На правах рукописи

Швец Александр Валерьевич

ВЗАИМОДЕЙСТВИЕ ИНФОРМАЦИОННЫХ И ЛИНГВИСТИЧЕСКИХ МЕТОДОВ В ЗАДАЧАХ АНАЛИЗА КАЧЕСТВА НАУЧНЫХ ТЕКСТОВ

Специальность 05.13.17 – Теоретические основы информатики

АВТОРЕФЕРАТ
диссертации на соискание ученой степени
кандидата технических наук

Москва – 2015

Работа выполнена в Федеральном государственном бюджетном учреждении науки Институте системного анализа Российской академии наук.

Научный руководитель: **Осипов Геннадий Семенович,** доктор физико-математических наук, профессор

Официальные оппоненты: **Сулейманов Джавдет Шевкетович,** доктор технических наук, академик, Академия наук Республики Татарстан, директор Научно-исследовательского института «Прикладная семиотика» АН РТ

Князева Анна Анатольевна, кандидат технических наук, младший научных сотрудник лаборатории численного моделирования и высокопроизводительных ресурсов томского филиала Института вычислительных технологий СО РАН

Ведущая организация: Федеральное государственное бюджетное образовательное учреждение высшего профессионального образования «Национальный исследовательский университет «МЭИ»

Защита состоится «7» октября 2015 года в 16 часов 30 минут на заседании диссертационного совета Д 002.073.01 на базе Федерального государственного учреждения «Федеральный исследовательский центр «Информатика и управление» Российской академии наук» (ФИЦ ИУ РАН) по адресу: 119333, г. Москва, ул. Вавилова, д. 44, корп. 2.

С диссертацией можно ознакомиться в библиотеке ФИЦ ИУ РАН и на официальном сайте ФИЦ ИУ РАН: http://www.ipiran.ru/.

Автореферат разослан «____» августа 2015 г.

Ученый секретарь
диссертационного совета Д 002.073.01,
доктор технических наук, профессор С.Н. Гринченко

ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

Актуальность темы исследования. В открытой научной печати регулярно появляются тексты, которые не прошли должную проверку перед изданием. Они не соответствуют требованиям, предъявляемым к научным публикациям, содержат различные нарушения или вовсе являются псевдонаучными. Такие тексты встречаются в научных журналах (обычно не рецензируемых), в материалах конференций и в других источниках научной сферы (под источниками научной сферы понимаются издания открытой печати и информационные ресурсы, которые позиционируют себя как научные). В большинстве случаев нарушения приводят к снижению ясности изложения, что вводит в заблуждение как исследователей, которые знакомятся с новой для них научной областью, так и аналитиков, работающих с большими объемами данных, у которых нет возможности рассматривать каждый текст детально. Существующие методы автоматического анализа текстов не ориентированы на проверку качества анализируемых текстов. Они позволяют выполнять поиск релевантных запросу публикаций, структурировать данные, извлекать полезную информацию, однако отсутствие этапа, на котором определяется надежность источника и возможность использования содержащейся в нем информации, часто приводит к некорректным, необъективным результатам. В связи с этим требуется разработка методов и программных средств автоматического определения признаков, характеризующих качество текстов научной сферы, и выявления псевдонаучных текстов. Под качеством понимается совокупность характеристик, включающих оценку лексики и множества синтактико-семантических структур текста, оценку наличия лингвистических ошибок, оценку наличия псевдонаучных фрагментов, оценку формальной структуры текста, т. е. наличия в тексте необходимых разделов (например, описания результатов). Настоящая работа посвящена созданию методов интеллектуального анализа текстов, направленных на решение указанных задач, что свидетельствует о ее актуальности.

Извлечение признаков, характеризующих качество текста, опирается на лингвистические структуры, выделяемые в тексте посредством синтаксического и семантического анализа, а также на информационные методы: статистические, индуктивного порождения гипотез и машинного обучения. Множество признаков формируется на основе лексических, морфологических, синтаксических и информационных характеристик текстов научной сферы.

Научная задача. Разработка нового метода определения качества текстов научной сферы, основанного на автоматическом выявлении лексических, синтаксических, морфологических и информационных признаков.

Предмет исследования – методы автоматического обнаружения признаков, характеризующих качество текстов научной сферы.

Основной целью является автоматизация определения качества текстов научной сферы. Для достижения цели работы поставлены следующие **задачи**:

1. Выполнить анализ методов определения различных характеристик и свойств текстов научной сферы.
2. Разработать метод автоматического формирования общенаучного словаря устойчивых словосочетаний.
3. Разработать метод автоматического выявления структуры научной публикации.
4. Разработать метод автоматического обнаружения лингвистических ошибок.
5. Разработать метод автоматического определения псевдонаучных фрагментов текстов научной сферы.
6. Сформировать признаковое пространство для автоматического определения научных и псевдонаучных текстов.
7. Проверить экспериментально разработанные методы.

Методы исследования. В диссертации использованы методы интеллектуального анализа текстов, статистические методы, методы машинного обучения, методы снижения размерности признакового пространства, индуктивные методы порождения гипотез, метод реляционно-ситуационного анализа текстов.

Научная новизна и результаты, выносимые на защиту.

1. Разработан новый метод автоматического формирования общенаучного словаря устойчивых словосочетаний.
2. Разработан новый метод автоматического выявления структуры научной публикации.
3. Разработан новый метод обнаружения нарушений правил согласования, нарушений синтаксической и семантической связности, лексической избыточности, нарушений последовательности изложения.

4. Впервые разработан метод автоматического выявления псевдонаучных фрагментов текстов научной сферы.
5. Сформировано множество признаков, характеризующих качество текстов научной сферы.
6. Построено множество правил для обнаружения псевдонаучных текстов.

Теоретическая значимость работы состоит в создании новых методов автоматического выявления признаков, характеризующих качество текстов научной сферы, на основе взаимодействия информационных и лингвистических методов.

Практическая значимость. Результаты работы могут применяться в системах поддержки принятия решений при отборе заявок, проектов, приеме отчетов, статей для публикации в научных журналах и в трудах конференций, а также для решения иных задач интеллектуального анализа информации. Разработанные методы извлечения признаков научного текста и метод обнаружения псевдонаучных текстов могут применяться в системах поиска и анализа научной информации.

Реализация результатов работы. Разработанные методы определения качества текстов научной сферы реализованы в виде программных средств и внедрены в следующие организации:

- Государственная публичная научно-техническая библиотека (информационная система «ЭКБСОН»);
- ООО «Национальный цифровой ресурс «Руконт» (электронно-библиотечная система «Руконт»);
- ООО «Научно-издательский центр ИНФРА-М» (электронно-библиотечная система «Znanium.com»);
- ЗАО «РосИнтернет технологии» (система интеллектуального поиска и анализа научных публикаций «Exactus Expert»).

Разработанные методы, правила и алгоритмы использованы в рамках научно-исследовательских работ по следующим проектам Минобрнауки РФ, программам ОНИТ РАН и грантам РФФИ:

1. «Создание программного комплекса информационно-аналитической поддержки научно-технической деятельности на основе вычислительного семантического поиска и анализа неструктурированной текстовой информации» (*ФЦП, № 07.551.11.4003, 2011-2013 гг.*);

2. «Разработка вычислительных методов объективной оценки качества научно-технических документов на естественных языках» (*ФЦП, № 14.514.11.4018, 2012-2013 гг.*);
3. «Исследование и разработка методов и алгоритмов связанности сложно-структурированных данных в научно-технической сфере» (*ФЦП, № 14.514.11.4024, 2012-2013 гг.*);
4. «Развитие методов и технологии семантического поиска и анализа научных публикаций Exactus Expert» (*в рамках проекта 2.9 ОНИТ РАН 2012-2013 гг.*);
5. «Исследование методов и разработка моделей и средств оценки научных текстов на основе их когнитивных структур» (*грант РФФИ № 14-29-05028-офи_м, 2014-2016 гг.*).

Достоверность результатов подтверждена проведенными вычислительными экспериментальными исследованиями программных средств, реализующих предложенные методы, правила и алгоритмы.

Апробация результатов исследования. Основные положения работы докладывались и обсуждались на следующих научных конференциях:

- XVI Международная научная конференция «Решетневские чтения», ноябрь 2012, г. Красноярск.
- Тринадцатая национальная конференция по искусственному интеллекту с международным участием КИИ-2012, октябрь 2012, г. Белгород.
- Пятая международная конференция «Системный анализ и информационные технологии» (САИТ-2013), сентябрь 2013, г. Красноярск.
- 20-я Международная конференция «Библиотеки и информационные ресурсы в современном мире науки, культуры, образования и бизнеса», июнь 2013, г. Судак.
- III Всероссийская научная конференция молодых ученых с международным участием «Теория и практика системного анализа» (ТПСА’14), май 2014, г. Рыбинск.
- Восемнадцатая международная научно-практическая конференция «SCIENCE ONLINE: электронные информационные ресурсы для науки и образования», май 2014, г. Белек.

– 7-я Международная конференция «Интеллектуальные системы» IEEE (The 7th IEEE International Conference Intelligent Systems, IS’2014 IEEE, Warsaw), сентябрь 2014, г. Варшава.
– Шестая международная конференция «Системный анализ и информационные технологии» (САИТ-2015), июнь 2015, г. Светлогорск.

Публикации. По теме диссертации опубликовано 9 работ, из них 4 в рецензируемых изданиях, рекомендованных ВАК РФ и приравненных к ним, и 2 зарегистрированные программные системы.

Структура и объем работы. Диссертация состоит из введения, трех глав, заключения, списка использованных источников и приложения. В приложении приведены описания программ, реализующих алгоритмы, предложенные в работе. Работа изложена на 120 страницах машинописного текста, содержит 21 таблицу и 24 рисунка. Список использованных источников включает 94 наименования.

КРАТКОЕ СОДЕРЖАНИЕ РАБОТЫ

Во **введении** обоснована актуальность темы, определен предмет исследования, сформулированы цель и задачи исследования, научная новизна, теоретическая и практическая значимость полученных результатов, а также приведены данные о структуре и объеме диссертации.

В **первой главе** рассматриваются нарушения в текстах научной сферы, которые могут приводить к снижению ясности изложения текста и к отсутствию возможности оценить исследование, применить описанные методы и воспроизвести эксперименты. В первом параграфе приводится типология нарушений в научных публикациях и исследуется возможность их выявления с помощью анализа лексики и синтаксических структур.

Среди множества нарушений в текстах научной сферы можно выделить следующие типы:

– Нарушение требований к лексике научного текста;
– Нарушение структуры научного текста;
– Нарушение правил согласования;
– Нарушение синтаксической и семантической связности;
– Лексическая избыточность (употребление плеоназмов);
– Нарушение последовательности изложения.

Наличие или отсутствие в тексте определенного нарушения является признаком, характеризующим качество текста. Систематизация предложенных в работе признаков представлена на рис. 1.

Рисунок 1 – Признаки, характеризующие качество текстов научной сферы

Приведем несколько примеров предложений из научных статей, содержащих различные нарушения. Курсивом выделены места в предложениях, характеризующие нарушения.

Пример 1 (нарушение требований к лексике): «И что об этом *думают* сами *языковеды*? Не стану добавлять имеющуюся *словесную чепуху* с целью придания наукообразия ссылками на разнообразные мнения на сей счет. Их *без труда* можно *найти* в Интернете».

Пример 2 (нарушение правил согласования): «Такие факторы как *возраст*, *образование*, социальный *статус* обычно *оказывает* существенное влияние на речевое поведение носителя языка».

Пример 3 (нарушение семантической связности): «Сформулировать и *доказать о свойствах* прямоугольных треугольников».

Пример 4 (лексическая избыточность): «То, что я назвал понятием, в этих *школах* обычно называют *содержанием* понятия, хотя *содержание* этого *содержания* может несколько варьироваться от *школы* к *школе* и соответственно отличаться от моего».

Обозначенные в примерах ошибки могут быть выявлены путем анализа лексики и синтаксических и семантических структур, которые могут быть выделены в тексте автоматически с помощью методов обработки естественного языка.

На рис. 2 представлена предлагаемая в настоящей работе схема извлечения рассмотренных признаков, характеризующих качество текста. Сначала происходит формирование базовых средств, а именно формирование общенаучного словаря устойчивых словосочетаний, выявление маркеров структурных разделов и формирование правил, характеризующих лингвистические ошибки. Затем выполняется анализ конкретного текста и извлечение его характеристик с помощью синтаксического и семантического анализа. После этого применяются методы выявления нарушений, которые оперируют с извлеченными характеристиками и сформированными базовыми средствами. В результате происходит формирование множества признаков, характеризующих качество анализируемого текста. Все методы, соответствующие процессам на рис. 2, за исключением синтактико-семантического анализа текста, разработаны в рамках настоящей работы и описаны в главе 2.

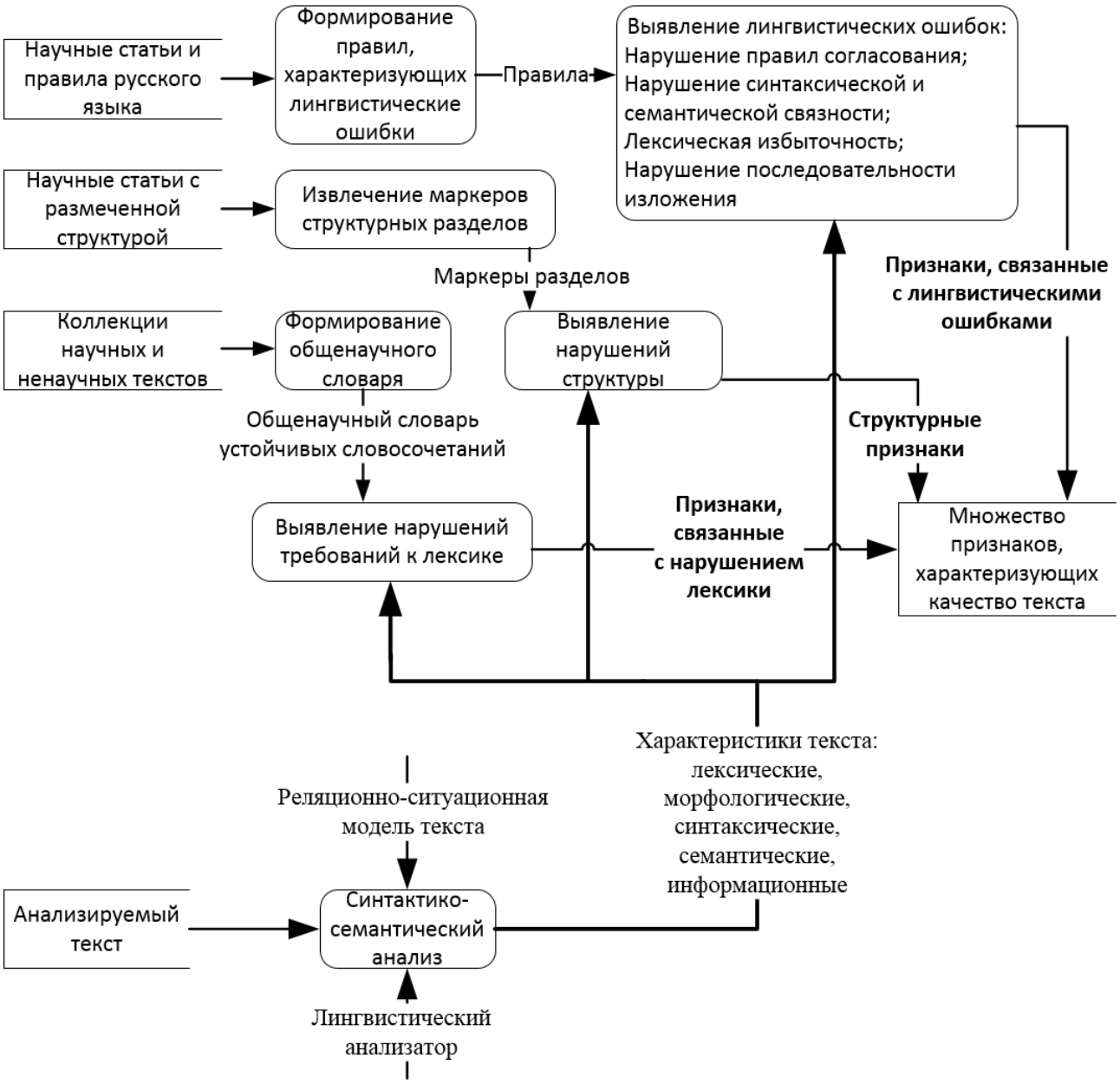

Рисунок 2 – Предлагаемая в работе схема выявления признаков, характеризующих качество текстов научной сферы

Во втором параграфе первой главы рассматриваются методы, позволяющие в некоторой степени выполнять автоматический анализ качества научных текстов. В заключительной части главы приведены основные выводы и сформулированы задачи исследования.

Вторая глава посвящена разработке методов выделения признаков, характеризующих качество текстов научной сферы, а именно разработке метода автоматического формирования общенаучного словаря устойчивых словосочетаний, метода автоматического выявления структурных разделов научной публикации и метода автоматического обнаружения лингвистических ошибок.

Особенностью всех предложенных методов является оперирование с полуструктурированными данными, которые формируются в результате синтактико-семантического анализа и представляют собой текст с установленными свойствами его элементов. Для выполнения синтаксического анализа текста используется его реализация в системе АОТ[1]. Для извлечения семантических характеристик текста в работе применяется метод реляционно-ситуационного анализа[2], который основан на теории коммуникативной грамматики русского языка и теории неоднородных семантических сетей.

Приведем основные алгоритмы, соответствующие разработанным методам (нумерация алгоритмов сохранена).

Пусть S^+ и S^- – множества предложений научных и ненаучных текстов, таких что $\left|S^+\right| \le \left|S^-\right|$. Требуется построить словарь словосочетаний W, в большей степени характерных для предложений множества S^+. Предлагается следующий алгоритм.

Алгоритм 2.1 **(алгоритм формирования общенаучного словаря устойчивых словосочетаний).**

Шаг 1. Задать множества $W^+=\{\varnothing\}$ и $W^-=\{\varnothing\}$ – множества словосочетаний, входящих в предложения множеств S^+ и S^- соответственно.

Шаг 2. Выполнить синтактико-семантический разбор каждого предложения множества S^+, расширяя множество W^+ словосочетаниями с синтаксическими и семантическими связями.

Шаг 3. Для каждого встретившегося словосочетания w_i $(i=\overline{1,\left|W^+\right|})$ подсчитать количество его вхождений n_i в множество предложений S^+ и определить значение функции $n^+(w)$ в точке w_i так, что $n^+(w_i)=n_i$. Пусть $n^+(w)=0$ для словосочетаний $w \notin W^+$.

Шаг 4. Выполнить синтактико-семантический разбор каждого предложения множества S^-, расширяя множество W^- словосочетаниями с синтаксическими и семантическими связями.

Шаг 5. Для каждого встретившегося словосочетания w_j $(j=\overline{1,\left|W^-\right|})$ подсчитать количество его вхождений m_j в множество предложений S^- и

[1] Сокирко А. В. Семантические словари в автоматической обработке текста: По материалам системы ДИАЛИНГ: Дисс. … канд. тех. наук. Москва, 2001. 120 с.

[2] Осипов Г. С., Смирнов И. В., Тихомиров И. А. Реляционно-ситуационный метод поиска и анализа текстов и его приложения // Искусственный интеллект и принятие решений. – 2008. – № 2. – С. 3-10.

определить значение функции $n^-(w)$ в точке w_j так, что $n^-(w_j)=m_j$. Пусть $n^-(w)=0$ для словосочетаний $w \notin W^-$.

Шаг 6. Сформировать множество W путем добавления в него словосочетаний $w \in W^+$, для которых выполняются неравенства $n^+(w) > n^-(w)$ и $n^+(w) > 1$.

Сложность алгоритма равна $O(|S^+|+|S^-|)$.

Словарь построен автоматически на базе Национального корпуса русского языка (НКРЯ)[3]. Получено свыше 500 тысяч словарных единиц. Исследована зависимость объема словаря от размера научного подкорпуса предложений. Показано, что 80% словаря формируется при анализе 45% предложений подкорпуса, разбор каждого последующего предложения добавляет незначительное число словосочетаний, что говорит о высокой полноте полученного словаря.

Проведены эксперименты, показывающие, что использование сформированного словаря позволяет отличить научные статьи от научно-популярных и ненаучных публикаций, однако не всегда возможно выявить псевдонаучные тексты.

Рассмотрена типичная структура научной публикации. Она включает в себя разделы, соответствующие формату IMRAD[4]: «Постановка проблемы», «Методы», «Результаты», «Выводы». Проанализированы современные методы, предназначенные для структурирования текста в соответствии с перечисленными разделами. Показаны их недостатки и выявлены ограничения на применение для определения наличия разделов.

Опишем предложенные в работе алгоритмы выявления маркеров и определения наличия разделов. Положим, $M_I=\{\emptyset\}$, $M_M=\{\emptyset\}$, $M_R=\{\emptyset\}$, $M_D=\{\emptyset\}$ – множества, которые необходимо заполнить маркерами, характеризующими разделы «Постановка проблемы», «Методы», «Результаты» и «Выводы» соответственно. Пусть S_I, S_M, S_R, S_D – множества предложений обучающей выборки, соответствующих указанным структурным разделам. Для выявления маркеров предлагается следующий алгоритм.

[3] Национальный корпус русского языка [Электронный ресурс]: URL: http://ruscorpora.ru/ (дата обращения: 31.05.2015).

[4] Sollaci L.B., Pereira M.G. The introduction, methods, results, and discussion (IMRAD) structure: a fifty-year survey //J. Med. Libr. Assoc. 2004. V.92(3) P.364-371.

Алгоритм 2.3 **(алгоритм выявления маркеров структурных разделов).**

Шаг 1. Выполнить синтактико-семантический разбор каждого предложения множества S_I, расширяя множество M_I словосочетаниями с синтаксическими и семантическими связями.

Шаг 2. Повторить шаг 1 для множеств S_M, S_R, и S_D.

Шаг 3. Определить степень принадлежности разделу каждого маркера m_i множества M_I ($i = \overline{1,|M_I|}$), используя метод сглаживания Лапласа, $V_{m_i} = \frac{n_{m_i} + \alpha}{N_{m_i} + k\alpha}$, где n_{m_i} – число вхождений маркера m_i в множество предложений S_I, N_{m_i} – общее число вхождений маркера m_i во все предложения обучающей выборки, k – число различных разделов (в этом случае $k=4$), α – произвольный коэффициент сглаживания (положим, $\alpha = 0{,}25$).

Шаг 4. Удалить из множества M_I маркеры со степенью принадлежности, не превышающей значение 0,5.

Шаг 5. Повторить шаги 3-4 для маркеров множеств M_M, M_R, и M_D. Маркеры построены.

Сложность алгоритма равна $O(|S_I|+|S_M|+|S_R|+|S_D|)$.

Пусть оценка соответствия некоторого текста структурному разделу вычисляется по совокупности маркеров этого раздела, входящих в текст, и зависит от маркера с максимальной оценкой принадлежности, от средней оценки принадлежности маркеров и от количества встретившихся маркеров, соответствующих разделу:

$$E = \begin{cases} \mathrm{V}_{\max} - \dfrac{\mathrm{V}_{\max} - \mathrm{V}_{\mathrm{avg}}}{5}, \textit{если } \dfrac{\mathrm{n}}{N} > C \\ 0, \textit{иначе} \end{cases} \quad (1)$$

где n – число маркеров раздела в тексте с повторениями, N – общее число семантических и синтаксических конструкций в тексте, V_{max} – максимальная оценка принадлежности, V_{avg} – средняя оценка принадлежности, C – константа, задающая приемлемое относительное количество маркеров. Константа C задается на этапе обучения, своя для каждого структурного раздела.

Пусть T – произвольный текст. Для определения наличия в нем структурного раздела предлагается следующий алгоритм.

Алгоритм 2.4 **(алгоритм определения наличия структурного раздела в тексте).**

Шаг 1. Разделить текст T на фрагменты F_i равной длины.

Шаг 2. Выбрать один из фрагментов F. Для этого фрагмента выполнить синтактико-семантический разбор предложений и построить множество словосочетаний M_F.

Шаг 3. Найти пересечение множеств $M_F^I = M_F \cap M_I$.

Шаг 4. Вычислить значение E по формуле (1), используя степени принадлежности маркеров множества M_I, входящих в множество M_F^I.

Шаг 5. Повторить шаги 2-4 для каждого фрагмента F_i. Максимальное значение E и будем считать оценкой наличия раздела в тексте.

Сложность алгоритма равна $O(N)$, где N – число семантических и синтаксических конструкций в тексте.

Экспериментально установлены три интервала значений оценки E, которые определяют наличие раздела и имеют следующие обозначения: «присутствует», «вероятно, отсутствует», «отсутствует».

В основе метода обнаружения лингвистических ошибок лежит некоторое множество правил R, с помощью которых можно выявить нарушения правил согласования, нарушения семантической связности, последовательности изложения и др. Для формирования множества правил в работе предложен следующий алгоритм.

Алгоритм 2.5 **(алгоритм формирования правила, характеризующего лингвистические ошибки).**

Шаг 1. Выбрать одно из правил русского языка r'.

Шаг 2. Исследовать примеры предложений из множества S^+, удовлетворяющих данному правилу, и примеры предложений с нарушением правила из множества S^-.

Шаг 3. Извлечь условия, выполнение которых свидетельствует о наличии ошибки. При формировании условий степень обобщения ограничивается множеством правильных предложений S^+.

Шаг 4. В выборке научных текстов выделить предложения S^0, для которых выполняются полученные условия.

Шаг 5. Если среди выделенных предложений содержатся правильные предложения ($S^0 \cap S^+ \neq \emptyset$) или обнаруживаются предложения с нарушениями S^-, которые не были выделены ($S^- \setminus S^0 \neq \emptyset$), и есть возможность уточнить условия, так чтобы правило покрывало меньше

предложений из S^+ и больше из S^-, то уточнить правило и выполнить шаг 4. Правило r является результатом последовательного итерационного уточнения условий.

С использованием описанного алгоритма получено 9 правил, покрывающих основные нарушения. Рассмотрим одно из правил: «*Если в состав предложения входят однородные подлежащие, принадлежащие к разному грамматическому роду, и сказуемое в форме глагола прошедшего времени единственного числа, то предложение содержит нарушение согласования сказуемого с однородными подлежащими*». Приведем пример предложения, найденного автоматически по этому правилу: «Несмотря на то, что все преобразования … существовали в разных видах в разных местах, … *выбор* и *ответственность* за него *ложился* на реформатора».

Тексты с низкой синтаксической и семантической связностью могут быть обнаружены в результате лингвистического анализа: они содержат большое число слов, отделенных от синтаксического дерева (отсутствует связь со словом-родителем) и не входящих в семантическую сеть. Следующее правило позволяет выявить такие тексты: «*Если в тексте превышено допустимое количество слов, не связываемых со словами-родителями, то степень синтаксической и семантической связности текста является низкой*». Допустимое количество несвязанных слов устанавливается автоматически при обучении на выборке научных статей.

Проведенные эксперименты подтверждают, что разработанные методы применимы для обнаружения различных нарушений и отступлений от норм научного текста.

В **третьей главе** выполняется исследование применимости разработанных методов. Поставлена задача обнаружения псевдонаучных текстов.

В первом параграфе приводится определение псевдонауки, используемое в настоящей работе. Под *псевдонаукой* понимается любая методология или система взглядов, которая претендует на то, чтобы считаться научной, но не соблюдает принципы доказательности и аргументированности, не соответствует ни нормам научного знания, ни какой-либо области действительности, а ее предмет либо не существует, либо сфальсифицирован[5,6].

[5] Кувакин В.А. Интернет пресс-конференция члена Комиссии РАН по борьбе с лженаукой и фальсификацией научных исследований, 04.05.2010 г.

[6] RationalWiki [Электронный ресурс]: URL: http://rationalwiki.org/ (дата обращения: 31.05.2015)

Второй параграф третьей главы посвящен разработке метода автоматического определения псевдонаучных фрагментов, заключающегося во взаимодействии лингвистических методов, используемых для выявления признаков классификации, и информационных методов, в частности, статистического метода, позволяющего устанавливать значимость признаков, и метода машинного обучения, который необходим для определения принадлежности фрагмента множеству псевдонаучных текстов.

В связи с тем, что псевдонаучные высказывания могут составлять лишь часть публикации, предлагается разбивать статьи на небольшие фрагменты текста, близкие по объему, и классифицировать их отдельно. Разбиение текста выполняется таким образом, чтобы фрагменты состояли из абзацев, поскольку абзац обычно несет в себе законченную мысль, и, как правило, позволяет получить представление о корректности входящих в него высказываний.

Множество признаков классификации формируется автоматически с помощью лингвистического анализатора на этапе обучения, описанного ниже. В качестве признаков классификации предлагается использовать:

– слова (например, "торсионный", "гармонизировать", "чрезвычайно", "неправота");

– словосочетания с синтаксическими и семантическими связями (например, "повсеместное наличие", "необъяснимая аномалия", "усматривать в модели", "убедительно показать", "память воды");

– обобщения словосочетаний (например, "память <сущ.>", "<прил.> аномалия", "усматривать в <сущ.>");

– триграммы (например, "я якобы сразу", "и почти нигде", "совершенно очевидно то", "сейчас наукой доказано").

Для придания большей значимости признакам, характерным лишь для псевдонаучных текстов, всем признакам назначаются веса, которые вычисляются для каждого фрагмента текста с помощью статистической меры *TF-IDF*[7], приведенной ниже.

Описаны условия формирования обучающей выборки для предметно независимой классификации.

В качестве классификатора выбран метод опорных векторов[8] (SVM – support vector machine), который хорошо зарекомендовал себя при классификации текстовой информации.

---

[7] Salton, G., Buckley, C. Term-weighting approaches in automatic text retrieval. Information Processing & Management. 24(5), 513-523 (1988).

[8] Cortes, C., Vapnik, V. Support-vector networks. Machine Learning. 20(3), 273(1995).

Приведем алгоритм обнаружения псевдонаучных фрагментов в тексте. Пусть T_x – множество псевдонаучных текстов обучающей выборки, T_y – множество научных текстов обучающей выборки, таких что предметные области множества T_y включают в себя все предметные области множества T_x. Тогда обучение состоит из следующих шагов.

Алгоритм 3.1 **(алгоритм обучения классификатора для обнаружения псевдонаучных фрагментов).**

Шаг 1. С помощью метода, представленного в разделе 2.2.2, выделить в текстах списки использованных источников и удалить их, получив два новых множества $T_x^{'}$ и $T_y^{'}$, состоящих лишь из авторского текста.

Шаг 2. Разделить тексты множеств на непересекающиеся фрагменты, длина которых не превосходит среднюю длину абзаца *l*, так, что $T_x^{'} = \bigcup x_i$, где x_i – псевдонаучный фрагмент, аналогично $T_y^{'} = \bigcup y_i$, где y_i – научный фрагмент.

Шаг 3. Задать множество $F=\{\varnothing\}$ – множество признаков для классификации. Выполнить синтактико-семантический анализ каждого фрагмента из множеств $T_x^{'}$ и $T_y^{'}$, расширяя множество *F* следующими признаками классификации *t*: словами, составляющими фрагменты, словосочетаниями с синтаксическими и семантическими связями, их обобщениями и триграммами.

Шаг 4. Каждому фрагменту $d \in T_x^{'} \cup T_y^{'}$ поставить в соответствие вектор длины $|F|$, состоящий из весов признаков, вычисленных по формуле *TF-IDF*: $tfidf(t,d,D) = tf(t,d) \cdot idf(t,D)$, где $tf(t,d) = \frac{n_t}{\sum_k n_k}$, где n_t – число вхождений признака *t* в фрагмент *d*, а в знаменателе – общее число признаков в данном фрагменте, $idf(t,D) = \log \frac{|D|}{|d_i \supset t|}$, где $|D|$ – количество фрагментов в обучающей выборке; $|d_i \supset t|$ – количество фрагментов, в которых встречается *t* (когда $n_t \neq 0$). Задать множество $IDF = \bigcup idf(t,D)$, необходимое для последующей классификации.

Шаг 5. Выполнить обучение с помощью алгоритма SVM на полученных векторах. В результате будет получена линейная модель классификации *M*.

Наибольшей алгоритмической сложностью в предложенном методе обучения обладает алгоритм SVM: в худшем случае она равна $O(N^3)$, в среднем – $O(N^2)$, где N – число обучающих примеров.

Алгоритм классификации тестового фрагмента d состоит в следующем.

Алгоритм 3.2 **(алгоритм классификации фрагмента текста).**

Шаг 1. Выполнить синтактико-семантический анализ предложений фрагмента d и извлечь признаки $F_d = \{t_i\}$.

Шаг 2. Для всех признаков, входящих в пересечение $F_d \cap F$, вычислить значение *tfidf*.

Шаг 3. Заполнить вектор длины $|F|$, используя вычисленные значения *tfidf*, и выполнить классификацию фрагмента с помощью модели *M*, полученной на этапе обучения.

Составлена коллекция псевдонаучных публикаций, содержащая более 4 тысяч статей, взятых из источников, рекомендованных комиссией РАН по борьбе с лженаукой и фальсификацией научных исследований. Эксперимент проводился на обучающей выборке, состоящей из 220 научных статей (3767 фрагментов) и 60 псевдонаучных статей (9001 фрагмент). Тестовая выборка состояла из 12624 авторефератов докторских диссертаций (1300K фрагментов) и 4203 статей псевдонаучных сообществ (130K фрагментов). Поскольку тестовая выборка значительно больше обучающей выборки, перекрестная проверка не проводилась.

Результаты классификации фрагментов приведены в табл. 1.

Таблица 1 – Результаты классификации фрагментов публикаций

Классы фрагментов публикаций		Настоящий класс	
		псевдонаучный (130K)	научный (1300K)
Оценка классификатора	псевдонаучный	83%	28%
	научный	17%	72%

Согласно табл. 1, точность классификации для класса «псевдонаучный» равна 0,75, полнота – 0,83, F_1-мера – 0,79. Все показатели качества классификации, такие как точность (precision), полнота (recall) и F_1-мера (F_1-measure), вычисляются согласно стандартным формулам[9].

[9] Powers, David M. W. Evaluation: From Precision, Recall and F-Factor to ROC, Informedness, Markedness and Correlation. Journal of Machine Learning Technologies. – V. 2(1). – pp. 37–63. – 2011.

Тексты тестовой выборки группировались по доле текста, которую занимают фрагменты, классифицированные как псевдонаучные. На рис. 3 и рис. 4 представлены полученные группы, каждой из которых соответствует отдельный столбик. Над столбиком указано относительное количество текстов в этой группе, под ним – та часть текста, которую занимают псевдонаучные фрагменты.

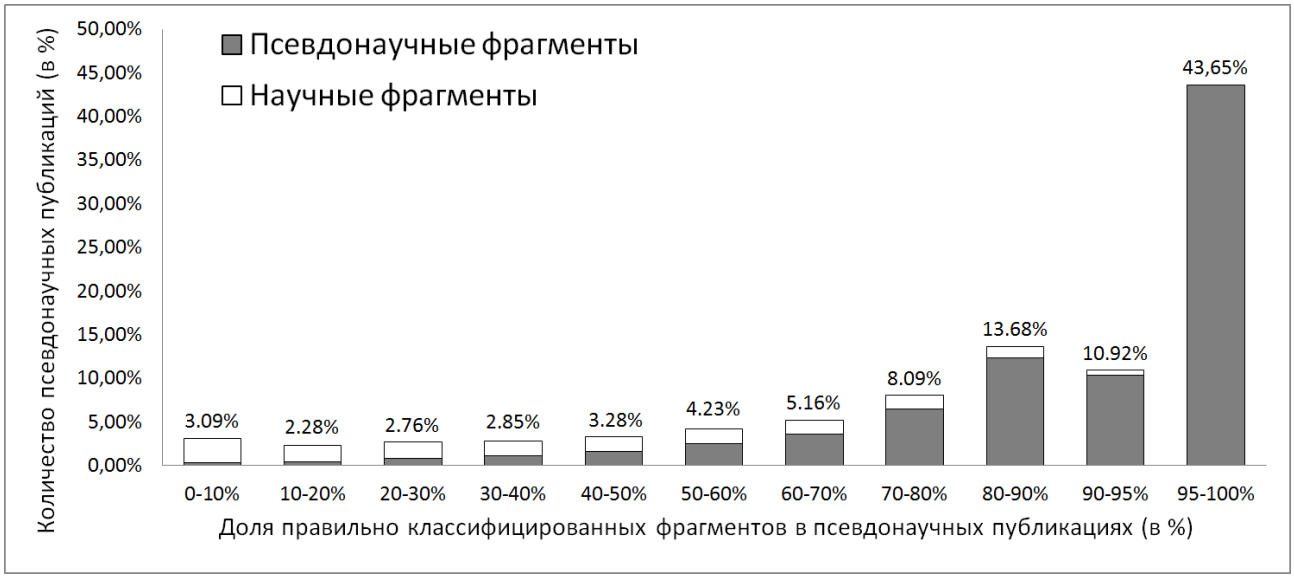

Рисунок 3 – Количество псевдонаучных фрагментов в псевдонаучных статьях (в %)

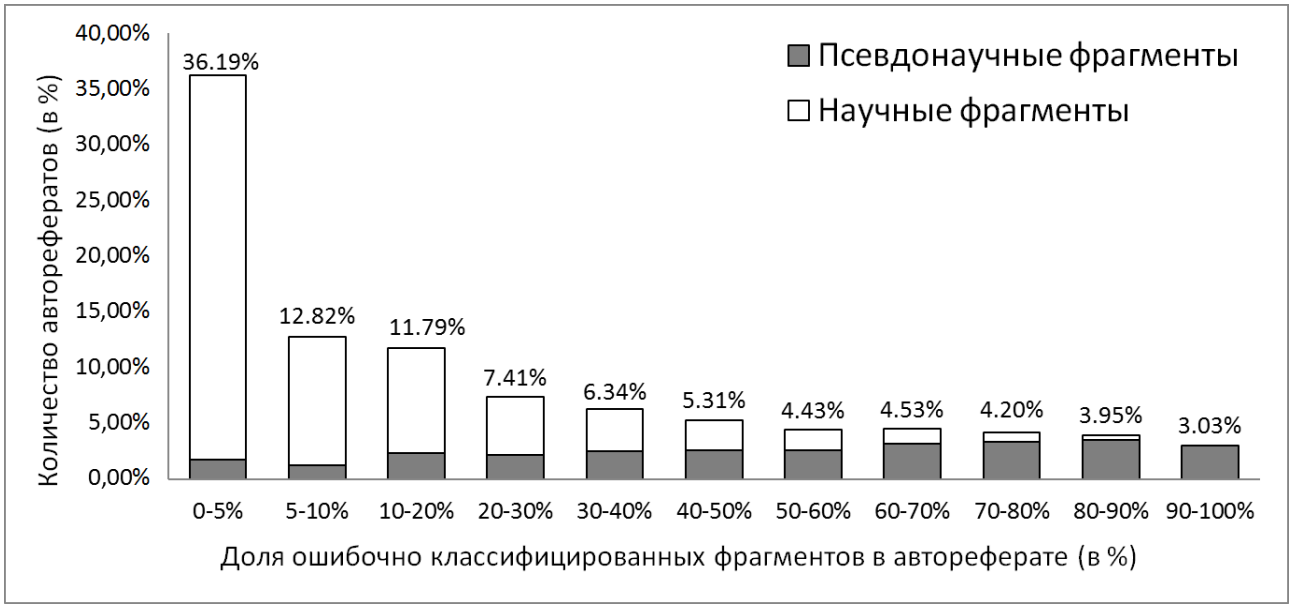

Рисунок 4 – Количество псевдонаучных фрагментов в научных публикациях (в %)

Примечание – окрашенная серым часть каждого столбика на рис. 3-4 соответствует диапазону, указанному под ним, который обозначает количество фрагментов в тексте, классифицированных как псевдонаучные. Так, левый столбик на рис. 4 означает, что количество авторефератов, в которых лишь от 0 до 5% фрагментов классифицировано неправильно, составляет 36.19% от всех авторефератов.

Результаты демонстрируют эффективность предложенного метода: большая часть псевдонаучных фрагментов классифицирована правильно и небольшая часть научных фрагментов классифицирована неправильно.

Согласно рис. 3, в большинстве псевдонаучных статей значительную часть текста составляют псевдонаучные фрагменты, тогда как в большинстве научных статей, согласно рис. 4, неправильно классифицированные фрагменты занимают лишь небольшой процент текста. В связи с этим показатель относительного количества псевдонаучных фрагментов в тексте выбран как дополнительный признак, характеризующий качество текстов научной сферы.

В третьем параграфе третьей главы формируется признаковое пространство и выполняется снижение его размерности с помощью индуктивного алгоритма AQ[10] с целью выявления признаков, значимых с точки зрения обнаружения псевдонаучных текстов. Применение алгоритма показало, что наиболее информативными являются структурные признаки, признаки, связанные с лингвистическими ошибками (нарушение связности текста, некоторые нарушения согласования, лексическая избыточность), а также признак, определяющий относительное количество псевдонаучных фрагментов в тексте. Признаки имеют дискретные значения, примеры приведены в табл. 2.

Таблица 2. Примеры признаков, характеризующих качество текстов

Название признака	Значение
Относительное количество устойчивых общенаучных словосочетаний	приемлемое
	заниженное
	низкое
Описание методов исследования	присутствует
	вероятно, отсутствует
	отсутствует
Количество нарушений синтаксической и семантической связности	низкое
	высокое

В четвертом параграфе третьей главы с использованием первого этапа ДСМ-метода[11] выполняется индуктивное построение множества правил для обнаружения псевдонаучных текстов.

Опишем формирование правил и признаков псевдонаучного текста в терминах ДСМ-метода. Пусть множество $O=\{o_i\}$ – множество объектов,

[10] The aq21 natural induction program for pattern discovery: Initial version and its novel features / Janusz Wojtusiak, Ryszard S. Michalski, Kenneth A. Kaufman, Jaroslaw Pietrzykowski // ICTAI. – 2006. – P. 523-526.

[11] Финн. В.К. ДСМ-метод как средство анализа каузальных зависимостей в интеллектуальных системах. // НТИ, № 11, 2000.

объектом в данном случае является текст научной сферы; множество $P=\{p_j\}$ – множество признаков, каждый из которых обладает своим множеством допустимых значений $p_j^1,...,p_j^{n_j}$. Один объект обладает одним значением каждого признака, которое называется свойством объекта. Каждый объект представляется в виде вектора свойств $o = \{p_1^{i_1},...,p_n^{i_n}\}$. Пусть $O^+ \subset O$ – положительные примеры (псевдонаучные тексты), $O^- \subset O$ – отрицательные примеры (научные тексты). Тогда гипотезы $H^+ = \bigcup p_j^{n_j}$, полученные на первом этапе ДСМ-метода и представляющие собой объединения значений некоторых признаков, и будут являться условиями принадлежности к множеству O^+: если объект o обладает признаками, которые удовлетворяют гипотезе $h \in H^+$ и не удовлетворяют ни одной гипотезе множества H^-, то $o \in O^+$. Каждая гипотеза $h \in H^+$ является интегральным признаком псевдонаучного текста.

В результате применения описанного индуктивного метода на обучающей выборке текстов получено 3 тысячи интерпретируемых правил, с которыми в дальнейшем может работать эксперт-аналитик. Ниже приведено несколько примеров:

1) **Если** (количество псевдонаучных фрагментов = *очень высокое*)
   И (описание методов исследования = *отсутствует*)
   То (публикация является псевдонаучной).

2) **Если** (выводы исследования = *вероятно отсутствуют*)
   И (количество устойчивых общенаучных словосочетаний = *заниженное*)
   То (публикация является псевдонаучной).

3) **Если** (количество нарушений синтаксической и семантической связности = *завышенное*)
   И (количество плеоназмов = *высокое*)
   И (список цитируемой литературы = *отсутствует*)
   И (выводы исследования = *вероятно отсутствуют*)
   То (публикация является псевдонаучной).

4) **Если** (количество псевдонаучных фрагментов = *очень высокое*)
   И (количество цитирований работ одного автора = *завышенное*)
   То (публикация является псевдонаучной).

Проведено сравнение метода выявления псевдонаучных текстов по построенным правилам с базовым методом, заключающимся в распределении текстов по классам в зависимости от показателя относительного количества

псевдонаучных фрагментов в тексте, рассмотренного во втором параграфе третьей главы. Выполнена процедура перекрестной проверки, результаты классификации представлены в табл. 3.

Таблица 3. Сравнение результатов классификации разными методами

	Точность	Полнота	F_1-мера
Разделение по количеству псевдонаучных фрагментов	0,68	0,89	**0,77**
ДСМ-метод (множество правил)	0,78	0,88	**0,83**

Таким образом, построенное множество правил позволяет проводить классификацию текстов с большим значением F_1-меры, что говорит о целесообразности применения разработанных в настоящей работе методов извлечения признаков, характеризующих качество текстов научной сферы.

В заключительной части третьей главы приводится сравнение методов машинного обучения, подходящих для решения задачи классификации текстов научной сферы для обнаружения псевдонаучных текстов на основе сформированного пространства признаков. Проведен комплекс экспериментов на различных выборках с применением перекрестной проверки и с многократными прогонами для усреднения результатов.

Наиболее высокие значения F_1-меры достигают метод опорных векторов и деревья решений[12]. Нейронные сети[13] (трехслойный персептрон) позволяют решать задачу с высокой точностью лишь при обучении на большом числе данных, ДСМ-метод, напротив, лучше работает при небольшой обучающей выборке. При этом все методы показывают высокие значения F_1-меры, что говорит о применимости сформированного пространства признаков к автоматическому обнаружению псевдонаучных текстов.

В **заключении** приводятся основные результаты, полученные в работе.

В **приложении** описаны реализованные программные модули, которые внедрены в программный комплекс интеллектуального поиска и анализа научных публикаций «Exactus Expert» и использованы при тестировании разработанных методов.

[12] Murthy S. Automatic construction of decision trees from data: A multidisciplinary survey. Data Mining and Knowledge Discovery. – 1998. – V. 2(4). – pp. 345-389.
[13] Hertz, J., Palmer, R. G., Krogh. A. S. Introduction to the theory of neural computation, Perseus Books. – 1990. – 327 p.

ОСНОВНЫЕ РЕЗУЛЬТАТЫ РАБОТЫ

1. Разработан новый метод автоматического формирования общенаучного словаря устойчивых словосочетаний.
2. Разработан новый метод автоматического выявления структуры научной публикации.
3. Разработан новый метод обнаружения нарушений правил согласования, нарушений синтаксической и семантической связности, лексической избыточности, нарушений последовательности изложения.
4. Впервые разработан метод автоматического выявления псевдонаучных фрагментов текстов научной сферы.
5. Сформировано множество признаков, характеризующих качество текстов научной сферы.
6. Построено множество правил для обнаружения псевдонаучных текстов.

ПУБЛИКАЦИИ АВТОРА ПО ТЕМЕ ИССЛЕДОВАНИЯ

Публикации автора в изданиях, входящих в перечень ВАК и приравненных к ним:

1. Shvets, A. A Method of Automatic Detection of Pseudoscientific Publications // Proceedings of the 7th IEEE International Conference Intelligent Systems (IS'2014 IEEE). Advances in Intelligent Systems and Computing (AISC). – Warsaw, 2015. – Vol. 2. – P. 533-539.
2. Osipov, G., Smirnov, I., Tikhomirov, I., Sochenkov, I., Shelmanov, A., and Shvets, A. Information Retrieval for R&D Support / Paltoglou, Georgios, Loizides, Fernando, Hansen, Preben (Eds.) Professional Search in the Modern World. Lecture Notes in Computer Science (LNCS). – 2014. – Vol. 8830. – P. 45-69.
3. Швец А.В., Кузнецова Ю.М., Осипов Г.С., Латышев А.В. Метод и алгоритм обнаружения признаков лингвистических дефектов в научно-технических текстах // Информационные технологии и вычислительные системы. – 2013. – № 2. – С. 79-87.
4. Кузнецова Ю.М., Осипов Г.С., Чудова Н.В., Швец А.В. Автоматическое установление соответствия статей требованиям к научным публикациям // Труды ИСА РАН. – 2012. – Т. 62. – Вып. 3. – С. 132-138.

Зарегистрированные программные системы:

5. Швец А.В., Смирнов И.В. Программа оценки соответствия структуры научно-технического документа предъявляемым требованиям (свидетельство № 2013613411, 2013 г.).
6. Смирнов И.В., Девяткин Д.А., Тихомиров И.А., Швец А.В. Программа выявления связей между научно-техническими документами (свидетельство № 2013613409, 2013 г.).

Публикации в сборниках докладов российских и международных конференций:

7. Швец А.В. Формирование признакового пространства в задачах автоматического анализа научных текстов // Труды шестой международной конференции «Системный анализ и информационные технологии» (САИТ-2015). Светлогорск, 2015. – Т. 1. – С. 222-228.
8. Швец А.В. Метод автоматического выявления псевдонаучных публикаций // Теория и практика системного анализа: Труды III Всероссийской научной конференции молодых ученых с международным участием (ТПСА’14). – Рыбинск, 2014. – Т. 2. – С. 186-193.
9. Швец А.В. Экспериментальный метод автоматического определения уровня качества научных публикаций // Труды пятой международной конференции «Системный анализ и информационные технологии» (САИТ-2013). Красноярск, 2013. – Т. 1. – С. 304-312.

Личный вклад соискателя: в работах 1–9 автору принадлежат результаты, относящиеся к методам и алгоритмам выявления признаков, характеризующих качество текстов научной сферы.

Швец Александр Валерьевич (Россия)

ВЗАИМОДЕЙСТВИЕ ИНФОРМАЦИОННЫХ И ЛИНГВИСТИЧЕСКИХ МЕТОДОВ В ЗАДАЧАХ АНАЛИЗА КАЧЕСТВА НАУЧНЫХ ТЕКСТОВ

1. Разработан новый метод автоматического формирования общенаучного словаря устойчивых словосочетаний.
2. Разработан новый метод автоматического выявления структуры научной публикации.
3. Разработан новый метод обнаружения нарушений правил согласования, нарушений синтаксической и семантической связности, лексической избыточности, нарушений последовательности изложения.
4. Впервые разработан метод автоматического выявления псевдонаучных фрагментов текстов научной сферы.
5. Сформировано множество признаков, характеризующих качество текстов научной сферы.
6. Построено множество правил для обнаружения псевдонаучных текстов.

Shvets Alexander (Russia)

INTERACTION OF INFORMATIONAL AND LINGUISTIC METHODS IN PROBLEMS OF ANALYSIS OF QUALITY OF SCIENTIFIC TEXTS

1. A new method of automatic forming of general scientific vocabulary of set expressions has been developed.
2. A new method of automatic identification of structure of a scientific publication has been developed.
3. A new method of detection of violations of rules of agreement, violations of syntactic and semantic coherence, lexical redundancy, and detection of violations of text order has been developed.
4. For the first time a method of automatic detection of pseudoscientific fragments of texts of scientific area has been developed.
5. A set of features that characterize the quality of texts of scientific area has been formed.
6. A set of rules for detection of pseudoscientific texts has been built.