распознавание в текстовых документах объектов, фактов и их атрибутов. Все результаты являются новыми.

Рекомендации по использованию: разработанные методы, алгоритмы и программные средства могут быть использованы при построении систем информационного поиска и автоматического анализа текста.

Область применения: Полученные результаты внедрены в ряде учреждений Республики Беларусь, а также в учебный процесс в Учреждении образования «Брестский государственный технический университет».

РЭЗЮМЭ

Крапівін Юрый Барысавіч

«Метады і алгарытмы аўтаматычнага распазнавання васпраізведзеных фрагментаў тэкставых дакументаў»

Ключавыя словы: інфармацыйны пошук, натуральная мова, аналіз тэксту, лінгвістычны працэсар, машынны пераклад, алгарытм, запазычаны фрагмент, плагіят.

Мэта працы: распрацоўка прынцыпаў, метадаў, алгарытмаў і праграмных сродкаў рашэння задачы аўтаматычнага распазнавання ЗФ тэкставых дакументаў у шматмоўным інфармацыйным асяроддзі.

Метады даследвання і ўжыванае абсталяванне: метады камп’ютарнай лінгвістыкі, тэорыя фармальных моў, аналізу і сінтэзу тэкста, аб’ектна-арыентаванае праграмаванне, экспертнае тэставанне.

Атрыманыя вынікі і іх навізна: распрацавана канцэпцыя аўтаматычнага распазнавання лексічна- і семантычна-запазычаных фрагментаў, арыентаваная, увогуле, на пошукавую прастору, якая ўключае як тэкставыя Интэрнэт-дакументы, так і дакументы з базы дадзеных карыстальніка, на шматмоўнасць інфармацыйнага асяроддзя і cross-language-функцыянальнасць. Распрацаваны алгарытм аўтаматычнага распазнавання мовы тэкставых дакументаў, які патрабуе непрацаёмкага статыстычнага і лінгвістычнага аналіза моўных дадзеных. Прапанавана стратэгія выкарыстання функцыянальнасці машыннага пераклада і яго агульны алгарытм для пошукавых вобразаў тэкставых дакументаў, якія будуюцца аўтаматычна. Распрацаваны алгарытм рашэння задачы аўтаматычнага распазнавання лексічна-запазычаных фрагментаў тэкставых дакументаў, у аснову якога паложаны паняцце лексічна-аднолькавых, з дакладнасцю да кананічных формаў слоў і адносін сінаніміі, сказаў і працэдуры пабудовы адваротных індэксаў, а таксама агульны алгарытм рашэння задачы аўтаматычнага распазнавання семантычна-запазычаных фрагментаў тэкставых дакументаў, заснаваны на сістэме ведаў, арыентаванай на распазнаванне ў тэкставых дакументах аб'ектаў, фактаў і іх атрыбутаў. Усе вынікі з’яўляюцца новымі.

Рэкамендацыі да выкарыстання: распрацаваныя метады, алгарытмы і праграмныя сродкі могуць быць выкарыстаны пры пабудове сістэм інфармацыйнага пошуку і аўтаматычнага аналізу тэксту.

Вобласць ужывання: атрыманыя вынікі ўкарэнены ў шэрагу ўстаноў Рэспублікі Беларусь, а таксама ў навучальны працэс Установы адукацыі «Брэсцкі дзяржаўны тэхнічны ўніверсітэт».

SUMMARY

Krapivin Yury

«Methods and algorithms of automatic recognition of the reproduced fragments of the text documents»

Keywords: information retrieval, natural language, analysis of text, linguistic processor, machine translation, algorithm, adopted fragment, plagiarism.

The purpose of the work: to develop principles, methods, algorithms and software applications to solve the task of the automatic recognition of AF of the text documents in multilingual information environment.

Research methods and used equipment: computational linguistics methods, subject-domain specific languages methodology, analysis and synthesis of text, object-oriented programming, expert-based testing.

Obtained results and their novelty: a concept of automatic recognition of lexically and semantically AF, that is oriented, in general case, to the search space containing both text Internet-documents and user's database documents, to the multilinguality of the information environment and cross-language functionality, is developed. An algorithm of automatic recognition of the language of the text documents, which needs the effortless statistic and linguistic analysis of language data, is developed. A strategy of application of machine translation functionality, and its general algorithm of automatically constructed search profiles of the text documents are proposed. An algorithm of solving the task of automatic recognition of lexically AF of the text documents, which is based on the concept of lexically equivalent sentences, with precision up to canonical forms of words and synonymy relations, and the inverted indices construction procedures, as well as a general algorithm of solving the task of automatic recognition of semantically AF of the text documents, which is based on the knowledge system, oriented to the recognition of the concepts, actions and their attributes in the text documents, are developed. All results are new.

Recommendations for use: developed methods, algorithms and software applications can be used to construct systems of information retrieval of text and an automatic analysis of text.

Application area: presented results are implemented in a number of institutions of the Republic of Belarus, as well as in the teaching process at Brest state technical university.