На правах рукописи

ПАЧКОВСКАЯ
Светлана Валерьевна

ФОРМИРОВАНИЕ КОНТЕНТА РЕФЕРАТА ПРИ АВТОМАТИЧЕСКОМ РЕФЕРИРОВАНИИ НАУЧНОГО ТЕКСТА

05.13.11 – Математическое и программное обеспечение вычислительных машин, комплексов и компьютерных сетей

АВТОРЕФЕРАТ
диссертации на соискание ученой степени
кандидата технических наук

Красноярск – 2010

Работа выполнена в Восточно-Сибирском государственном технологическом университете

Научный руководитель: доктор технических наук, доцент *Найханова Лариса Владимировна*

Официальные оппоненты: доктор технических наук, профессор *Доррер Георгий Алексеевич*

доктор технических наук, профессор *Ноженкова Людмила Фёдоровна*

Ведущая организация: ГОУ ВПО «Петрозаводский государственный университет» (г. Петрозаводск)

Защита диссертации состоится 12 марта 2010г. в 14-00 часов на заседании диссертационного совета ДМ 212.099.05 при Сибирском федеральном университете по адресу: 660074 г. Красноярск, ул. Киренского, 26, ауд. – УЛК-1-15.

С диссертацией можно ознакомиться в библиотеке Сибирского федерального университета по адресу: г. Красноярск, ул. Киренского, 26, ауд. – Г2-74.

Автореферат разослан 10 февраля 2010г.

Ученый секретарь
диссертационного совета
канд. техн. наук, проф. Е.А. Вейсов

ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

Актуальность темы. Анализ работ, посвященных автоматическому реферированию, показал, что их можно разделить на две категории.

К первой категории относятся работы В.А. Чижаковского, Э.М. Добрускиной, Р.Г. Пиотровского, Л.Н. Беляевой, О.А. Бородиной, Д.И. Блюменау, Н.И. Гендиной, Д.Г. Лахути, В.А. Яцко, в которых реферат формируется в соответствии с заранее заданной жесткой структурой. При этом для его формирования используются различного рода диагностирующие конструкции, не требующие глубокого семантического анализа исходного текста. Работы второй категории посвящены созданию реферата на основе компрессии исходного текста; исследования направлены на решение задачи понимания смысла текста, и их можно разбить на две группы.

В исследованиях первой группы, описанных в работах У. Хана, И. Мани, И.В. Николаевой, решение задачи основано на применении статистического анализа, института весовых коэффициентов и сопоставления фразовых шаблонов без применения глубокого семантического анализа. В исследованиях второй группы семантический анализ исходного текста требует не только грамматики и словари для морфологического и синтаксического разбора, но и онтологические тезаурусы, позволяющие делать логические выводы на основе временных, пространственных, каузальных и других законов. В работах А.З. Панкратовой, А.А. Харламова, А.Е. Ермакова, Д.М. Кузнецова для понимания смысла текста строятся семантические сети, а в работах А.М. Бледнова, А.В. Корхова, В.А. Тузова, О.В. Корховой выполняется более глубокий семантический анализ на основе метода семантических падежей К. Филмора и модели В.А. Тузова.

В основе многих систем автореферирования текста лежит статистический анализ текста, а для выявления смысла текста используется построение семантической сети исходного текста. Однако большинство разработок носят экспериментальный характер, не многие существующие системы поддерживают русский язык, генерируют сглаженные рефераты и предполагают опору на знания. Сервисы реферирования ориентированы либо на узкую предметную область, либо требуют участия человека, хотя можно выделить системы, в которых сделана попытка использования знаний на основе онтологического подхода, например для разрешения проблем омонимии.

Таким образом, необходимо отметить, что в настоящее время переход от экспериментальных лабораторных исследований по созданию качественных рефератов к их использованию в решении прикладных задач реального мира до сих пор не произошел. Поэтому проблема автоматического реферирования научного текста остается актуальной. Как правило, в автореферировании выделяются задачи формирования контента реферата и построения дискурса текста реферата.

Цель данной работы – разработка и исследование алгоритмов автоматического формирования контента реферата научного текста, позволяющих улучшить смысловое качество реферата и повысить эффективность процессов обработки данных и знаний в компьютерных системах и сетях.

Для достижения поставленной цели в работе решаются следующие задачи.

1. Обзор и анализ существующих решений в области автоматического реферирования текста.
2. Выбор и усовершенствование методов лингвистического анализа научного тек-

ста.

3. Разработка алгоритмов выбора предложений в реферат и алгоритма составления логической последовательности предложений в реферате.

4. Апробация разработанных моделей и алгоритмов.

Методы исследования. Методологической и теоретической основой выполненного исследования послужили положения теории искусственного интеллекта, логики предикатов первого порядка, нечеткой логики, ситуационного моделирования, теории автоматов и математической лингвистики.

Научная новизна. Научная новизна работы заключается в развитии методов автоматического построения реферата и состоит из следующих элементов.

1. Новизна разработанного алгоритма выбора предложений из анализируемого текста в реферат заключается в применении онтологической базы знаний, позволяющей сформировать совокупность предложений реферата, отражающих смысловой аспект анализируемого текста.

2. Новизна алгоритма формирования логической последовательности предложений в реферате заключается в том, что сформированный контент реферата является информативным и обладает достаточно точным изложением содержания документа.

Практическая ценность исследования состоит в том, что применение алгоритмов формирования контента реферата обеспечит повышение качества реферата в системе автоматического реферирования, применение которой в глобальной сети Интернет, библиотечных системах позволит усовершенствовать процессы обработки данных и знаний в компьютерных системах и сетях.

Публикации. Основные результаты диссертационной работы опубликованы в 8 печатных работах, из которых 7 статей и одно свидетельство об официальной регистрации программы для ЭВМ.

Апробация результатов исследования. Основные положения и результаты диссертационной работы докладывались и обсуждались на Всероссийской научно-технической конференции «Информационные системы и модели в научных исследованиях, промышленности и экологии» (Тула, 2007, 2009); Международной научно-технической мультиконференции «Актуальные проблемы информационно-компьютерных технологий, мехатроники и робототехники» (Таганрог, 2009); Всероссийской научно-практической конференции «Системы автоматизации в образовании, науке и производстве – AS’2009» (Новокузнецк, 2009); Всероссийской научно-технической конференции «Теоретические и прикладные вопросы современных информационных технологий» (Улан-Удэ, 2008-2009) и на ежегодных конференциях преподавателей, сотрудников и аспирантов ВСГТУ. Материалы диссертационных исследований используются в научных исследованиях Центра дистанционного образования Воронежского государственного технического университета при разработке интеллектуальных систем поддержки принятия решений в части построения семантической сети предметной области решаемой задачи; в учебной деятельности Восточно-Сибирского государственного технологического университета при разработке учебного курса «Естественно-языковые системы» по специальности 230105 – «Программное обеспечение вычислительной техники и автоматизированных систем» в виде курса лекций и методических указаний к лабораторным работам.

Структура и объем работы. Диссертация состоит из введения, четырех глав, за-

ключения, списка литературы и шести приложений, содержит 126 страниц текста, 25 рисунков и 24 таблицы. В список литературы вошло 131 наименование.

СОДЕРЖАНИЕ РАБОТЫ

Во введении обосновывается актуальность выбранной темы, определяются цель, задачи и методы исследования, излагаются научная новизна и практическая ценность полученных результатов.

В первом разделе дан обзор существующих методологий решения задачи автоматического реферирования научного текста и систем автоматического реферирования текста, рассмотрена классификация рефератов, проанализированы методы и системы автореферирования текста, описаны проблема и постановка задачи.

Исследованиями по автоматическому реферированию начали заниматься более 50 лет назад. К настоящему времени разработано достаточно много методов автореферирования, которые можно разделить на методы квазиреферирования и генерирования рефератов. Первые основаны на экстрагировании, т.е. выделении из текста наиболее информативных фрагментов, передающих основной смысл текста, вторые – на выделении наиболее существенной информации из текстов документов.

Теория и методика реферирования-экстрагирования были разработаны в конце 70-х – начале 80-х гг. группой исследователей Ленинградского института культуры: Д.И. Блюменау, Н.И. Гендиной, И.С. Добронравовым, Д.Г. Лахути и др. В рамках этой методики разработаны три вида методов: статистические, позиционные и индикаторные. Статистические методы основаны на использовании статистических параметров для оценки информативности различных элементов текста (слов, предложений), прежде всего, по частоте встречаемости слов в тексте; вес предложения определяется как сумма частот входящих в него значимых слов. Позиционные методы опираются на предположение о том, что информативность предложения находится в зависимости от его позиции в тексте документа. Индикаторные методы основаны на функциональной идентификации фраз первичного документа с помощью индексации их специальными словами – маркерами, индикаторами и коннекторами, образующими лексический аппарат теории экстрагирования.

Для реализации метода генерирования рефератов требуются мощные вычислительные ресурсы, грамматики и словари для синтаксического разбора и генерации естественно-языковых конструкций, онтологические справочники, отражающие соображения здравого смысла, и понятия, ориентированные на предметную область.

На сегодняшний день разработаны системы автореферирования текста, такие как промышленная система Newsblaster (Колумбийский университет, США), система Prosum (British Telecommunication Laboratories), инструмент для автоматического аннотирования документов МЛ Аннотатор (МедиаЛингва), система «Аналитический курьер», модуль Extractor, выделяющий из представленного ему на вход текста наиболее информативные именные группы, система TextAnalyst (Микросистемы) и целый ряд других. Кроме того, разработаны такие инструменты, как функция AutoSummarize в Microsoft Office, Inxight Summarizer (компонент поискового механизма AltaVista), системы IBM Intelligent Text Miner, Oracle Context. Большинство разработанных систем автоматического реферирования используют метод составления выдержек, т.е. выделяют и выбирают оригинальные фрагменты из исходного документа и соединяют их в корот-

кий текст.

Самым распространенным подходом в существующих системах является комбинированный, использующий «усиленные» статистические алгоритмы, которые предполагают нахождение различных частот слов и словосочетаний, таких как частота встречаемости в главе, начале или конце текста. Не многие существующие системы поддерживают русский язык, генерируют сглаженные рефераты и предполагают опору на знания. Сервисы реферирования ориентированы либо на узкую предметную область, либо требуют участия человека.

Результаты проведенного обзора показывают, что в большинстве методологий и готовых программных продуктов автореферирования текста проводится поверхностный семантический анализ. На выходе реферат представляется в виде несвязанных предложений, т.е. в виде выдержек или тезисов текста.

В связи с этим необходимо разработать семантический анализ, который позволит более глубоко понимать смысл текста. На наш взгляд, это возможно только при применении онтологического подхода к семантическому анализу, и в работе поставлена следующая задача.

Пусть заданы исходный научный текст $T = < t_1, t_2,...,t_n >$, состоящий из последовательности предложений; V_T – размер текста T в символах, включая пробелы и другие специальные знаки; O – онтологический тезаурус по предметной области текста T, Г – лингвистическое обеспечение, p – требуемое процентное сжатие текста, ε – погрешность размера полученного текста.

Требуется построить производный текст (реферат повествовательного типа) $R = < r_1, r_2,...,r_k >$, состоящий из последовательности предложений, адекватно передающий смысл текста T без потерь основных информационных единиц и удовлетворяющий заданным значениям p и ε, при следующих ограничениях:

1) $p \in [5,30]$; 2) $V_R \in \left[p(1-\varepsilon)V_T, p(1+\varepsilon)V_T \right]$; 3) $\varepsilon \le 5\%$.

Во втором разделе рассматриваются широко известные методы лингвистического анализа текста, применяемые в работе, приведена обобщенная схема решения задачи автоматического реферирования.

Для выполнения автоматического реферирования научного текста будем считать, что выполнена его предварительная обработка и известны исходные данные в виде лексем с морфологической информацией и графов зависимостей предложений текста.

С целью повышения статистических характеристик терминов выполняется анализ словосочетаний исходного текста с применением онтологии предметной области. Для выделения словосочетаний построен конечный автомат, который осуществляет поиск и формирование набора словосочетаний в разрезе различных моделей. Конечный автомат включает следующие основные групповые состояния: определение и исключение абстрактных прилагательных из словосочетаний; разделение композиционных словосочетаний на простые; определение синонимов термов; замена термов с низкой частотой встречаемости на термы-синонимы с максимальной частотой встречаемости; перерасчет частот встречаемости термов. После этого строится семантическая сеть текста в виде взвешенного графа:

$$S = (V, U, W), \qquad (1)$$

где V – множество вершин, каждой из которых соответствует граф G^F семантической окрестности некоторого термина;

W – множество весов дуг $u \in U$, отражающих семантическую близость вершин.

Построение графов семантической окрестности. Пусть имеем сформированное множество именных словосочетаний (термов) $E = \{e_1, e_2, .., e_m\}$. Граф G^F семантической окрестности термина t^η является древовидным представлением класса эквивалентности K^E по отношению R общности терминов множества E относительно термина $t^\eta \in E$, который назовем несущим словом, т.е.

$$K^E = \{x \mid (t^\eta, x) \in R, x \in E\}. \quad (2)$$

Несущее слово t^η будет располагаться в корневой вершине графа G^F. Таким образом, граф G^F семантической окрестности имеет вид:

$$G^F = \left(V^F, U^F, W^F\right), \quad (3)$$

где V^F – множество вершин, каждой из которых приписаны термины (лексемы или словосочетания) из одного класса эквивалентности;

U^F – множество дуг графа;

W^F – множество весов вершин графа.

Вес графа G^F определяется по формуле:

$$w\left(K^E\right) = \sum_{x \in K^E} w(x), \quad (4)$$

где $w(x)$ – вес каждого терма, входящего в класс эквивалентности:

$$w(x) = \frac{\sum_{x \in K^E} f(x)}{\max\limits_{x \in \tilde{T}} \{f(x)\}} + \frac{f(x)}{\sum_{x \in K^E} f(x)}. \quad (5)$$

В формуле (5) $f(x)$ – частота встречаемости терма x в рассматриваемом тексте.

На рисунке 1 представлен граф семантической окрестности термина «система». Между смежными вершинами графа установлены родовидовые отношения. В корневой вершине располагается несущее слово «система». В смежных с корнем вершинах располагаются словосочетания, зависимые от несущего слова.

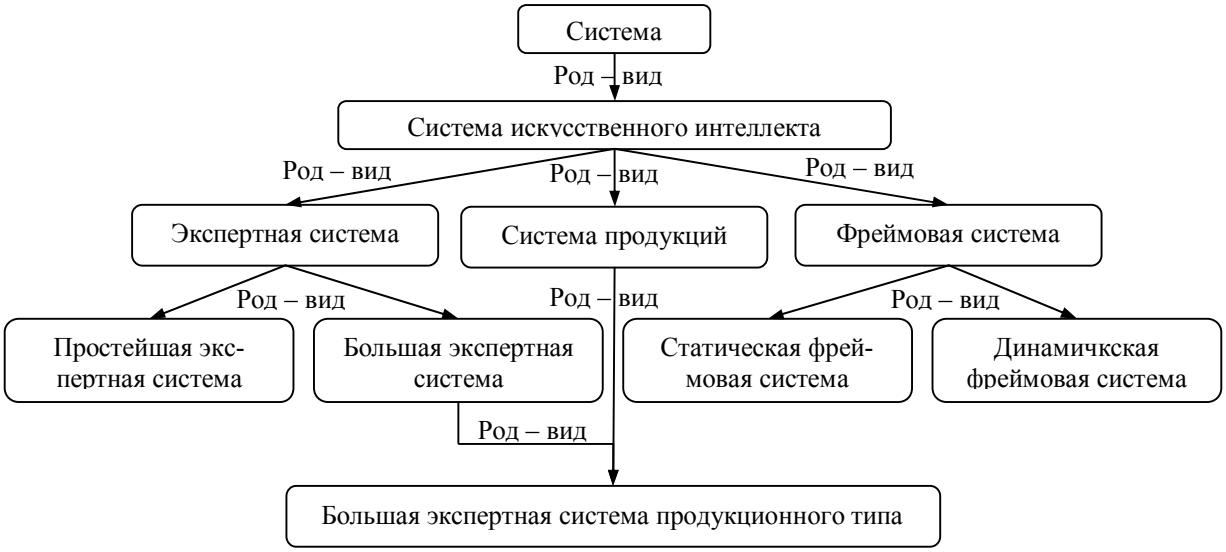

Рисунок 1 – Граф семантической окрестности термина «система»

Таким образом, значимые графы семантической окрестности, обладающие наибольшими весовыми коэффициентами, почти всегда определяют термины, речь о которых идет в анализируемом тексте. Кроме того, достоинство графа семантической окрестности в том, что при его построении определяются родовидовые отношения между терминами. Это важно для определения семантики текста.

Построение модифицированного графа зависимостей. Модифицированный граф G^M зависимостей является ориентированным графом, построенным на основе термов (и описывающих их векторов), лексем (и векторов их морфологической информации) одного предложения. Граф имеет вид:

$$G^M = \left(V^M, U^M\right), \qquad (6)$$

где V^M – множество вершин, выраженных термами и лексемами предложения соответственно;

U^M – множество дуг, определяющих семантическое отношение между термами, находящимися в вершинах графа.

На рисунке 2 приведен пример модифицированного графа G_A^G зависимостей предложения А = «Данные в процессах компьютерной обработки проходят преобразование от исходной формы данных до базы данных на машинных носителях информации через представления данных на машинных языках».

Термины, выраженные отглагольными существительными, в графе заменены соответствующими глаголами. В вершинах графа зависимостей располагаются термы-словосочетания или термы-лексемы. Из примера видно, что между вершинами, содержащими термины «данные» и «процессы компьютерной обработки» имеется вершина, содержащая глагол «проходить», между вершинами с терминами «исходная форма данных» и «базы данных» – вершина с глаголом «преобразовать» и т.д.

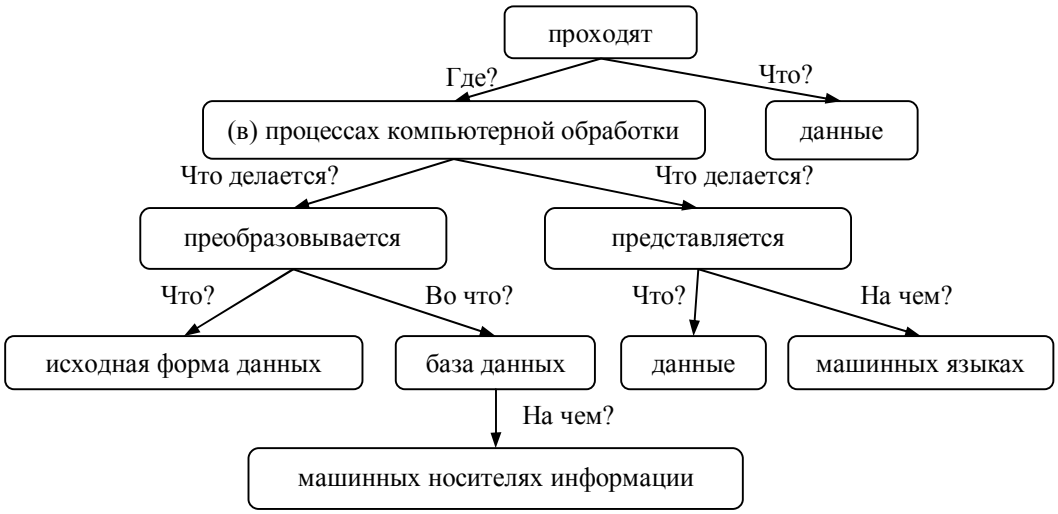

Рисунок 2 – Пример модифицированного графа зависимостей предложения

Таким образом, можно сделать вывод, что в модифицированном графе зависимостей G^M каждая пара термов, представляющих собой устойчивые именные словосочетания, определяется семантическим отношением, выраженным глаголом или глагольной группой.

Соединение графов семантической окрестности. Связи между графами семантической окрестности устанавливаются на основе анализа модифицированных графов зависимостей и определения семантических отношений между наиболее значимыми тер-

минами графов.

Опишем алгоритм построения семантической сети S научного текста на основе соединения графов семантической окрестности. Из множества графов семантической окрестности выделим подмножество графов $\{G^F\}$, веса которых больше среднего веса всех построенных графов семантической окрестности.

В двух различных графах семантической окрестности $G_s^F, G_t^F \in \{G^F\}$, $s \neq t$, $s,t \in |\{G^F\}|$ рассмотрим последовательно все пары $(v_s^F, v_t^F) \in V_s^F \times V_t^F$, исключая случай, когда $v_s^F = v_t^F$. Если среди них встречается пара (v_s^F, v_t^F), такая что $(v_s^F, v_t^F) = (v_i^M, v_{i+2}^M)$ или $(v_s^F, v_t^F) = (v_{i+2}^M, v_i^M)$, где $v_i^M, v_{i+2}^M \in V^M$ – вершины некоторого модифицированного графа G^M, то соединим вершины v_s^F и v_t^F дугой с пометой, имеющей значение вершины v_{i+1}^M. Вершины v_i^M и v_{i+2}^M содержат термины, между которыми имеется семантическое отношение, определенное термином вершины v_{i+1}^M.

Практика показала, что между вершинами семантической сети *S* могут быть кратные дуги, которым приписаны различные семантические отношения. При слиянии кратных дуг вес результирующей дуги будет равен количеству кратных дуг. Помета такой дуги будет содержать множество семантических отношений. В качестве основной связи между этими графами выберем связь с наибольшим весом, т.е. $w^G = max\{p^{G^F}\}$.

В том случае, если остались изолированные графы семантической окрестности, то для установления связей используется онтология. В ней производится поиск фреймов, содержащих термины, идентифицирующие вершины v_s^F и v_t^F графов $G_s^F, G_t^F \in \{G^F\}$. В семантической сети знаков-фреймов онтологии осуществляется поиск пути между этими фреймами. Как правило, связь представляет собой цепочку фреймов. Вес этой связи будет зависеть от длины цепочки.

Таким образом, будет построена взвешенная семантическая сеть текста, в узлах которой располагаются графы семантической окрестности терминов. Дуги сети, помеченные глаголами, отражают отношения между терминами. Фрагмент семантической сети текста приведен на рисунке 3. Такая семантическая сеть отражает основной смысл научного текста.

Третий раздел посвящен формированию контента реферата. Для этого вначале осуществляется выбор предложений-кандидатов из исходного текста в производный, а затем выстраивается их логическая последовательность. Рассмотрим первый алгоритм.

Первый алгоритм выбора предложений. Идея данного алгоритма заключается в том, что граф G^F, являющийся вершиной семантической сети *S* и имеющий наибольший вес, описывает семантическую окрестность термина, о котором идет речь в научном тексте. С этой точки зрения мы предположили, что так как в научном тексте, как правило, идет речь не об одном термине, а о нескольких, то, произведя поиск и анализ большевесных графов G^F, можно найти предложения, которые отражают смысл первичного текста.

Рисунок 3 – Фрагмент семантической сети текста

Алгоритм начинается с выбора в семантической сети *S* графа G^F с наибольшим весом, в котором отмечается вершина v_m с максимальным весом $\max\limits_{i\in|V^F|} w_i^F$. Среди всех дуг, инцидентных вершине v_m, выберем дугу u_k, обладающую максимальным весом: $w_k = \max\limits_{j\in|W^F|} w_j^F$. Если таких дуг несколько, необходимо анализировать вес второй концевой вершины v_l. Дугу (v_m, v_l) помечаем для ее исключения из дальнейшего анализа. Процесс повторяется для вершины v_l и т.д. В случае, если в процессе анализа текущая вершина не будет иметь исходящих дуг в вершины других графов G^F, анализ продолжается с вершины, обладающей наибольшим весом среди нерассмотренных ранее вершин.

Таким образом, семантическая сеть реферата строится как проекция семантической сети текста, которая однозначно определяет выбор предложений в реферат.

К примеру, пусть в семантической сети выбран граф «Знания» с наибольшим весом, в котором вершина «База знаний» имеет наибольший вес. Среди дуг, инцидентных данной вершине, выбрана дуга с наибольшим весом. Данная дуга связывает вершину «База знаний» с вершиной «Интеллектуальная система». Предложения, соответствующие данной связи, включаем в реферат. Далее рассматриваем дуги, связывающие вершину «Интеллектуальная система» с другими вершинами сети, и повторяем описанные выше действия.

Достоинство приведенного алгоритма заключается в том, что совокупность выбранных предложений действительно передает смысл исходного текста. Однако вторичный текст, являющийся проекцией исходного текста, обладает низкой связностью предложений и больше похож на совокупность основных выдержек из текста. Это послужило основанием для разработки другого алгоритма.

Второй алгоритм выбора предложений. В данном алгоритме реализуется следующая идея. Необходимо найти термины, о которых идет речь в первичном тексте, и

построить пути для них в сети *S*; затем после подсчета весов всех путей выбрать путь с наибольшим весом. Тогда полученный в результате этого процесса вторичный текст должен быть связным.

Отсортируем по убыванию веса W^F графов G^F. Найдем $\Delta_i \quad W_i^F - W_{i+1}^F$ и среди полученных Δ_i их среднее значение $\Delta_{avg} \quad = avg_i \Delta_i$, i – индекс веса в отсортированном списке. Для анализа выберем графы G^F, для которых выполняется условие $\Delta_i \le \Delta_{avg}$. В каждом выбранном графе для вершины $v_m \in V^F$, обладающей максимальным весом, найдем путь H наибольшей длины по сети *S*. Для каждого такого пути H вычислим ее вес P по формуле:

$$P = \sum_{i=1}^{n} w_i^H + \sum_{k=1}^{m} w_k^H , \qquad (7)$$

где $w_i^H = w_i^H / \sum_{i=1}^{n} w_i^H$ – нормированные веса вершин, входящих в путь H;

$w_k^H = w_i^H / \sum_{i=1}^{m} w_i^H$ – нормированные веса дуг, входящих в путь H.

Из множества найденных путей выберем путь с наибольшим весом *P*. Предложения, соответствующие выбранному пути, включим в реферат.

Совокупность предложений, полученная по второму алгоритму, отличается от первой. В данной совокупности наблюдается тема-рематическая цепочка предложений, отражающая тематические отношения в тексте. Но полученный реферат содержит избыточные предложения и по наполнению далек от реферата, составленного экспертом, вот почему был предложен третий алгоритм.

Третий алгоритм выбора предложений. Для устранения недостатков предыдущего алгоритма было решено ввести коэффициент значимости предложений, который позволит удалить избыточные предложения. Данный алгоритм почти полностью совпадает с предыдущим алгоритмом, отличие заключается только в процедуре выбора предложений во вторичный текст. Будем считать, что найден путь *H*. Отрезок, представленный двумя вершинами и дугой, однозначно определяет предложение исходного текста. Поэтому введем критерий значимости *k*-го предложения, который может быть вычислен по формуле:

$$kr_k^H = \frac{w_i^H + w_{i+1}^H}{\sum_{j=1}^{n} w_j^H} + \frac{w_{i,i+1}^H}{\sum_{j=1}^{m} w_j^H} \qquad (8)$$

Определим среднее значение вычисленных критериев предложений, соответствующих пути *H*:

$$kr_{avg} = \frac{\sum_{k=1}^{m} kr_k}{m} .$$

Тогда в реферат включим предложения, коэффициент значимости которых $kr_k \ge kr_{avg}$. Этот алгоритм позволяет удалить избыточные предложения. Кроме того, изменяя kr_{avg}, можно уменьшать или увеличивать объем вторичного текста. Для этого

нужно вычислить процент сжатия первичного текста p_1 как отношение количества предложений вторичного текста, полученного при использовании kr_{avg}, к количеству предложений первичного текста.

Коэффициенты уменьшения kr'_{avg} и увеличения kr''_{avg} объема вторичного текста можно найти по следующим формулам:

$$kr'_{avg} = \frac{kr_{avg} \cdot p_1}{p}; \qquad kr''_{avg} = \frac{kr_{avg} \cdot p}{p_1}, \tag{9, 10}$$

где p – заданный процент сжатия текста.

Полученный реферат обладает смысловой цельностью и отличается по составу от текстов, составленных с применением предыдущих алгоритмов. Реферат включает предложения, описывающие основные темы, затронутые в исходном тексте. Достоинством данного алгоритма является то, что его объем может быть легко изменен подбором значения коэффициента p_1.

Однако полученный реферат все еще далек от реферата, полученного экспертом в данной предметной области. В связи с этим предложен четвертый алгоритм выбора предложений в реферат.

Четвертый алгоритм выбора предложений. Алгоритм основан на использовании онтологии предметной области, представленной в виде семантической сети знаков-фреймов. В знаке-фрейме сконцентрированы знания о некотором термине. Сопоставляя значения слотов знака-фрейма с содержимым вершин семантической сети анализируемого научного текста, можно выбрать предложения из первичного текста, содержащие знания о текущем термине.

В семантической сети S выбираем граф G^F с наибольшим весом. В онтологии O найдем знак-фрейм Ф, в котором описан термин t, расположенный в корне графа G^F. Последовательно просматриваются значения слотов знака-фрейма. Если в семантической сети S найдено значение слота (термин t), то предложение, содержащее этот термин, включается в реферат.

Так, в построенной сети S выбран граф G^F с вершиной в корне «Знания», имеющей наибольший вес. Затем в онтологии найден соответствующий данному понятию знак-фрейм, представленный в таблице 1.

Таблица 1 – Упрощенное представление слотов знака-фрейма понятия «Знания»

НАЗВАНИЕ СЛОТА	ЗНАЧЕНИЕ СЛОТА
ИМЯ ТЕРМИНА	ЗНАНИЯ
ДЕФИНИЦИЯ	СОВОКУПНОСТЬ СВЕДЕНИЙ, ОБРАЗУЮЩИХ ЦЕЛОСТНОЕ ОПИСАНИЕ, СООТВЕТСТВУЮЩЕЕ НЕКОТОРОМУ УРОВНЮ ОСВЕДОМЛЕННОСТИ ОБ ОПИСЫВАЕМОМ ВОПРОСЕ, ПРЕДМЕТЕ, ПРОБЛЕМЕ И Т.Д.
СИНОНИМЫ	МЕТАДАННЫЕ
РОД	МЕТАЗНАНИЯ
ВИД	ДЕКЛАРАТИВНЫЕ ЗНАНИЯ ПРОЦЕДУРНЫЕ ЗНАНИЯ
	ЗАКОНОМЕРНОСТИ ФАКТЫ
ЦЕЛОЕ	БАЗА ЗНАНИЙ
ЧАСТЬ	ДАННЫЕ
ДЕЙСТВИЯ	ПРЕДСТАВЛЕНИЕ ЗНАНИЙ
…	…

Слот «Род» имеет значение «Метазнания», такой вершины в сети S не найдено, переходим к следующему слоту. Слот «Вид» имеет два значения: «Декларативные зна-

ния» и «Процедурные знания». В сети S найдены вершины, содержащие данные термины, и имеющие дуги, инцидентные вершине, которая содержит термин «Знание». Поэтому соответствующие предложения включаются в реферат. Далее последовательно друг за другом раскрываются термины, находящиеся в смежных вершинах: «Декларативные знания», «Процедурные знания», «База знаний» и другие. После этого рассматривается следующий граф G^F с наибольшим весом среди оставшихся. Процесс повторяется.

При помещении предложений во вторичный текст постоянно идет проверка на достижимость заданного значения p. При его достижении процесс прекращается.

Построенные по предложенным алгоритмам рефераты сравнивались с рефератом, построенным экспертом. Реферат, построенный по четвертому алгоритму, наиболее близок к реферату эксперта и обладает лучшими характеристиками качества.

Алгоритм построения логической последовательности предложений. Данный алгоритм также базируется на использовании онтологии предметной области и применяется к выбранной совокупности предложений.

Знаки-фреймы, с помощью которых происходил процесс выбора предложений во вторичный текст, по определению представляют собой иерархическую сеть. Это свойство положено в основу данного алгоритма и означает, что сначала необходимо выстроить иерархию терминов, затем в соответствии с ней определить последовательность предложений в реферате. Так, например, в интеллектуальную систему входит база знаний, база знаний состоит из знаний, и т.д. База знаний является частью интеллектуальной системы, поэтому ссылка на знак-фрейм термина «База знаний» будет записана в соответствующем слоте знака-фрейма термина «Интеллектуальная система». Знания являются частью базы знаний, поэтому ссылка на знак-фрейм термина «Знания» будет записана в слоте знака-фрейма термина «База знаний» и т.д. Тогда в реферат сначала будут выбраны предложения, содержащие термин «Интеллектуальная система», за ними последуют предложения с термином «База знаний», затем – предложения с термином «Знания» и т.д.

Таким образом, все предложения во вторичном тексте будут выстроены в логической последовательности. На наш взгляд, реферат, построенный с помощью предложенного алгоритма, получается связным и осмысленным.

В четвертом разделе приведены описание программного обеспечения и результаты вычислительных экспериментов.

Разработанная система JASS (Java Automatic Summarize System) осуществляет морфологический, синтаксический и семантический анализ естественно-языковых текстов, строит семантическую сеть текста и формирует его реферат. Для разработки программного обеспечения использовались объектно-ориентированный язык программирования JAVA, среда разработки программного обеспечения Eclipse IDE, фреймворк для визуализации графов JUNG (Java Universal Network/Graph Framework).

Для апробирования предложенных в работе алгоритмов рассматривались разные по виду, объему, содержанию научные тексты: монографии, диссертации, отчеты о НИР, учебно-методические пособия, конспекты лекций, объемы которых составляют минимум 10 страниц и могут превысить 100 страниц. Подготовленные тексты принадлежат следующим предметным областям: «Искусственный интеллект», «Информатика» и «Экономика», так как для данных предметных областей имеются построенные онто-

логии. Объем онтологии по искусственному интеллекту составляет 550 терминов, по информатике – 2500, по экономике – 1200 терминов.

В работе для оценки качественных показателей рефератов использован метод экспертной оценки. В качестве оцениваемых альтернатив экспертам было предложено множество рефератов текста, полученных с применением четырех алгоритмов. При этом в качестве критериев оценки альтернатив предлагались: связность – правильность следования предложений в тексте; осмысленность – выбор предложений, несущих основную смысловую нагрузку; полнота – полнота охвата всех разделов текста. Для оценки альтернатив экспертам была предложена лингвистическая шкала измерений.

Оценка *i*-й альтернативы *j*-м экспертом производилась по формуле:

$$v'_{ij} = 1 - \frac{(l-1)}{k}, \qquad (11)$$

где l – индекс значения лингвистической шкалы;

k – количество значений этой шкалы.

Для оценки *i*-й альтернативы всеми n экспертами используется формула:

$$s_i = \sum_{j=1}^{n} v'_{ij} \qquad (12)$$

Результаты экспертной оценки характеристик полученных рефератов, вычисленные по этим формулам, приведены в таблице 2.

Таблица 2 – Экспертная оценка характеристик качества рефератов

Характеристики качества	Алгоритмы			
	Первый	Второй	Третий	Четвертый
Связность текста	0,00	0,27	0,45	0,82
Осмысленность текста	0,18	0,18	0,82	1,00
Полнота текста	0,18	0,36	0,91	1,00
Среднее значение оценок	0,12	0,27	0,73	0,94

Как видно из таблицы, третий и четвертый алгоритмы показали лучшие характеристики. Эти алгоритмы подверглись вычислительным экспериментам, в которых задавались различные коэффициенты сжатия p исходного текста. Например, для p=11% реферат, полученный с использованием критерия значимости (третий алгоритм), дает 46% совпадения с рефератом, построенным экспертом, а реферат, полученный с использованием онтологии предметной области (четвертый алгоритм), – 85%. Это говорит о том, что четвертый алгоритм обладает лучшими характеристиками.

Таким образом, результаты описанных вычислительных экспериментов позволяют сделать вывод о корректности предложенных в работе алгоритмов формирования контента реферата, об адекватности смысла построенных рефератов; и можно сделать вывод, что для улучшения логической связности и информативности автореферата необходимо использовать онтологию анализируемой предметной области.

В **приложениях** приведены фрагмент исходного текста; рефераты, полученные с использованием предложенных алгоритмов с различным процентом сжатия; описание программного обеспечения; результаты вычислительных экспериментов.

ЗАКЛЮЧЕНИЕ

Основным результатом проведенных исследований является совершенствование методов автоматического формирования контента реферата научного текста, которое помогло улучшить смысловое качество реферата, что, в свою очередь, позволит повы-

сить эффективность процессов обработки данных и знаний в компьютерных системах и сетях.

Научные и практические результаты работы состоят в следующем.

1. Усовершенствован способ построения семантической сети текста путем соединения графов семантической окрестности посредством анализа модифицированных графов зависимостей и онтологии предметной области.

2. Разработана автоматная модель поиска словосочетаний различных моделей.

3. Разработаны и исследованы алгоритмы выбора предложений из анализируемого текста в реферат на основе семантического анализа.

4. Разработан алгоритм формирования логической последовательности предложений в реферате с использованием онтологии и графов семантической окрестности понятий.

5. Разработано программное обеспечение для апробации предложенных алгоритмов.

Результаты работы отражены в следующих публикациях.

Публикации в изданиях по перечню ВАК:

1. Машанова С.В. (Пачковская С.В.) Построение семантической сети текста в задаче автоматического реферирования / С.В. Машанова (С.В. Пачковская), С.Д. Данилова // Системы управления и информационные технологии: науч.-техн. журн. – М.; Воронеж: Научная книга, 2009 – №1.3(35). – С. 383-386.

Основные публикации:

2. Машанова С.В. (Пачковская С.В.) Технология автоматического реферирования текста / Л.В. Найханова, С.В. Машанова (С.В. Пачковская) // Информационные системы и модели в научных исследованиях, промышленности и экологии: мат-лы всерос. науч-техн. конф. – М.; Тула: Изд-во ТулГУ, 2007. – С.69-70.

3. Машанова С.В. (Пачковская С.В.) Автоматическое реферирование научного текста на основе использования онтологического тезауруса / Л.В. Найханова, С.В. Машанова (С.В. Пачковская) // Теоретические и прикладные вопросы современных информационных технологий: мат-лы всерос. науч.-техн. конф., – Улан-Удэ: Изд-во ВСГТУ, 2008. – С. 130-134.

4. Машанова С.В. (Пачковская С.В.) Соединение графов семантической окрестности / С.В. Машанова (С.В. Пачковская) // Искусственный интеллект. Интеллектуальные системы: мат-лы X междунар. науч.-техн. конф. – Таганрог: Изд-во ТТИ ЮФУ, 2009. – С. 282-285.

5. Машанова С.В. (Пачковская С.В.) Повышение коэффициентов значимости словосочетаний / С.В. Машанова (С.В. Пачковская) // Информационные системы и модели в научных исследованиях, промышленности и экологии: мат-лы V всерос. науч-техн. конф. – М.; Тула: Инновационные технологии, 2009. – С.11-14.

6. Машанова С.В. (Пачковская С.В.) Автоматная модель поиска словосочетаний в научном тексте / С.В. Машанова (С.В. Пачковская), О.Г. Шарагина, Л.К. Мадаева // Теоретические и прикладные вопросы современных информационных технологий: мат-лы всерос. науч.-техн. конф.: в 2 ч. – Улан-Удэ: Изд-во ВСГТУ, 2009. – Ч.II. – С. 343-348.

7. Машанова С.В. (Пачковская С.В.) Диагностирующие выражения для выявления в тексте маркеров, индикаторов и коннекторов / С.В. Машанова (С.В. Пачковская) // Системы автоматизации в образовании, науке и производстве: тр. VII всероссийской науч.-практ. конф. – Новокузнецк: Изд-во СибГИУ, 2009. – С. 429-433.

8. Машанова С.В. (Пачковская С.В.) Свидетельство о государственной регистрации программы для ЭВМ №2009615123. «Программа построения семантической сети текста для задачи автоматического реферирования» / С.В. Машанова (С.В. Пачковская), А.И. Ильинчик. – М.: Федеральная служба по интеллектуальной собственности, патентам и товарным знакам, 2009.

Пачковская Светлана Валерьевна
Формирование контента реферата при автоматическом реферировании научного текста
Автореф. дис. на соискание учёной степени кандидата технических наук.
Подписано в печать 8.02.2010г.
Формат 60×84 1/16. Усл.печ. л. 1,27. Тираж 100 экз.
Заказ № 23.

Издательство ВСГТУ 670013 г.Улан-Удэ, ул. Ключевская 40в.