И. В. Ковалев

МУЛЬТИЛИНГВИСТИЧЕСКИЙ АНАЛИЗ УНИКАЛЬНОСТИ ТЕКСТОВ НА БАЗЕ ЛЕКСИЧЕСКИ СВЯЗАННЫХ ИНФОРМАЦИОННЫХ КОМПОНЕНТОВ

Монография

Красноярск
2020

УДК 004
ББК 32.97
К 56

Рецензенты:

д-р техн. наук, проф. *С.В. Ченцов*
д-р филол. наук, проф. *Г. Д. Воскобойник*

Ковалев, И. В.
К56 Мультилингвистический анализ уникальности текстов на базе лексически связанных информационных компонентов: монография / И. В. Ковалев ; ОУ «ККДНиТ». - Красноярск, 2020. - 144 с.

ISBN 978-5-6045165-0-8

Монография посвящена компьютерному моделированию и разработке алгоритмических и информационных средств поддержки мультилингвистического анализа уникальности текстов на базе лексически связанных компонентов с формированием межязыковых/внутриязыковых ассоциативных связей. Для внедрения предложенной методики в процесс мультилингвистического анализа уникальности текстов разработана структура частотных лексически связанных словарей как средства информационно-терминологического обеспечения систем поддержки принятия решений.

Монография предназначена для специалистов в области информационных технологий, научных работников, а также для аспирантов и магистрантов, обучающихся по направлениям «Информатика и вычислительная техника», «Программная инженерия».

УДК 004
ISBN 978-5-6045165-0-8

ОГЛАВЛЕНИЕ

ВВЕДЕНИЕ

Развитие современной науки и технологий, включая и технологии производства, и информационные технологии, неизменно влечет за собой появление новых научных текстов описательного свойства [1-22]. Научные статьи, монографии, технические документации, учебники, самоучители – это далеко не весь спектр литературы и интернет источников, сопровождающих развитие новых научных направлений и нового технологического оборудования, программных продуктов, а так же тех или иных изменений в информационных и технологических процессах [23-45].

Огромные объемы и интенсивность появления новых научных и специальных текстов, как правило, на разных языках (зависит от страны-производителя, места проживания авторов и т.п.) и необходимость коммуникации с зарубежными партнерами как в производственной, так и научной сфере, вынуждают часто выполнять анализ текста на уникальность (в последние годы происходит ужесточение требований к корректному авторству и уникальности текстов, предъявляемых как академическим сообществом, так и интернет пользователями и производителями интернет-ресурсов, практически во всех областях знаний) [46].

И если требования к профессиональным навыкам ученых и специалистов, причастным к публикационной активности, их знанию и опыту остаются неизменно высокими, то требования к уникальности публикуемых текстов на родном или любом иностранном языке с каждым годом все возрастают. При этом от специалиста не требуется глубокого знания того или иного иностранного языка, требования, как правило, ограничиваются предметной областью профессиональной деятельности [47-56].

В рамках решения этой проблемы при переводе текста на иностранные языки во многих случаях должна учитываться профессиональная ориентация текста. Одним из основных моментов в данном случае является использование характерной для научной/учебной специальности, исследовательского направления и ряда новых технологий [57].

Современное развитие компьютерного моделирования и информационных технологий на сегодняшний день позволяет разработать эффективный программно-алгоритмический аппарат для создания компьютерных интерактивных систем мультилингвистического анализа уникальности текстов, в том числе, при использовании разноязычной иностранной терминологической

лексики [58-69]. При этом используется современный подход в рамках мультилингвистической технологии, базирующийся на лексически связанных компонентах как внутри одного языкового поля, так и на уровне межязыкового взаимодействия [70].

Лежащая в основе диссертационного исследования, мультилингвистическая адаптивно-обучающая технология (МЛ-технология) разработана специально для создания подобных систем. Она базируется на адаптивной модели обучаемого Л.А. Растригина и электронных мультилингвистических частотных словарях. К ее достоинствам, с точки зрения анализа уникальности текстов, можно отнести возможность генерации ассоциативных полей между терминами-аналогами различных языков [71, 72]. Однако МЛ-технология, как и многие другие, созданные для обучения [73-87], не использует в полной мере лексические зависимости в рамках одного или многих языков, поэтому генерация внутриязыковых ассоциативных полей происходит исключительно на этапе применения знаний на практике и имеет стихийный характер. Если же такая генерация будет происходить организованно и непосредственно в процессе реального времени, качество такого процесса, а, следовательно, и компьютерных систем в значительной степени возрастет.

Описанные проблемы в сфере науки и образования, в информационных технологиях определи актуальность работы, заключающуюся в необходимости компьютерного моделирования и построения методики процесса мультилингвистического анализа уникальности текстов на базе лексически связанных компонентов, разработки соответствующего модельно-алгоритмического аппарата и информационной поддержки, что и дает простор для дальнейшего исследования.

Объектом исследования являются компьютерные модели и системы анализа уникальности текстов, структуры данных информационной поддержки систем и сопутствующие им алгоритмы.

Предмет исследования – это модели и алгоритмы, реализующие процесс мультилингвистического анализа уникальности текстов на базе лексически связанных компонентов, базирующихся на модели данных и алгоритмах обработки текстов с помощью многоязычных частотных словарей.

Целью исследования является компьютерное моделирование и разработка алгоритмических и информационных средств поддержки мультилингвистического анализа уникальности текстов на базе лексически связанных компонентов с формированием межязыковых/внутриязыковых ассоциативных связей.

В монографии рассмотрены следующие задачи, обусловленные целью исследования: теоретико-информационный анализ структуры информационной базы при компьютерном моделировании процесса мультилингвистического анализа уникальности текстов; разработка алгоритмических и информационных средств поддержки мультилингвистического анализа уникальности текстов на базе лексически связанных компонентов с формированием межязыковых/внутриязыковых ассоциативных связей; разработка структуры частотных лексически связанных словарей как средства информационной поддержки мультилингвистического анализа уникальности текстов; разработка алгоритмов структурно-параметрического синтеза информационно-терминологического базиса как совокупности лексически связанных компонентов; формирование информационно-терминологического базиса предложенной структуры и реализация его в виде электронного двухблочного мультилингвистического лексически связанного словаря на основе разработанных алгоритмов и других компьютерных методов обработки информации.

Следует отметить, что для современных информационных и управляющих систем характерен значительный рост объемов многоязычной (мультилингвистической) информации, хранящейся в банках данных предприятий и корпораций. В мультинациональных корпорациях требуется своевременное предоставление управленческому персоналу документов, принадлежащих различным языковым множествам, в процессе подготовки и принятия решения. Значительно возрастают требования к оперативности и качеству функционирования систем поиска мультилингвитсической информации, используемых лицом, принимающим решения (ЛПР) в различных экономических и производственных ситуациях.

Предлагаемые ранее методы решения поставленных задач зачастую оказываются недостаточно эффективными, так как не позволяют осуществить поиск и сортировку больших объемов информации, характеризующей сложившиеся производственные ситуации, не обеспечивают возможность мультилингвистического поиска данных принятия решения в информационно-управляющих системах (ИУС).

Указанное обстоятельство требует разработки новых эффективных моделей и алгоритмов поиска мультилингвистической информации, а также анализа уникальности многоязычных текстов на базе лексически-связанных компонентов.

1 Компьютерная модель процесса анализа текста на базе лексически связанных компонентов

1.1 Электронные частотные профессионально-ориентированные словари

Целью применения компьютерного моделирования процесса мультилингвистического анализа уникальности текстов является выявление идентичности текстовых фрагментов специализированного, в том числе и иностранного текста как на одном, так и на нескольких иностранных языках. Это может выполняться как для учебных, научных, так и для других профессиональных целей. Основными компонентами средств поддержки данного процесса являются электронные частотные словари (ЭЧС), построенные по мультилингвистическому принципу, и компьютерные системы, реализующие анализ уникальности текстов на базе специализированной терминологической лексики [68, 88-95].

ЭЧС – это именно та информация, порции которой используются для установления идентичности текстовых фрагментов, а частоты позволяют находить наиболее употребительные аналоги в других языках, то есть те слова и словосочетания, которые чаще всего встречаются в текстах узкой предметной области. Это и позволяет при проверке уникальности текста на одном языке, выявлять аналоги на другом языке, и учитывая частотные свойства текстов этого языка, делать вывод о степени уникальности текста в том случае, если он будет переведен на этот язык.

В таблице 1.1 приведен пример представления информации в частотном словаре.

Таблица 1.1 – Фрагмент частотного словаря

combined method, 49	комбинированный метод
research, 1020	исследование
mathematical analysis, 177	математический анализ
expert approach, 110	экспертный подход

В мультилингвистических частотных словарях учитываются частотные свойства многоязычных терминов [81]. Пример фрагмента мультилингвистического словаря по системному анализу можно наблюдать в таблице 1.2:

Таблица 1.2 – Фрагмент мультилингвистического частотного словаря

tool, 61	Werkzeug, 4 *n*, Mittel, 17 *n*	средство, 100, инструмент, 1
basic tool, 61	Hauptmittel, 17 *n*	основное средство, 100
topic, 9	Thema, 4 *n*, Frage, 23 *f*	тема, 9, вопрос, 62
topology, 13	Topologie, 1 *f*	топология, 8
total, 129	Ergebnis, 13 *n*, zusammenfassend, 2	итог, 44, полный, 190
tracing, 9	Nachführung, 1 *f*	слежение, 1

Словарь в широком смысле является существеннейшим компонентом модели восприятия речи [96-108]. При восприятии речи основной оперативной единицей выступает слово. Из этого следует, в частности, что каждое слово воспринимаемого текста должно быть отождествлено с соответствующей единицей внутреннего словаря слушающего (или читающего). Естественно считать, что уже с самого начала поиск ограничен некоторыми подобластями словаря. Собственно фонетический анализ звучащего текста в типичном случае дает лишь некоторую частичную информацию о возможном фонологическом облике слова, и такого рода информации отвечает не одно, а определенное множество слов словаря; следовательно, возникает задача (а) выделить соответствующее множество по тем или иным параметрам и (б) в пределах очерченного множества (если оно выделено адекватно) произвести "отсев" всех слов, кроме того единственного, которое и соответствует наилучшим образом данному слову распознаваемого текста. Одна из стратегий "отсева" - исключение низкочастотных слов. Отсюда следует, что словарь для восприятия речи - это частотный словарь. Отчасти по этой причине, когда тождество слова во многом опирается на совпадение формальное, графическое, недостаточно учитывается семантика. В результате оказываются смещенными, искаженными и частотные характеристики; например, если слова из сочетания "друг друга" составитель частотного

словаря включает в общую статистику употребления слова "друг", то едва ли это оправданно: учитывая семантику, мы должны признать, что в составе сочетания это уже другие слова, а точнее, что самостоятельной словарной единицей выступает лишь само по себе сочетание в целом. Именно поэтому рассматриваются как четко формализованные лексические единицы, так и устоявшиеся лексемы [45, 46].

При организации ЭЧС, составляющих информационную базу поддержки мультилингвистической технологии преследуются следующие цели: отразить некоторые важные качественные и количественные аспекты употребительной лексики в различных терминологических областях (например, по системному анализу на английском, немецком, русском и китайском языке, полученные в результате статистического анализа и описания текста [69, 77, 79, 80, 90]). В результате проведенного статистического исследования и описания текстов по системному анализу объемом около 30000 слов был составлен список слов объемом более 2500, адекватность которых была установлена для всех четырех языков, а затем около 2000 слов были включены в частотный словарь, являющийся непосредственным компонентом компьютерной системы, реализуемой на базе МЛ технологии. Так как почти все частотные словари одноязычны, что значительно снижает степень активности слова при определенном значении слова в словарном минимуме, был проведен анализ многоязычных текстов на систематическом уровне. Словарь организован как англо-немецко-русский (и китайский – в последней редакции 2020 года), однако принцип его построения в виде мультилингвистической базы данных обеспечивает возможность использования его как двуязычного в любом выбранном варианте.

Необходимо отметить, что первая версия трехъязычного (английский -немецкий - русский) словаря создавалась в рамках сотрудничества с коллегами института германистики Рурского университета г. Бохума, Германия. Основная работа, связанная с проверкой и коррекцией немецкоязычной части словаря, была выполнена профессором Ковалевым И.В. в сотрудничестве с профессором Х.-Р. Флуком в Рурском университете г. Бохума (Германия) в сентябре-октябре 1999 г., июле 2000 г. и июле 2001 г. Эти визиты были бы невозможны без финансовой поддержки исследовательского фонда аэропорта во Франкфурте на Майне и Немецкой службы академического обмена (DAAD).

Позднее словарь был дополнен информацией о внутриязыковых лексических связях.

1.1.1 Терминологические стандарты и международная регламентация терминологии

Лексический состав научно-технических текстов характеризуется широким использованием терминов (слов, словосочетаний, фраз), являющихся инструментом, при помощи которого оперируют научно-техническими понятиями, и относительно малым разнообразием лексических групп. Основным признаком терминов является их определяемость, иными словами, формой существования термина является определение. В общем случае под термином понимается слово или словосочетание любой сферы употребления, выражающее научное понятие, имеющее определение и пригодное к применению в науке и технике. Все термины делятся на простые, сложные и словосочетания. Термины объединяются в терминологические системы, выражающие понятия определенной отрасли науки и техники [8, 20, 86].

В настоящее время в связи с появлением новых наук, открытием новых явлений возникают значительные трудности выделения общенаучной, общетехнической, отраслевой и узкоспециальной терминологии. Общенаучные и общетехнические термины - это термины, используемые в нескольких областях науки и техники. Отраслевые термины — это термины, присущие только одной какой-либо отрасли знаний. Узкоспециальные термины - это термины, имеющие значения, характерные для какой-специальности данной отрасли [85, 86].

В условиях, когда требуется давать названия все новым и новым явлениям и понятиям, многие слова общего языка приобретают определенную качественную специфику, происходит специализация их значений при переводе на другой язык и унификация переводных эквивалентов. Нередко в научно-технических текстах встречаются слова и словосочетания, относящиеся к другим функциональным системам [93].

Термины должны сохранять такие черты, как абстрагированный характер, однозначность и системность. Однако, многие термины и терминологические системы как лингвистические знаки до сих пор имеют такие недостатки, как многозначность (один термин имеет два и более значений), синонимия (для одного понятия существуют два и более термина), противоречие термина понятию, длиннота и

неудобопроизносимость термина, загруженность терминологии иноязычными терминами и т.д., поэтому необходимо проводить большую систематическую работу по созданию кратких, по возможности интернациональных терминов для новых понятий (отметим, что у математических знаков, которыми оперируют в теории информации и других областях, такие недостатки обычно отсутствуют) [85, 86].

Термин может быть создан на базе слов родного языка или заимствован как из нейтрального терминологического банка (международные греко-латинские терминоэлементы), так и из другого языка, он должен отражать признаки данного понятия и вводить в модель семантического треугольника знак - понятие - объект. Значение термина для специалиста равно значению понятия.

Стремление упорядочить терминологию приняло значительный размах в конце XIX века, когда появились тысячи новых терминов и технических понятий. Причем, для выражения одних и тех же понятий использовались разные термины. Так, например, известно, что к моменту начала работы Международной электротехнической комиссии (МЭК) в 1880 г. для электрического сопротивления применялось 15 разных единиц. Недостатки терминологии затрудняют взаимопонимание между специалистами, проведение учебного процесса, вызывают излишние затраты труда при работе со специальной литературой [86].

По мнению ряда авторов [79, 85, 86, 93], работа над терминологией должна быть построена по следующей схеме: 1) выявление основ данной науки или отрасли техники; 2) отбор понятий, их систематизация и группировка; 3) построение классификации понятий; 4) определение понятий; 5) отбор терминов из числа имеющихся синонимов с учетом принципов установления эффективной коммуникации; 6)построение новых терминов. Следует отметить, что хороший термин должен представлять определенные и упорядоченные концептуальные единицы и занимать свое собственное место в терминосистеме данного языка, его следует точно соотнести с другими терминами, он моносемичен, контекстуально независим, свободен от эстетических и эмоциональных признаков.

В настоящее время терминоведение является сложившейся научной дисциплиной. В роли термина может выступать всякое слово; подавляющее большинство терминов являются именами существительными или сочетаниями слов с именем с именем существительным в качестве основного слова. Все термины одной отрасли знания образуют терминологию, из которой можно выделить

терминосистемы, микросистемы, ряды терминов и сами термины, взаимосвязанные на экстралингвистическом, т.е. понятийном уровне.

Разработка единых методов создания терминов, единой терминологии содействует более активному участию специалистов разных стран в решении актуальных проблем современной науки и техники, повышает их информационную вооруженность. Вопросы выяснения, уточнения, пересмотра смысла терминов, установления их познавательного значения являются одними из наиболее актуальных, в существенной степени обуславливающих прогресс научного знания. Следует помнить, что значительная часть научного языка - это слова естественного языка, получившие специальный смысл, а также слова, употребляемые в их обычном значении. Широкое и частое употребление терминов является характерной чертой функционального стиля научно-технической литературы [86].

В последнее время остро встает вопрос о стандартизации, унификации, нормализации и инвентаризации терминов, создании таких терминологических словарей, которые бы в полной мере удовлетворяли нужды специалистов, работающих в различных областях науки и техники, отражали бы такие лексикографические принципы, как современность, информативность, многоязычность, дидактичность и др., совершенствовали бы процесс обработки научно-технической информации.

В нашей стране принято свыше 750 000 терминологических стандартов по различным областям знаний. Во ВНИИКИ создан справочный банк терминов в виде карточек, микрофильмов. В банк вводятся только стандартизированные термины. Терминологические стандарты содержат сам термин в начальной форме, его определение, эквиваленты на иностранных языках. Термины, установленные стандартом, обязательны для применения в документации всех видов, в научно-технической, учебной и справочной литературе. Ниже в таблице 3 даются примеры терминов на трех языках, приведенные в ряде Государственных стандартов Союза ССР, в том числе ГОСТ 19472-80. Сети телефонные. Термины и определения; ГОСТ 21415075. Конденсаторы постоянной ёмкости. Термины и определения; ГОСТ 13699-80. Запись и воспроизведение информации. Термины и определения; ГОСТ 23150-78. Коммутация каналов и коммутация сообщений. Термины и определения; и другие стандарты [22]. Примеры терминов на различных языках представлены в таблице 1.2.

Таблица 1.2 – Примеры терминов

Русский язык	Английский язык	Немецкий язык
База	Base region	Basis
Местная телефонная система	Local telephone system	Örtliches Telephonsystem
Импульсный коэффициент	Impulse coefficient	Impulsverhältnis

Важную роль в унификации специальной терминологии играют международные научно-технические организации, в том числе Международный союз электросвязи (МСЭ), Международная электротехническая комиссия (МЭК), Международная организация стандартизации (МОС) и др. Так, 160 странами мира - членами МСЭ принят Регламент радиосвязи - основной документ, регулирующий использование электросвязи и обязательный для всех членов. Первая глава Регламента радиосвязи посвящена терминологии, которой необходимо придерживаться при международном сотрудничестве, всего определено более 180 терминов. Например, [22]:

1) наземная радиосвязь. Любая радиосвязь, за исключением космической радиосвязи или радиоастрономии;

2) terrestrial Radiocommunication. Any Radiocommunication other than space radiocommunication or radio astronomy;

3) всемирное координированное время (UTC). Шкала времени, основанная на секунде в системе единиц, определенная и рекомендованная Международным консультативным комитетом по радио и поддерживаемая Международным бюро времени;

4) coordinated Universal Time (UTC). Time scale, based on the second, as defined and recommended by the International Radio Consultative Committee, and maintained by the International Time Bureau;

5) служба космической эксплуатации. Служба радиосвязи, предназначенная исключительно для эксплуатации космических кораблей, в частности, для целей космического слежения, комической телеметрии и космического телеуправления;

6) space Operating Service. A radiocommunication service concerned exclusively with the operation of spacecraft, in particular space tracking, space telemetry and space telecommand;

7) спутниковая служба стандартных частот и сигналов времени. Служба радиосвязи, использующая космические станции, установленные на спутниках Земли, для тех же целей, что и служба стандартных частот и сигналов времени;

8) standard Frequency and Time Signal-Satellite Service. A radiocommunication service using space stations on earth satellites for the same purpose as those and the standard frequency and time signal service;

9) судовой аварийный передатчик. Судовой передатчик, используемый исключительно на частоте бедствия для нужд, связанных с бедствием, срочностью или безопасностью;

10) ship's Emergency Transmitter. A ship's transmitter to be used exclusively on a distress frequency for distress, urgency or safety purposes.

Проблемы научно-технической терминологии, ее правильного употребления, понимания и толкования становятся все более актуальными, особенно в связи с развитием новых научных перспективных направлений, проникновением средств цифровой техники во многие области науки и техники, значительными достижениями в микроэлектронике и микропроцессорной технике, а также после создания Международной информационной службы по научным и техническим переводам (МИС «Интеринформперевод»).

Актуальной задачей является разработка отраслевых информационных терминологических банков данных – электронных терминологических словарей и проектирование информационных справочных систем, работающих в диалоговом режиме. Подготовка таких систем, содержащих стандартизированные термины и символы, является первой фазой стандартизации в отрасли и достижения международной согласованности, что приведет к повышению технического уровня и качества продукции.

Вместе с тем следует помнить, что в любом языке происходит постоянное взаимодействие между терминологией и нетерминологией [22].

Анализ методов образования английских научно-технических терминов позволяет сделать вывод о том, что основными способами их образования являются синтаксический, семантический и морфологический, а также заимствования из других языков и отраслевых терминологий [45].

Следует отметить, что большинство современных английских терминов образованы синтаксическим способом. Практика показывает, что в области электросвязи многокомпонентные термины составляют более 80 процентов всего лексикона - более 15 тысяч лексических единиц, что находит свое отражение в появлении все большего числа многокомпонентных терминологических сочетаний, например, такие термины, как hypothetical reference connection (гипотетическое эталонное соединение), subtelephone telegraphy (подтональная телеграфия), character signal (сигнал знака), equal-length code

(равномерный код), theoretical duration of a significant interval (теоретическая длительность значащего интервала), start-stop telegraph signal (стартстопный телеграфный сигнал), multi-channel voice-frequency telegraphy (многоканальное тональное телеграфирование), earth-return double phantom circuit (заземленная двойная фантомная цепь) [45].

Заимствования из других языков в английской терминологии в рассматриваемой области незначительны, в среднем такие лексические единицы составляют не более 3-4%. Наиболее распространенным типом морфологического способа образования является суффиксация. Так образуется 5-18% терминов [45].

Оптимизация работы над английским научно-техническим текстом требует учета словообразовательных отношений. Их изучение приводит к повышению эффективности и продуктивности работы.

В современной научно-технической литературе на английском языке наблюдается тенденция увеличения количества разного рода сокращений (аббревиации) всех морфологических классов слов и словосочетаний, образование новых слов путем сокращения существующих (в настоящее время зафиксировано более 250 тысяч сокращений в странах, пишущих на латинице), что является следствием информационной оптимизации сообщений и стремления коммуникации к минимизации сложности знаков.

В некоторых текстах на долю сокращений приходится более 50% всех словоупотреблений и более 15% словаря, что позволяет в несколько раз сжимать как объем передаваемой информации, так и графическое представление отдельных текстов, а также пополнять лексический состав языка.

Существуют сокращения буквенные, инициальные, слоговые, смешанные (сочетания букв и слогов), усеченные слова, стяжения, а также сокращения с помощью средств символики.

В качестве иллюстрации рассмотрим такие примеры возможных вариантов сокращений, наиболее характерных для английских научных текстов, как AASIR (advanced Atmospheric Sounder and Imaging Radiometer), B/W (Bothway), C/N (Carrier-to-Noise Ratio), CORODIM (Correlation of the Recognition of Degradation with Intelligibility Measurements), FAX (Facsimile), HDX (Half Duplex), Marots (Maritime Test Satellite), RD (Research and Development), RTT (Radioteletype), UTC (Coordinated Universal Time) [45].

Таблица 1.3 – Примеры технических аббревиатур

DER REG	Register does not operate;	Регистр не работает
ICI	Here is;	Здесь
OPH	Out of phase on system;	Несовпадение по фазе в системе
PERC	Permanent call on;	Непрерывный вызов в
RS	Reception switched over to;	Прием переключен на

Сокращения часто используются в служебных документах, при международном сотрудничестве они утверждаются на соответствующих конференциях и форумах. Так, согласно [45], в принятой на VIII Пленарной ассамблее Международного консультативного комитета по телеграфии и телефонии (Малага - Торремолинос, 1984 г.) Рекомендации R. 90 содержится 79 аббревиатур для использования всеми службами, ответственными за техническое обслуживание телеграфного оборудования, в том числе и приведенные в таблице 4.

1.1.2 Многокомпонентные термины и их анализ

При работе с лексикой английской научно-технической литературы наибольшую трудность для понимания и перевода представляют многокомпонентные термины - терминологические словосочетания, созданные лексическим и синтаксическим способом, т.е. представляющие собой словосочетания, образованные по определенным моделям.

Способ создания терминов в виде цепочки слов все шире входит в практику. Это происходит по объективным причинам, связанным с тем, что, во-первых, у любого языка существуют ограниченные ресурсы в плане лексических единиц и, во-вторых, результаты научно-технической революции приводят к новым открытиям и явлениям, требующим точных определений и наименований. Замечено, что в эпоху научно-технической революции пополнение лексики языка осуществляется, в основном, за счет специальной терминологии, которая является наиболее подвижной частью лексико-семантической системы языка; ежегодно такое увеличение составляет около 1000

новых терминов. Лексико-синтаксический способ составляет серьёзную конкуренцию таким традиционным методам, как семантическому и морфологическому, в научно-технических текстах такие конструкции относятся к числу самых употребительных. Терминологические словосочетания представляю собой семантически целостные сочетания двух и более числа слов, связанных с помощью предлога или беспредложным способом. Они могут быть устойчивыми и свободными сочетаниями [45].

При переводе терминологических словосочетаний необходимо четко уяснить, в каком порядке следует раскрывать значение данного словосочетания. Терминологические словосочетания строятся из сочетания существительного обычно в единственном числе (ядра словосочетания) с другими частями речи, которые могут стоять до и после него [45].

Особую трудность при переводе представляют беспредложные терминологические словосочетания, состоящие из цепочки слов, не связанных между собой какими-либо служебными словами (артиклями, предлогами и т.д.). В беспредложном терминологическом словосочетании главным словом является последнее, все слова, стоящие слева от него, играют второстепенную роль - роль определения. Период беспредложных терминологических словосочетаний надо начинать с главного слова [5]. Рассмотрим примеры, изображенные на рисунке 1.1:

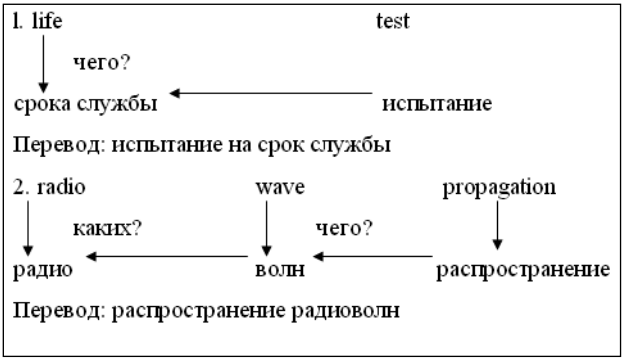


Рисунок 1.1 – Примеры последовательностей перевода беспредложных терминологических словосочетаний

Беспредложные терминологические словосочетания могут нести больший объем информации. Рассмотрим примеры, изображенные на рисунке 1.2:

Рисунок 1.2 – Пример последовательности перевода сложного беспредложного терминологического словосочетания

Обычно терминологические словосочетания классифицируются по их лексическому составу [4, 5]. Существуют терминологические словосочетания, состоящие только из существительных, из прилагательных и существительных, трехкомпонентные (содержащие различные комбинации причастий, существительных, наречий, герундия, инфинитивов и т.д.).

При переводе следует уяснить состав терминологического словосочетания, определить главное слово, перевести данное терминологическое словосочетание, а затем, отредактировав его, дать необходимый эквивалент на русском языке. При этом следует учитывать, что порядок слов в русских эквивалентах терминологических словосочетаний часто не совпадает с тем порядком слов, который используется в английских терминологических словосочетаниях. Так, например, определения, находящиеся слева от основного слова, при переводе довольно часто приходится ставить справа от основного слова [22]:

1) single-sideband full carrier emission - однополосное излучение с неподавленной несущей;

2) one-hop Pedersen-ray propagation - односкачковое распространение посредством луча Педерсена.

В терминологических словосочетаниях, состоящих из длинной цепочки слов, могут быть отдельные звенья, которые между собой обычно соединяются дефисом [22], как это проиллюстрировано на рисунке 1.3.

Рисунок 1.3 – Пример терминологического словосочетания с использованием знака дефиса.

Предложные терминологические словосочетания характеризуются тем, что главное слово в таких терминологических словосочетаниях стоит до предлога, а слова, стоящие за предлогом, играют роль определения. Например:

1) class of emission - класс излучения;
2) the development of new methods - создание новых методов;
3) the development of a new device - разработка нового прибора;
4) inclination of an orbit - наклонение орбиты.

Следует обратить внимание на то, что перевод главного слова зависит от значения его определения. Например:

1) noise from extra-terrestrial sources - шум внеземных источников;
2) noise in parts of radio links - шумы на участках радиолиний.

В составе терминологических словосочетаний могут быть также предложные словосочетания, являющиеся определением к главному слову. Например:

1) signal-to-noise ratio - отношение сигнал/шум;
2) point-to-point communication - направленная (радио) связь.

Предложные терминологические словосочетания могут быть выражены существительным, определением к которому обычно являются существительные с предлогом или герундий с предлогом, которые, как правило, стоят справа от определяемого слова. Например:

1) margin of start-stop apparatus - исправляющая способность стартстопных аппаратов;
2) office of destination - станция назначения;
3) noise from extra-terrestrial sources - шум внеземных источников;
4) method of signaling - способ передачи сигналов вызова.

Анализ терминологических словосочетаний приводит к выводу о том, что их модели определяются числом компонентов, что, в свою очередь, влияет на мотивированность терминологического словосочетания. Работа с научно-техническими текстами показывает, что наиболее частотными терминологическими словосочетаниями являются те, которые состоят из 2-3 компонентов, что является характерным для любой отраслевой терминологии [43]. На практике встречаются терминологические словосочетания, состоящие из четырех и большего числа компонентов [85].

Терминологические словосочетания позволяют передавать информацию в более сжатом виде, а тенденция к свертыванию речевого сообщения является характерным признаком стиля научно-технической литературы, поскольку в любом сообщении (высказывании) имеется определенная степень избыточности. Более того, терминологические словосочетания позволяют осуществлять межфразовые связи между предложениями и абзацами [85].

Владение методами построения терминологии на английском языке, знание типовых моделей повышает эффективность работы с текстами, скорость чтения, уменьшает затраты труда.

1.2 Анализ текстов для формирования ЭЧС

Возникающие направления языкознания в той или иной степени связаны с какими-то прикладными задачами. В настоящее время практически сформировалось направление прикладной лингвистики, связанное с преподаванием иностранных языков (педагогическая лингвистика, лингводидактическое направление) [46].

В ходе описания языка были получены некоторые самостоятельные результаты. А, кроме того, в русло нового направления попали некоторые разработки, появившиеся ранее, например, функциональная грамматика, лингвистика связного текста и, возможно, другие. Здесь тоже можно увидеть параллель с компьютерной лингвистикой, вобравшей в себя достижения структурализма, дескриптивизма и ряда других направлений.

Рассмотрим особенности педагогической лингвистики в связи с задачами, стоящими перед этим направлением.

Основной особенностью описания языка в целях преподавания является учет психологических свойств человека, связанных с речевой деятельностью. Эти особенности связаны со свойствами памяти (запоминания, хранения информации, ее активизации), понимания речи и ее порождения и с особенностями общения, понимаемыми в более широком аспекте (социальные, межличностные отношения и т.п.). Если сопоставить эти психологические сведения с информацией о структуре языка, представляемой грамматикой, то во многих случаях будут обнаружены нестыковки – факт, известный всем преподавателям неродного языка (особенно если для преподавателя он является родным). В первую очередь они касаются активной речевой деятельности, т.к. при этом требуется использовать информацию о соответствии языковых структур замыслам говорящего [85].

Традиционная грамматика предлагает сведения о значениях языковых единиц, которые не всегда могут быть «повернуты» в сторону речепроизводства. В качестве примера можно привести синонимы, информации о которых чаще всего недостаточно для правильного выбора в речи.

Аналогичные проблемы стояли и перед компьютерной лингвистикой, которая тоже при решении некоторых задач должна была моделировать речепорождение. Это заставило ее выработать свод правил «от смысла к тексту», ввести понятие глубинной семантики. И хотя далеко не все компьютерные правила оказались применимы в преподавании, обращение к ним было полезно, и взаимообогащение этих двух направлений продолжается и сейчас [85].

Еще одной особенностью педагогической лингвистики является специфика формулировки правил. В отличие от традиционного и компьютерного направлений от правил не требуется полнота, и их дополнения до полноты за счет списка исключений осуществляется только частично: человек не в состоянии усвоить большие списки. Некоторые исключения встречаются настолько редко, что при обучении языку ими можно пренебречь (например, устаревшие слова и конструкции). Однако свойства памяти требуют, чтобы как можно больше единиц было объединено в группы. Это заставляет выискивать максимальное количество связей, объединений [85].

Наряду с обычными правилами алгоритмического типа (взаимно-однозначными) используются и правила, которые можно назвать «односторонними»: они не дают полной определенности. Эти правила используются в дериватологии. Такая информация полезна при анализе текстов, т.е. помогает пониманию (хотя в основном в компьютерной лингвистике такие правила избегаются). Однако она может помочь и

при обучении активному речепорождению, поскольку ограничивает число возможных вариантов и помогает при запоминании. Например, есть большая разница между запоминанием производных слов целиком и запоминанием слова, состоящего из известного корня и одного из трех-пяти известных суффиксов [85].

В настоящее время прикладная лингвистика в наибольшей степени уделяет внимание работам в области NLP (Natural Language Processing). Среди них выделяются системы, автоматически сопоставляющие предложению его грамматическую структуру. Стоит отметить, что реально такого рода системы выступают как средства анализа текста, а не языка, и должны были бы представлять область NTP (Natural Text Processing). Язык (грамматика) либо берется как данное в каком-то из существующих вариантов, либо разрабатывается специально применительно к указанной задаче. К области NTP можно отнести и существенно иные работы, моделирующие восприятие речи: здесь также объектом приложения некоторой программы является текст, но на выходе - интерпретация последнего в терминах тех языковых единиц и категорий, которые, как можно полагать, используются носителем языка для смысловой записи данного текста. Строго говоря, такие работы уже не принадлежат прикладной лингвистике [97].

Языковую статистику можно определить как вспомогательную дисциплину языкознания, которая исследует количественные аспекты употребления языковой системы, в том числе, профессионально-ориентированной. Ранее в ряде работ успешно применялись при этом статистические методы [85, 93]. Языковая статистика дополняет качественные методы описания языка посредством дополнительных данных, характеризующих частоту языковых явлений, что весьма полезно в таких практических областях как информационный поиск, занятия по иностранным языкам. Математически этот подход позволяет моделировать профессионально-ориентированную языковую коммуникацию как вероятностный процесс, позволяя определить объективные параметры языковой дифференциации, которая находит выражение в различных субъязыках, профессиональных языках, профессиональных диалектах или стилях [86].

Языкознание использует статистические методы, прежде всего там, где речь идет о том, чтобы охватить язык функционально, в текстах, собрав отдельные отрывки в единое целое, что совершенно невозможно сделать как-либо иначе из-за большого разнообразия языковой коммуникации в различных профессиональных областях.

При автоматизации общего статистического анализа могут быть выделены следующие этапы: определение статистических элементов

(слово, фраза, предложение); определение абсолютной частоты элементов по единичной выборочной пробе и общей выборочной пробе; расчет относительной частоты и вероятности для основной совокупности терминов профессионального субъязыка; проверка достоверности полученных частотных характеристик путем вычисления стандартных отклонений и относительной ошибки; формализация результатов в виде списков, таблиц или графиков; интерпретация и обобщение результатов, вплоть до формулирования закономерностей [86].

Так как практически невозможно охватить всю общность предметно-языковой коммуникации даже только для одного языка и одной области, предметно-языковая статистика должна опираться на наиболее репрезентативные выборочные пробы, т.е. на письменные или устные предметно-типичные тексты. Каждый языково-статистический анализ начинается с выбора и подготовки соответствующей текстовой базы. При специфических постановках задач в рамках прикладного языкознания, например, при определении словарного запаса для заучивания на занятии по иностранному языку или при составлении вокабуляра для внутрипроизводственной документации, объем текстовой базы может быть сильно ограничен.

Необходимо также обращать внимание на вид текстов. Особенно пригодны для определения научно-технического основного словарного запаса учебники высшей и профессиональной школы обзорного характера. Они гарантируют систематический, пропорциональный и полный охват материала и необходимые языковые средства для его изложения, кроме того, они в меньшей степени подвержены влиянию со стороны индивидуального языкового употребления отдельных представителей профессии. Дальнейшее формирование текстовой базы основывается на использовании новых журналов не специального характера. Справочники, рефераты, прогрессивные сообщения, инструкции по применению и другие виды текстов, напротив, являются благоприятным исходным базисом для наблюдения предметно-языковых особенностей на уровне предложения и текста [86].

Первым результатом статистической обработки текста является абсолютная частотность. Она показывает, как часто возникает соответствующее явление в исследуемом тексте. Однако, она имеет малую ценность для дальнейших исследований при практическом использовании результатов или вообще для обобщенных высказываний, так как она напрямую зависит от объема выбранного текста [61]. Она служит исключительно как исходная величина, например, для расчета относительной частотности.

Относительная частотность – процентная величина, которая выражает долю языковой единицы в целом тексте. Она получается из деления абсолютной частотности на длину выборочной пробы, например, для слова с частностью 186 в одной выборочной пробе из N=50000 относительная частотность будет вычисляться как 186/50000=0,00372.

Другими словами, относительная частотность явления – отношение числа его действительного возникновения к числу его теоретически возможного появления. Если выборка по величине репрезентативна для предметного языка, тогда можно приравнять относительную частотность к вероятности языкового явления. Тогда она дает основание высказываниям о статистической структуре соответствующего субъязыка или о важности отдельных элементов для организации текста [86].

Особенно важный шаг при языково-стилистическом анализе – контроль достоверности определяемых данных. Для этого в распоряжении имеются различные способы контроля. В стилистической статистике и в предметно-языковой статистике учитываются прежде всего стандартные отклонения (погрешности), относительная ошибка и конфиденциальные границы.

Стандартная погрешность (средняя квадратная погрешность) – мера изменчивости средней частотности языкового явления в частичных выборочных пробах [86]. Для ее расчета служит формула

$$S = \sqrt{\frac{SAQ}{n-1}} \quad , \tag{1.1}$$

где S – стандартная погрешность, SAQ – сумма квадрата погрешности, n – число контрольных проб.

Относительная ошибка вычисляется прежде всего для определенных лексических единиц в частотных словарях, чтобы определить достоверность этих словарей [86]. Это происходит с помощью формулы:

$$|f - p| = Zp\sqrt{\frac{p(1-p)}{n}} \ , \tag{1.2}$$

где f – относительная частотность, p – вероятность, Zp – коэффициент для данного уровня доверия p, n – объем контрольной пробы.

В лингвистических работах применяются упрощенные варианты этой формулы, которые исходят из того, что у маленькой p разница (1-p) примерно равна единице [86]. Общеупотребительный вариант определения относительной ошибки выглядит следующим образом:

$$\delta = \frac{Zp}{\sqrt{nf}}, \tag{1.3}$$

где δ - относительная ошибка,
Zp – коэффициент для данного уровня доверия p,
n – объем выборки (выборочной пробы),
f – относительная частотность.

Расчет интервала доверия – утонченный вариант расчета относительной ошибки, с которой определяется нижняя и верхняя граница (p_1 и p_2) колебаний и средняя частотность. Существуют разные способы расчета, например [86]:

$$p_1 = \frac{fN+\frac{1}{2}Zp^2-Zp\sqrt{f(1-f)N+\frac{1}{4}Zp^2}}{n+Zp^2}, \tag{1.4}$$

$$p_2 = \frac{fN+\frac{1}{2}Zp^2+Zp\sqrt{f(1-f)n+\frac{1}{4}Zp^2}}{n+Zp^2}. \tag{1.5}$$

С помощью χ^2-теста определяется, характерны ли различия частотности, с которыми появляются языковые явления в различных выборочных пробах, или принадлежат ли выборочные пробы одной и той же основной совокупности (функциональному стилю, субъязыку, предметному языку, виду текста и т.д.). В большинстве случаев при этом идет речь о проверке (удостоверении в подлинности или фальсификации) основного предположения (нулевой гипотезы); например, ожидания того, что виды слов при организации текста приблизительно играют одинаковую роль. Контрольная величина χ^2 представляет собой сумму отнесенного к ожидаемым частотностям

квадрата разницы между наблюдаемой и ожидаемой частотностями для определенного числа переменных [86].

$$\chi^2 = \sum_{i=1}^{k} \frac{(f_{bi} - f_{ei})^2}{f_{ei}}, \quad (1.6)$$

где k – число переменных,
i – переменная,
f_{bi} – наблюдаемая частотность переменной,
f_{ei} – ожидаемая частотность переменной.

При сравнении выборочных проб ожидаемая частотность f_{ei} обычно приравнивается к средней частотности $\overline{x}$.

$$\chi^2 = \frac{\sum_{i=1}^{k}(x_i - \bar{x})^2}{\bar{x}}. \quad (1.7)$$

При отображении результатов исследований предметно-языковая статистика пользуется различными списками, таблицами, графиками. С помощью кругового изображения и ленточных диаграмм изображаются прежде всего части в процентных величинах. Для графического изображения количественных признаков, таких как длина слова или предложения, пригодны гистограммы и цепь многоугольников. Кривые с более или менее типичным течением превышают это простое сочетание частотностей по качественным и количественным признакам. Они позволяют распознавать функциональные связи между признаками и их частотностью, и частотность языковых явлений сама может стать признаком того, что характеризуется другими данными [86].

Существуют, например, следующие зависимости:

– между частотностями лексических единиц и их классами в словаре частотности;
– между частотностью и ее вероятностью возникновения в тексте;
– между частотностью и относительной ошибкой;
– между классами одного словаря частотностей и кумулятивным числом лексических единиц;
– между классами и классовой оболочкой;
– между частотностью и потенцией связи;
– между частотностью и степенью специализации предметного лексического запаса слов;
– между длиной текста и объемом лексики и т.д.

Статистические методы все шире и глубже проникают в самые различные области научной деятельности, в том числе и в такие, традиционно считавшиеся “неточными”, как языкознание [86]. Совокупность этих методов, используемых в науке о языке, или лингвостатистика, позволяет определить, какие языковые явления встречаются в речи или тексте чаще, а какие реже. Наиболее интенсивно статистическими методами изучается словарный состав языка. Сведения об употребительной лексике дают частотные словари. Частотный словарь регистрирует слова, словоформы или словосочетания, которые встретились в исследованном для его составления тексте (выборке). При этих единицах (т.е. словах, словоформах, словосочетаниях) в словаре указываются их частоты, т.е. цифры, показывающие, сколько раз каждая словарная единица встретилась в данном тексте [4, 5].

Составление частотного словаря требует значительных затрат времени и знакомства со статистической методикой наблюдений. Например, создатели частотного англо-русского словаря-минимума газетной лексики [4] придерживались следующей методики. Словарь составлялся на основе лингвостатистического анализа языка газет и журналов Великобритании и США. Были отобраны тексты общей длиной 200000 словоупотреблений (под словоупотреблением в лингвостатистике принято понимать слово как единицу длины текста) из разных газет и журналов. Из этих текстов вручную выписывались слова и словосочетания с различной степенью устойчивости. В наше же время создание частотных словарей подразумевается с помощью компьютера.

Текст обладает статистической структурой. Сущность ее состоит, с одной стороны, в том, что все слова и выражения, составляющие словарь текстов по данной специальности, а также грамматические формы и синтаксические конструкции, обладают определенной вероятностью появления в текстах этой специальности [4, 5].

С другой стороны, если разбить специальный текст на малые порции, то одна часть лингвистических единиц будет давать в этих текстах примерно одни и те же частоты, обнаруживая, таким образом, устойчивость и равномерность употребления. Другая группа лингвистических единиц дает неустойчивое и неравномерно употребление в отдельных порциях текста. Первую группу обычно составляют служебные слова и общеупотребительные словосочетания. Вторую группу образуют чаще всего слова и словосочетания, непосредственно связанные с содержанием текста данной

специальности (эти слова и словосочетания часто называют ключевыми).

Структура текстов разных специальностей неодинакова: в них заметно отличаются вероятности и распределения ключевых слов и словосочетаний, статистика же служебных слов и некоторых общеупотребительных слов и словосочетаний остается неизменной.

Если исследовать текст достаточно большого объема, можно обнаружить закономерности функционирования лексики данного языка и получить представление о ее количественной структуре. При таком анализе выявляются, например, две важнейшие лингвостатистические закономерности [4, 5].

Во-первых, в любом тексте, каким большим бы он не был, используется лишь незначительная часть словарного богатства языка.

Очевидно, что лексика ребенка намного беднее, чем словарь взрослого человека. И все же полученные результаты эксперимента удивляют: всего 25000 разных словоформ оказалось в 100000 писем, классных и домашних работ общей длиной 6 млн. словоупотреблений (данные анализа письменной речи школьников) [4, 5].

Специальные научно-технические и публицистические тексты также весьма отличаются по объему словаря. Анализ текстов на английском, румынском и молдавском языках показал, что словарь текстов публицистического характера примерно в 2,5 раза больше, чем словарь специальных текстов [4, 5]. Эти цифры свидетельствуют о том, что в разных сферах речевого общения используются и разные количества слов.

Вторая лингвостатистическая закономерность состоит в том, что даже и ограниченная часть лексики языка используется в речи (тексте) неравномерно. Одни слова употребляются чаще, другие реже, причем большая часть всего текста приходится на незначительное количество самых частых слов. Например, при записи и анализе телефонных разговоров были получены следующие результаты: 737 самых частых слов занимают свыше 95% всех словоупотреблений [4, 5].

Как уже отмечалось, в частотном словаре указывается количество случаев употребления слова в тех текстах, которые были проанализированы для составления словаря. Частотные словари различаются в зависимости от принципа размещения материала. Слова или словосочетания могут быть расположены по алфавиту – как в обычном словаре, с проставлением рядом со словом его частоты. Также слова и словосочетания могут быть расположены по убыванию частот, начиная от самого употребительного слова. Первый вариант словаря адресуется обучаемому, второй – обучающему. Обучаемый может

также работать со вторым вариантом словаря при самостоятельном изучении иностранного языка, например, при заучивании слов и словосочетаний порциями в зависимости от их частоты или при проверке владения словарными единицами, начиная с самых частых [4, 5].

1.3 ИТБ на основе лексически связанных компонентов

Информационно-терминологическая база МЛ-технологии строится на основе результатов анализа языкового материала. Под языковым материалом здесь следует понимать некоторое множество текстов интересующей разработчика предметной области анализируемого языка. Размер языкового материала может варьироваться в зависимости от средств анализа, наличия оригинальных текстов и необходимого количества терминов.

Основу построения ИТБ составляют ЭЧС, по данным которых определяется приоритетность тех или иных терминов. Использование таких ЭЧС, полученных путем анализа языкового материала, качественно улучшает процесс анализа иностранной лексики при исследовании уникальности текста.

Структура ИТБ МЛ-технологии разрабатывалась посредством ориентированных на данные методологий структурного анализа с учетом частотных характеристик ее элементов [78-82].

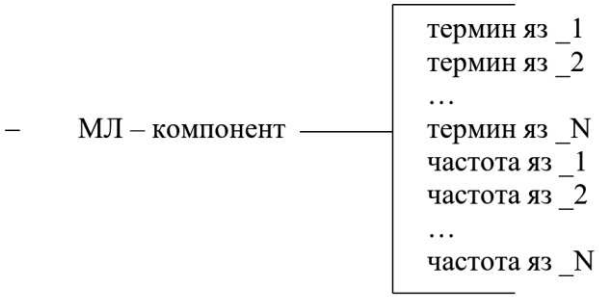
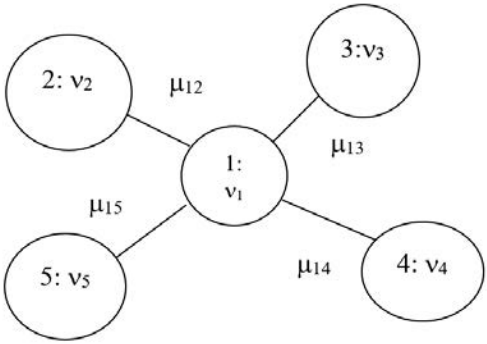


Рисунок 1.4 – Описание базисного информационного компонента с использованием нотации DSSD

Терминологическое множество, соответствующее базисному информационному компоненту МЛ-технологии, в [35-45] описывалось как: МЛ-компонент = {термин яз_1, термин яз_2, ..., термин яз_N, частота яз_1, частота яз_2, ..., частота яз_N}.

DSSD [33,34] использует аналогичную нотацию, а именно множественную скобку как это показано на рисунке 1.4.

Таким образом, ИТБ будет представлен совокупностью объектов, структура которых схематически представлена на рисунке 1.5.



лексемы: 1 – основная лексема; 2,3,4,5 – связанные лексемы (лексические связи: 1-2, 1-3, 1-4, 1-5)
количественные характеристики: ν_I – абсолютная частота i-ой лексемы; μ_{ik} – абсолютная частота сочетания; i–ой и k–ой лексем

Рисунок 1.5 – ЛСК в структуре ИТБ

Такие комплексные объекты называются лексически связанными компонентами (ЛС-компонентами) ИТБ [35, 37, 38]. Лексема, связанная со всеми без исключения лексемами ЛС-компонента ИТБ называется основной лексемой, лексемы же, имеющие только одну связь – связанными лексемами.

Таким образом, МЛ-компонент уже не будет играть роль базисного информационного компонента. Но эту роль будет играть лексически связанный компонент. Опишем его структуру, ориентируясь на данные. Для простоты, воспользуемся уже известным нам описанием МЛ-компонента.

ЛС-компонент = {МЛ-компонент (основная лексема), МЛ-компонент (связанная лексема №1), МЛ-компонент (связанная лексема №2), …}

В нотации DSSD [108] ЛС-компонент будет выглядеть, как это показано на рисунке 1.6:

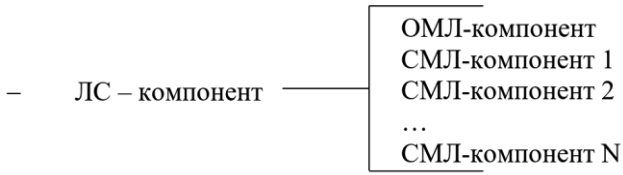


Рисунок 1.6 – Описание базисного информационного компонента с использованием нотации DSSD

Но при этом, структура МЛ-компонента, базирующегося на основной лексеме (основного) должна быть изменена, иначе теряется информация о лексических связях, а значит и смысл всей методики.

ОМЛ-компонент = {термин яз_1, термин яз_2,..., термин яз_N, частота яз_1, частота яз_2,..., частота яз_N, сочетание 1_яз_1, сочетание 1_яз_2,..., сочетание 2_яз_1, сочетание 2_яз_2,…, сочетание K_яз_N, частота 1_яз_1, частота 1_яз_2,…, частота 2_яз_1, частота 2_яз_2,…, частота K_яз_N}.

На рисунке использована нотация DSSD для описания ОМЛ-компонента. На первый взгляд структура ОМЛ-компонента выглядит громоздкой. Возникает вопрос о том, насколько ресурсоемко будет формировать и эксплуатировать такие компоненты. Но при ближайшем рассмотрении становится очевидным, что ОМЛ-компоненты включили в себя МЛ-компоненты, образованные от словосочетаний. Таким образом, исходный объем ИТБ не изменился, как не изменилась и ресурсоемкость его формирования. Ресурсоемкость в процессе анализа текста увеличится незначительно, так как сам процесс носит итеративный характер, а на каждой итерации анализа текстового материала такой элемент всего один. Структура остальных МЛ-компонентов (связанных) не претерпела изменений.

Лексические сочетания и их частотные характеристики, полученные путем анализа нейросетевого ИТБ [87], были включены в структуру ОМЛ-компонента затем, чтобы сохранить принцип мультилингвистичности. Дело в том, что для словосочетания на иностранном языке (в момент установления ассоциативного поля), согласно методике, должен быть сформирован аналог на другом (иностранном) языке. Это не что иное, как языковой аналог данного словосочетания. Но ввиду различного синтаксиса языков, входящих в ИТБ, генерировать такие языковые аналоги простой комбинацией лексем совершенно недопустимо. Поэтому для сохранения принципа

мультилингвистичности необходимо иметь точные языковые аналоги лексических сочетаний как отдельных понятий предметной области.

1.4 Анализ уникальности на базе ЛСК

Методика строится на совместном применении двух алгоритмов [57]:

– базового алгоритма [63], за тем исключением, что в качестве элементов текстового материала выступают не лексемы, а ЛС-компоненты [38];

– алгоритм построения внутриязыковых ассоциативных полей.

Алгоритм построения внутриязыковых ассоциативных полей прост и состоит в последовательной подаче к анализу элементов ЛС-компонента. Последовательность такова:

1) основная лексема – перевод, текстовый блок-подсказка (ТБП) на иностранном языке (в рамках МЛ-технологии);

2) связанная лексема – перевод, ТБП на иностранном языке (в рамках МЛ-технологии);

3) лексическое сочетание основной и связанной лексем – перевод сочетания, ТБП на иностранном языке (языковой аналог именно лексического сочетания, но не лексем по отдельности);

4) переход к следующей связанной лексеме;

5) переход к следующему лексически связанному компоненту.

При использовании данной методики базовый алгоритм будет в значительной степени модифицирован. Во-первых, и как уже говорилось, основным элементом анализируемой информации будет не термин, но лексически связанный компонент, поэтому возникает необходимость в разработке механизма формирования порций текстовой информации, который бы обеспечил целостность лексически связанных компонентов как базисных. Во-вторых, основываясь на алгоритме построения внутриязыковых ассоциативных полей можно говорить о том, что зависимость лексем как в одном языковом поле, так и в разных языках будет неоднородной, так как акцент делается на основные лексемы. Таким образом, возникает необходимость в модификации алгоритмической поддержки процесса анализа уникальности текстов на базе лексически связанных компонентов.

Решение задачи разработки методики анализа уникальности текстов на основе лексически связанных компонентов базируется на следующих основных выводах:

1) при компьютерном моделировании электронные частотные словари являются основой информационного обеспечения мультилингвистической технологии анализа уникальности текстов на основе лексически связанных компонентов;

2) разработку мультилингвистических электронных частотных словарей необходимо осуществлять в соответствии с тенденциями изменения спроса на рынках услуг по проверке текста на уникальность и согласно направлениям и проблематике научных и научно-технических исследований (материалов статей, монографий, диссертаций), необходимых для развития ключевых направлений научной и образовательной деятельности в сфере сохранения и приумножения научно-технического потенциала;

3) информационно-терминологический базис как совокупность лексически связанных компонентов может быть построен на основе электронных частотных словарей с данными как о внутриязыковых лексических связях, так и на уровне взаимодействия разных языков;

4) методика мультилингвистического анализа текста на основе лексически связанных компонентов требует модификации базового алгоритма и разработки алгоритмов формирования информационно-терминологического базиса как совокупности лексически связанных компонентов.

2. Алгоритмы формирования информационно-терминологического базиса как совокупности лексически связанных компонентов

Для выработки критерия отбора основных лексем рассмотрим ИТБ в момент времени t_{n+1}, когда прошло уже некоторое время, с тех пор как базис был пройден. Допустим, что ИТБ был построен как совокупность ЛС-компонентов, тогда основные лексемы согласно приведенного выше адаптивно-обучающего алгоритма будут заучены много лучше, чем их связанные лексемы. Вероятность знания лексем определяется согласно формуле:

$$p_{зні}^{n+1} = p_{зні}(t_i^{n+1}) = e^{-\alpha_i^{n+1} t_i^{n+1}} \quad , \tag{2.1}$$

Значения $p_{зн\ i}^{n+1}$ позволяют определить какие из лексем ИТБ более других подходят на роль основных, но это возможно только к концу обучения, при условии, что базис уже сформирован как совокупность ЛС-компонентов. Перед нами стоит обратная задача – сформировать ИТБ до начала обучения [39]. Однако, по частотным характеристикам лексем и лексических сочетаний возможно попытаться оценить значение $p_{зн\ i}^{n+1}$ для всех лексем базиса, и, исходя из этого набора значений, построить критерий для выделения основных лексем.

Предположим, что нам удалось вычислить оценку $\hat{p}_{зн\,i}^{n+1}$ для каждой лексемы, тогда учтя относительную частоту лексем q_i, построим искомый критерий:

$$L_i = \hat{p}_{зн\,i}^{n+1} q_i \to \max_{i \in N} \quad , \tag{2.2}$$

Теперь найдем значение $\hat{p}_{зн\,i}^{n+1}$.

$$p_{зн\,i}^{n+1} = e^{-\frac{b_i^{n+1}}{\hat{i}k \frac{\sum_k p_{зніk}^{n+1} \mu_{ik}}{\sum_k \mu_{ik}} + 1} t_i^{n+1}} \quad , \tag{2.3}$$

где t_i^{n+1}– время с последнего заучивания i-ой лексемы, о котором нам ничего не известно;

b_i^{n+1} – скорость забывания i-ой лексемы к моменту времени t_i^{n+1}, о которой нам так же ничего не известно, и которая вычисляется итеративно в процессе обучения;

$p_{знik}^{n+1}$ – вероятность знания k-ой лексемы, которая лексически связана (т.е. порождает ассоциацию), с i-ой лексемой.

Здесь нужно заметить, что задача выработки критерия для выделения основных лексем сводится к тому, чтобы для каждой лексемы было рассчитано значение (7), и это значение было тем больше, чем больше эта лексема подходит на роль основной [34, 39, 60]. Это обстоятельство позволяет нам во многом смириться с частичным незнанием элементов выражения (2.3). Так, например, для разбивки ИТБ на ЛС-компоненты совершенно не важны значения t_i^{n+1} и b_i^{n+1}, которые мы заменим соответственно на условную безразмерную единицу времени и на b_0 (получаемое из условия p = 0.5, t = 1) для всех лексем ИТБ, так же примем $\hat{h} = 1$.

Что касается вероятности $p_{знik}^{n+1}$, ее мы можем оценить через относительную частоту k-ой лексемы, т.к. к концу обучения, согласно адаптивно-обучающему алгоритму, вероятности знания всех лексем будут прямо пропорциональны их относительным частотам.

Таким образом, выражение для $\hat{p}_{\text{зн i}}^{n+1}$ будет выглядеть так:

$$\hat{p}_{\text{зн i}}^{n+1} = \mathrm{e}^{-\frac{b_0}{k\frac{\sum_k q_k \mu_{ik}}{\sum_k \mu_{ik}}+1}}, \qquad (2.4)$$

где $b_0 \approx 0.7$ (получаемое из условия p = 0.5, t = 1);

q_k и μ_{ik} содержаться в ИТБ.

Подставив выражение (2.4) в (2.2), получим коэффициент значимости основных лексем, который определяет меру эффективности использования i-ой лексемы как основной в ЛС-компоненте:

$$L_i = e^{-\frac{0,7}{k\frac{\sum_{j=1}^{k} q_j \mu_{ij}}{\sum_{j=1}^{k} \mu_{ij}}+1}} q_i \quad , \qquad (2.5)$$

где μ_{ik} – относительная частота сочетания i–ой и k–ой лексем, отражает силу ассоциативной связи;

q_i – относительная частота, выражающая долю лексической единицы в тексте, подвергшемуся статистической обработке при составлении частотного словаря, $0<q_i<1$.

Таким образом, мы получили значение L_i для каждой лексемы ИТБ, теперь необходимо определить количество основных лексем и задачу о формировании ИТБ как совокупности ЛС-компонентов можно будет считать решенной.

Задачу о нахождении оптимального количества основных лексем возможно решить с помощью перебора [34, 60], для этого введем функцию качества ИТБ как совокупности ЛС-компонентов от числа основных лексем:

$$L(n) = \sum_{i=1}^{N} q_i e^{-\frac{0,7}{k\frac{\sum_{j=1}^{k} q_j \mu_{ij}}{\sum_{j=1}^{k} \mu_{ij}}+1}}, \quad (2.6)$$

L(n) показывает сумму взвешенных вероятностей знания лексем по всему базису, естественно, что чем больше эта сумма, тем более удачно построен базис. Отсюда, максимизируя L(n), получаем оптимальное значение количества основных лексем.

Итак, данный алгоритм можно разделить на три основные фазы [60].

1. Подготовка ИТБ.

Для каждой лексемы ИТБ вычисляется значение коэффициента значимости по следующей формуле (2.5).

1.2 ИТБ упорядочивается по убыванию значения Li (таким образом, чем меньше будет порядковый номер лексемы, тем выше вероятность, образования на ее основе ЛС-компонента).

1.3 Данные о лексических связях упорядочиваются по убыванию значения $q_k \mu_{ik}$ (тем самым увеличивается вероятность попадания в ЛС-компонент тех из связанных лексем, которые более всего могут улучшить качество ИТБ).

2. Поиск оптимального количества основных лексем.

2.1 Осуществляется перебор возможного количества основных лексем k от 1 до объема ИТБ (возможно сужение интервала поиска разработчиком).

2.2 Для текущего значения k, определяются основные лексемы (k первых лексем ИТБ).

2.3 Для выбранных основных лексем определяются связанные лексемы (как правило, задается максимум количества связанных лексем).

2.4 Вычисляется значение функции качества ИТБ согласно формуле (2.6).

2.5 Если перебор окончен, идем в пункт 2.6, иначе возврат к пункту 2.1.

2.6 Определяем максимум функции качества от числа ЛС-компонентов (оптимальное число основных лексем k_{max}).

3. Формирование ИТБ как совокупности ЛС-компонентов (искомый ИТБ получается при прохождении пунктов 2.2 и 2.3 для k_{max} основных лексем).

Нисходящий алгоритм формирования ЛС-компонентов дает «хорошие» результаты [42, 56], однако из пунктов 2.2 и 2.3 следует, что связанные лексемы определяются согласно порядку основных лексем. Естественно, что связанная лексема, являясь частью одного ЛС-компонента, уже не может быть частью другого, даже если она "подходит" к нему больше (речь идет о конечном значении L(n)). Таким образом, возникает задача о нахождении наиболее "подходящих" связанных лексем для ЛС-компонентов в процессе их формирования [53]. Эта задача решается от обратного, т.е. не подбирая для основных лексем связанные, но наоборот. Алгоритм, работающий по такому принципу, получил название восходящего алгоритма формирования ЛС-компонентов [51] и так же может быть поделен на три основные фазы.

2.1. Восходящий алгоритм формирования ИТБ

1. Подготовка ИТБ.

1.1 Для каждой лексемы ИТБ вычисляется значение L_i.

1.2 ИТБ упорядочивается по убыванию значения L_i.

2. Поиск оптимального количества основных лексем.

2.1 Осуществляется перебор возможного количества основных лексем k от 1 до объема ИТБ (возможно сужение интервала поиска разработчиком).

2.2 Для текущего значения k, определяются основные лексемы (k первых лексем ИТБ).

2.3 Осуществляется перебор "не основных" (потенциально связанных) лексем и для каждой "не основной" лексемы l рассчитывается множество значений приращения функции качества, полученные путем ее включения во все возможные ЛС-компоненты. Расчет осуществляется по формуле:

$$\Delta L_i = q_i e^{-\frac{0,7}{(k+1)\frac{q_l\mu_{il}+\sum_{j=1}^{k} q_j\mu_{ij}}{\mu_{il}+\sum_{j=1}^{k}\mu_{ij}}+1}} - L_i^k \qquad (2.7)$$

где L_i^k – коэффициент значимости i-ой основной лексемы, образующей ЛС-компонент с k связанными лексемами;

Далее определяется ЛС-компонент для которого ΔL_i будет максимальным, лексема l включается в его состав в качестве связанной (как правило, задается максимум количества связанных лексем).

2.4 Подсчитывается значение функции качества.

2.5 Если перебор окончен, идем в пункт 2.6, иначе возврат к пункту 2.2.

2.6 Определяем максимум функции качества от числа ЛС-компонентов (оптимальное число основных лексем k_{max}).

3. Формирование ИТБ как совокупности ЛС-компонентов.

3.1 Незадействованные в ЛС-компонентах лексемы из числа основных (k_{max}) помечаем как "не основные". Нахождение наиболее "подходящих" связанных лексем порождает свободные элементы из числа потенциально основных лексем, что во многом ухудшает L(n); поэтому незадействованные в ЛС-компонентах лексемы из числа основных (k_{max}) помечаем как "не основные".

3.2 Для полученного значения k_{max} основных лексем осуществляем шаги 2.2 и 2.3 и получаем тем самым искомый ИТБ.

2.2 Сравнительный анализ приведенных алгоритмов

Проведем сравнительный анализ алгоритмов на 3-х ИТБ одинаковой структуры, но различного объема [54].

Настраиваемые параметры базиса:

– максимальное количество связей, приходящихся на одну лексему (10);

– максимальное значение абсолютной частоты лексем (100/50000);

– максимальное значение частоты сочетаний лексем (20/50000);

– объем материала, по которому произведен частотный анализ (50000);

– коэффициент связанности лексем (1);

Таблица 2.1 – Результаты теста №1 (объем базиса 1000 терминов)

Название алгоритма	Н-алгоритм	В-алгоритм

min L(n)	0,496118513639315	0,496118513639315
max L(n)	0,496118565143325	0,496118569546250
Основных лексем	188	172
Свободных элементов	320	324
Итераций	4981096	2000498500
Время исполнения (сек)	3,41	130,75

Таблица 2.2 – Результаты теста №2 (объем базиса 2000 терминов)

Название алгоритма	Н-алгоритм	В-алгоритм
min L(n)	0,994799407656066	0,994799407656066
max L(n)	0,994799517816764	0,994799524607485
Основных лексем	410	394
Свободных элементов	545	538
Итераций	19930343	21542653063
Время исполнения (ч:м:с:сс)	00:00:18:22	00:23:28:16

Таблица 2.3 – Результаты теста №3 (объем базиса 5000 терминов)

Название алгоритма	Н-алгоритм	В-алгоритм
min L(n)	2,47712640220252	2,47712640220252
max L(n)	2,47712664878646	2,47712666142042
Основных лексем	845	822
Свободных элементов	1867	1886
Итераций	124568303	253340376005
Время исполнения (ч:м:с:сс)	00:01:17:50	04:35:58:57

Приведем сводную таблицу результатов тестирования для времени исполнения и L(n).

Таблица 2.4 – Сводная таблица результатов тестирования

Название алгоритма	Н-алгоритм	В-алгоритм
ΔL(n) (1000)	0,00000005150401	0,000000055906935
ΔL(n) (1000), %	100	108,54
ΔL(n) (2000)	0,000000110160698	0,000000116951419
ΔL(n) (2000), %	100	106,16
ΔL(n) (5000)	0,00000024658394	0,0000002592179
ΔL(n) (5000), %	100	105,12
Δt (1000), %	100	$38,34 * 10^2$
Δt (2000), %	100	$76,7 * 10^2$
Δt (5000), %	100	$213,65 * 10^2$

Оценить насколько улучшает структуру базиса тот или иной алгоритм мы не можем, ввиду ограниченной специфики L(n), которая служит для нахождения оптимального числа основных лексем и не может использоваться как абсолютный показатель качества ИТБ; но мы можем, используя экстремальные значения этой функции сравнить алгоритмы формирования ЛС-компонентов между собой [55].

В таблице 4 приведены результаты 3-х экспериментов (ΔL(n) = max L(n) – min L(n)), согласно которым В-алгоритм превосходит Н-алгоритм на 8,54; 6,16; 5,12 %, соответственно. С чем связано снижение этого превосходства сложно сказать, в виду того, что проведено всего три эксперимента. Но можно предположить, что негативное влияние свободных элементов из числа основных лексем (которые мы записываем как связанные) на L(n) становится сильнее с ростом ИТБ и постепенно подавляет положительное влияние "наиболее подходящих" связанных лексем. Поскольку ИТБ, с которыми работает ЛСК-методика, предметно-ориентированы и их объем, как правило, не превышает 5000 терминов, ограничимся на том, что В-алгоритм превосходит Н-алгоритм по качественным показателям на 5-10 %.

Что касается времени исполнения, то здесь заметна тенденция: при увеличении объема ИТБ в n раз, время исполнения алгоритмов возрастает в геометрической прогрессии, причем для В-алгоритма это время возрастает в n раз быстрее (исключая помехи на больших объемах ИТБ).

Несмотря на большую, негативную разницу по времени исполнения, В-алгоритм формирует более качественную структуру ИТБ. Поскольку операция формирования ИТБ производится только один раз, показатель качества много важнее времени исполнения алгоритма.

Таким образом, В-алгоритм может успешно использоваться при формировании ИТБ как совокупности ЛС-компонентов. Единственным его серьезным недостатком является не время исполнения, а то, что заранее невозможно предугадать, сколько именно основных лексем будет в ИТБ. В случае, когда разработчик выставляет жесткие требования к количеству основных лексем (ЛС-компонентов) или когда важно время исполнения (очень большие объемы ИТБ), следует использовать Н-алгоритмы формирования ЛС-компонентов, в противном случае, предпочтение следует отдавать В-алгоритмам.

В главе 2 решена важная проблема, заключающаяся в разработке алгоритмического аппарата мультилингвистического анализа уникальности текстов для систем, ориентированных на формирование внутриязыковых групп ассоциативно связанных элементов. Решение этой проблемы базируется на следующих основных выводах и результатах:

- разработано модельное и алгоритмическое обеспечение для мультилингвистического анализа уникальности текстов для систем, ориентированных на формирование внутриязыковых групп ассоциативно связанных элементов на основе формирования лексически связанных компонентов в структуре ИТБ;
- показано, что лексически связанные компоненты обеспечивают целостность восприятия базисных компонентов ИТБ при анализе текста на уникальность;
- разработаны алгоритмы формирования информационно-терминологического базиса как совокупности лексически связанных компонентов;
- экспериментальное сопоставление нисходящего и восходящего алгоритмов формирования ИТБ как совокупности лексически связанных компонентов позволило выявить их преимущества и недостатки;
- реализация алгоритмов показала, что выбор наиболее эффективного из них во многом зависит от ситуации и требований, предъявляемых к базису при мультилингвистическом анализе уникальности текстов.

3 Реализация предложенной методики, алгоритмов и структур при мультилингвистическом анализе текстов на уникальность

3.1 ЛСК-методика и примеры ее реализации

Из ряда работ [54-65] следует, что появление методики обучения иностранной лексике посредством построения внутриязыковых ассоциативных полей (ЛСК-методики) повлекло за собой разработку новых алгоритмов и структур данных. Информационно-терминологическое обеспечение данной методики представляет собой совокупность лексически связанных компонентов (ЛС-компонентов). Такие компоненты содержат в себе одну основную лексему, несколько связанных лексем, соответствующее им число лексических сочетаний, а также частотные характеристики и языковые аналоги перечисленных элементов.

ЛСК-методика позволяет формировать при анализе текстов сложные строго организованные системы ассоциативных связей между лексемами одного (или нескольких) языков непосредственно в процессе анализа. Для ЛСК-методики характерна строгая организация материала, тщательная обработка текстов предметных областей, выявление нетривиальных зависимостей между терминами, а так же их частотных свойств, применение современных алгоритмов и структур данных, как в организации процесса анализа текста, так и в организации ИТБ на различных языках.

Несмотря на то, что ЛС-компоненты были разработаны специально для информационно-терминологической поддержки ЛСК-методики, они могут быть применимы в построении информационных систем и структур данных, которые эту методику «формально» не поддерживают; примером этого могут служить глобальные информационно-поисковые системы [58], используемые в современных сервисах анализа текстов на уникальность.

Рассмотрим применение ЛС-компонентов на различных примерах.

Мультилингвистический частотный словарь. Построение словаря – самый простой способ применения ИТБ, для этого не нужно ни обрабатывать данные, ни использовать дополнительные алгоритмы и программные средства. Необходимо лишь рациональным образом

организовать визуальное представление словаря и сформировать его структуру; в случае электронного словаря (ЭЧС) необходимо также обеспечить методы навигации. В случае ЛС-компонентов, в структуре словаря выделяются основные и связанные лексемы, упорядочивание терминов производится по алфавиту (для основных лексем) с указанием частоты.

Интерактивная система мультилингвистического анализа текста. Для создания интерактивной системы на основе ИТБ как совокупности ЛС-компонентов необходимо:

– разработать и реализовать программную оболочку системы;

– разработать сервисные механизмы, в том числе механизмы подачи материала и проведения проверочных мероприятий;

– внедрить в структуру ПО алгоритм на основе ЛС-компонентов;

– обеспечить средства взаимодействия с ИТБ.

Создание интерактивной системы включает в себя использование всех типов алгоритмов и моделей данных, разработанных для ЛСК-методики.

Мультилингвистическая поисковая система. Применение ЛС-компонентов в организации мультилингвистической поисковой системы дает базу для построения более точных алгоритмов поиска информации и анализа уникальности текстов [58].

Рассмотрим упрощенный пример работы такой системы. Информационно-терминологическая база системы представлена совокупностью мультилингвистических ИТБ на основе ЛС-компонентов нескольких предметных областей науки и техники.

Входные данные:

– запрос пользователя на родном ему языке, состоящий традиционно из ключевых слов и выражений;

– настраиваемые параметры поиска, такие как язык искомых текстов, которые будут подвергнуты анализу уникальности, и их предметная область.

Алгоритм работы ЛСК-методики:

1) согласно заданной предметной области система подключает ИТБ и производит поиск терминов запроса (по одному и в сочетаниях);

2) если термин попадает в один из ЛС-компонентов, то запрос перестраивается уже с учетом структуры этого компонента, таким образом, что текущий термин получает максимальный приоритет, например, 10, основная лексема – 5, остальные связанные лексемы – 2;

3) согласно установленных приоритетов и выбранного пользователем языка, производится процедура поиска и анализа текстов.

Выходные данные:

Список найденных текстов, упорядоченный согласно установленных системой приоритетов.

Таким образом, область искомых текстов сужается до указанной пользователем, независимо от выбранного языка. При этом для формирования запроса нет необходимости в точном знании иностранной терминологии.

3.2 Процедура построения частотного словаря на основе лексически связанных компонентов

Как бы стремительно ни развивалась наука и техника, основу анализа иностранной терминологии всегда будут составлять словари. Они видоизменяются, перекладываются на компьютерные системы, делятся по тематикам, но по-прежнему представляют собой не что иное, как совокупность терминов и их языковых аналогов. Что касается специализированных словарей, таких, например, как словарь по медицине, космонавтике или математическому моделированию, то они могут быть построены с помощью программных систем, анализирующих тексты. Как правило, в основе таких систем лежит частотный анализ терминов.

Словарь, полученный таким образом, перед тем как выйти в печать, проходит контроль со стороны лингвистов и специалистов конкретных предметных областей; чем более качественно реализована программная часть, тем более поверхностный характер несет их работа. Одним из таких словарей является разработанный и составленный в 2004 году англо-немецко-русский частотный словарь по системному анализу [33]. Этот словарь является одним из результатов развития МЛ-технологии и представляет собой одну из форм ее информационно-терминологического базиса.

Мультилингвистичность словаря делает возможным в процессе его использования построение однозначных ассоциативных связей между языковыми аналогами английского, немецкого и русского языка (при условии, что обучаемый или уже знает один из иностранных языков, или изучает два иностранных языка одновременно). Здесь и далее русский язык рассматривается в качестве оригинального.

Применение ЛСК-методики требует, чтобы словарь был организован как совокупность групп элементов на основании их частотных характеристик и лексических связей [41]. Процесс обучения может протекать как с учителем, так и с помощью программных средств.

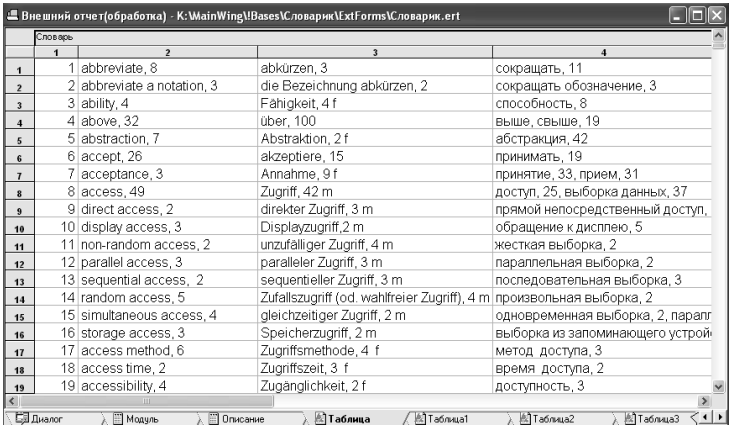
В рамках диссертационного исследования на основе предложенных алгоритмов и структур данных был разработан и реализован англо-немецко-русский словарь по системному анализу и информатике [36, 41]. В качестве первичной информационно-терминологической базы использовался ранее представленный словарь. Процедура построения включает следующие основные этапы:

1) получение и подготовка данных;
2) выбор и применение алгоритмов формирования ЛС-компонентов;
3) корректировка результатов и оформление словаря.

Получение необходимых данных для применения алгоритмов формирования ЛС-компонентов.

Для применения указанных алгоритмов необходимы следующие данные: абсолютные и относительные частоты лексем, данные о лексических связях, в том числе их частотные характеристики [41].

До обработки, словарь представляет собой таблицу, состоящую из терминов английского, немецкого и русского языков, так же определены частоты терминов, пример такой таблицы приведен на рисунке 3.1.


Внешний отчет(обработка) - K:\MainWing\Bases\Словарик\ExtForms\Словарик.ert

Словарь

	1	2	3	4
1	1	abbreviate, 8	abkürzen, 3	сокращать, 11
2	2	abbreviate a notation, 3	die Bezeichnung abkürzen, 2	сокращать обозначение, 3
3	3	ability, 4	Fähigkeit, 4 f	способность, 8
4	4	above, 32	über, 100	выше, свыше, 19
5	5	abstraction, 7	Abstraktion, 2 f	абстракция, 42
6	6	accept, 26	akzeptiere, 15	принимать, 19
7	7	acceptance, 3	Annahme, 9 f	принятие, 33, прием, 31
8	8	access, 49	Zugriff, 42 m	доступ, 25, выборка данных, 37
9	9	direct access, 2	direkter Zugriff, 3 m	прямой непосредственный доступ,
10	10	display access, 3	Displayzugriff,2 m	обращение к дисплею, 5
11	11	non-random access, 2	unzufälliger Zugriff, 4 m	жесткая выборка, 2
12	12	parallel access, 3	paralleler Zugriff, 3 m	параллельная выборка, 2
13	13	sequential access, 2	sequentieller Zugriff, 3 m	последовательная выборка, 3
14	14	random access, 5	Zufallszugriff (od. wahlfreier Zugriff), 4 m	произвольная выборка, 2
15	15	simultaneous access, 4	gleichzeitiger Zugriff, 2 m	одновременная выборка, 2, паралл
16	16	storage access, 3	Speicherzugriff, 2 m	выборка из запоминающего устрой
17	17	access method, 6	Zugriffsmethode, 4 f	метод доступа, 3
18	18	access time, 2	Zugriffszeit, 3 f	время доступа, 2
19	19	accessibility, 4	Zugänglichkeit, 2 f	доступность, 3

Диалог | Модуль | Описание | Таблица | Таблица1 | Таблица2 | Таблица3

Рисунок 3.1 – Частотный словарь до обработки

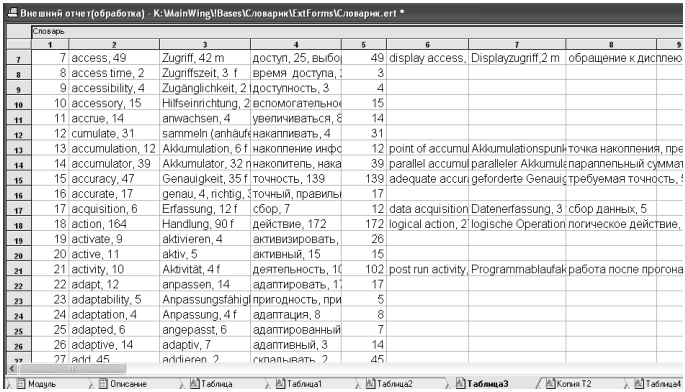
Для применения алгоритмов формирования ЛС-компонентов необходимо, чтобы каждый элемент словаря содержал в себе полную информацию о лексических связях в виде ссылки на связанные лексемы и частоты лексических сочетаний. Иначе говоря, информационный базисный компонент словаря необходимо привести к следующему виду:

ОМЛ-компонент = {термин яз_1, термин яз_2,..., термин яз_N, частота яз_1, частота яз_2,..., частота яз_N, сочетание 1_яз_1, сочетание 1_яз_2,..., сочетание 2_яз_1, сочетание 2_яз_2,…, сочетание K_яз_N, частота 1_яз_1, частота 1_яз_2,…, частота 2_яз_1, частота 2_яз_2,…, частота K_яз_N}.

Это возможно путем прогонки использованных при формировании словаря текстов через новые алгоритмы обработки. Но поскольку результаты работы программной обработки текстов при формировании первой версии словаря корректировались специалистами, такой подход не даст точных результатов; повторная экспертиза иностранных терминологий – задача достаточно трудоемкая и дорогостоящая, и так же не может гарантировать необходимой точности, если речь не идет о формировании словаря заново.

Необходимые данные можно получить без дополнительных затрат, проанализировав терминологический состав словаря. Здесь следует пояснить, что устойчивые лексические сочетания, которые мы ищем, отражают отдельные понятия языка, и наиболее востребованные из них вошли в состав словаря как отдельные элементы со своими частотами. Узкое место такого подхода состоит в том, что далеко не все термины составляющее лексические сочетания присутствуют в словаре. Однако частота этих терминов равна или, что скорее всего, превосходит частоту сочетаний, в которых они употребляются.

Это дает нам право пополнить словарь недостающими терминами и их языковыми аналогами. В рассматриваемом случае за основу были взяты английские термины, найден их адекватный перевод на другие языки, определены соответствующие частоты. На этом этапе словарь выглядел так как это показано на рисунке 3.2.



	1	2	3	4	5	6	7	8	9
7	7	access, 49	Zugriff, 42 m	доступ, 25, выбо	49	display access,	Displayzugriff,2 m	обращение к диспле	ю
8	8	access time, 2	Zugriffszeit, 3 f	время доступа,	3				
9	9	accessibility, 4	Zugänglichkeit, 2	доступность, 3	4				
10	10	accessory, 15	Hilfseinrichtung, 2	вспомогательное	15				
11	11	accrue, 14	anwachsen, 4	увеличиваться, 8	14				
12	12	cumulate, 31	sammeln (anhäufe	накапливать, 4	31				
13	13	accumulation, 12	Akkumulation, 6 f	накопление инфо	12	point of accumul	Akkumulationspunk	точка накопления, пре	
14	14	accumulator, 39	Akkumulator, 32 r	накопитель, нака	39	parallel accumul	paralleler Akkumula	параллельный суммат	
15	15	accuracy, 47	Genauigkeit, 35 f	точность, 139	139	adequate accur	geforderte Genauig	требуемая точность,	
16	16	accurate, 17	genau, 4, richtig,	точный, правильн	17				
17	17	acquisition, 6	Erfassung, 12 f	сбор, 7	12	data acquisition	Datenerfassung, 3	сбор данных, 5	
18	18	action, 164	Handlung, 90 f	действие, 172	172	logical action, 2	logische Operation	логическое действие,	
19	19	activate, 9	aktivieren, 4	активизировать,	26				
20	20	active, 11	aktiv, 5	активный, 15	15				
21	21	activity, 10	Aktivität, 4 f	деятельность, 1	102	post run activity,	Programmablaufak	работа после прогона	
22	22	adapt, 12	anpassen, 14	адаптировать, 1	17				
23	23	adaptability, 5	Anpassungsfähigl	пригодность, при	5				
24	24	adaptation, 4	Anpassung, 4 f	адаптация, 8	8				
25	25	adapted, 6	angepasst, 6	адаптированный	7				
26	26	adaptive, 14	adaptiv, 7	адаптивный, 3	14				
27	27	add, 45	addieren, 2	складывать, 2	45				

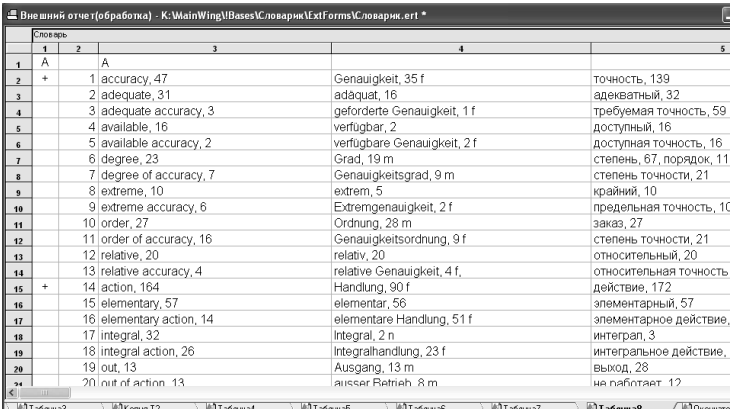
Модуль | Описание | Таблица | Таблица1 | Таблица2 | Таблица3 | Копия Т2 | Таблица4

Рисунок 3.2 – Словарь после этапа подготовки данных

Выбор и применение алгоритма формирования ЛС-компонентов.

Поскольку объем базиса невелик, а требования к количеству ЛС-компонентов отсутствуют, целесообразно применить восходящий алгоритм формирования ЛС-компонентов.

С помощью В-алгоритма было выявлено и сформировано 70 ЛС-компонентов, часть из них изображена на рисунке 3.3 (основные лексемы отмечены "+").



	1	2	3	4	5
1	A		A		
2	+	1	accuracy, 47	Genauigkeit, 35 f	точность, 139
3		2	adequate, 31	adäquat, 16	адекватный, 32
4		3	adequate accuracy, 3	geforderte Genauigkeit, 1 f	требуемая точность, 59
5		4	available, 16	verfügbar, 2	доступный, 16
6		5	available accuracy, 2	verfügbare Genauigkeit, 2 f	доступная точность, 16
7		6	degree, 23	Grad, 19 m	степень, 67, порядок, 11
8		7	degree of accuracy, 7	Genauigkeitsgrad, 9 m	степень точности, 21
9		8	extreme, 10	extrem, 5	крайний, 10
10		9	extreme accuracy, 6	Extremgenauigkeit, 2 f	предельная точность, 1
11		10	order, 27	Ordnung, 28 m	заказ, 27
12		11	order of accuracy, 16	Genauigkeitsordnung, 9 f	степень точности, 21
13		12	relative, 20	relativ, 20	относительный, 20
14		13	relative accuracy, 4	relative Genauigkeit, 4 f,	относительная точность
15	+	14	action, 164	Handlung, 90 f	действие, 172
16		15	elementary, 57	elementar, 56	элементарный, 57
17		16	elementary action, 14	elementare Handlung, 51 f	элементарное действие,
18		17	integral, 32	Integral, 2 n	интеграл, 3
19		18	integral action, 26	Integralhandlung, 23 f	интегральное действие,
20		19	out, 13	Ausgang, 13 m	выход, 28
21		20	out of action, 13	ausser Betrieb, 8 m	не работает, 12

Рисунок 3.3 – Словарь после обработки В-алгоритмом

Корректировка результатов и оформление словаря.

Таким образом, мы получили двухблочный словарь: первый блок состоит из ЛС-компонентов, второй – из элементов, которые не вошли в первый блок. Для удобства использования словаря следует выполнить сортировку блоков по алфавиту (ЛС-компоненты упорядочиваются по основным лексемам) и добавить соответствующие литеры [5].

Если словарь разработан для печати, то рационально добавить во второй блок терминологию первого блока со ссылками-номерами на соответствующие ЛС-компоненты. В таком случае, по второму блоку пользователь может производить поиск всей интересующей его терминологии, а по первому обучаться, согласно ЛСК-методики.

Так же рекомендуется провести корректировку терминологии в ручном режиме.

3.3 Двухблочная трехуровневая структура электронного англо-немецко-русского частотного словаря

Функциональное назначение. Данный программный продукт предназначен для исследования на уникальность переводных русскоязычных текстов, содержащих специальную терминологию английского и (или) немецкого языков в области системного анализа, информатики, вычислительной техники (с использованием межъязыковых ассоциативных связей). Также может быть использован в качестве справочного материала при переводе и анализе текстов как в рамках русского, так и английского или немецкого языков в отдельности (с использованием внутриязыковых ассоциативных связей) [36].

Структура словаря. Данный ЭЧС сформирован с помощью алгоритмов частотной обработки текстов и восходящего алгоритма формирования лексически-связанных компонентов. С помощью этих алгоритмов определена относительная частота лексем и лексических сочетаний, а также выявлено оптимальное количество и структура лексически связанных компонентов. ЭЧС состоит из двух блоков.

Блок лексически связанных компонентов. Первый блок содержит лексически связанные компоненты, которые представлены в виде двухуровневых структур данных, где на верхнем уровне находятся основные лексемы (крупный шрифт, зеленый фон), на нижнем –

связанные и сочетания (мелкий шрифт, песочный оттенок). На рисунке 3.4 изображен фрагмент блока лексически связанных компонентов [52].

1		**A**		
2	1	accuracy, 47	Genauigkeit, 35 f	точность, 139
15	14	action, 164	Handlung, 90 f	действие, 172
22	21	activity, 10	Aktivität, 4 f	деятельность, 102
25	24	adder, 51	Addierer, 47 m	сумматор, 127
38	37	address, 168	Adresse, 130 f	адрес, 168
51	50	algebra, 86	Algebra, 43 f	алгебра, 34
60	59	algorithm, 169	Algorithmus, 89 m	алгоритм, 126
73	72	analysis, 197	Analyse, 169 f	анализ, 327
74	73	comprehensive, 30	umfassend, 9	всесторонний, 30
75	74	comprehensive analysis, 9	umfassende Analyse, 7 f	комплексный анализ, 23
76	75	expert, 23	Experte, 3 m	эксперт, 55
77	76	expert analysis, 11	Expertenanalyse, 9 f	экспертный анализ, 26
78	77	factor, 68	Faktor, 44 m	коэффициент, 47, фактор, 31
79	78	factor analysis, 12	Faktoranalyse, 10 f	факторный анализ, 4
80	79	job, 50	Arbeit, 25 f	работа, 179
81	80	job analysis, 3	Arbeitsanalyse, 9 f	анализ работы, 13
82	81	network, 36	Netzwerk, 13 n	сеть, 9
83	82	network analysis, 9	Netzwerkanalyse, 2 f	сетевой анализ, 14
84	83	qualitative, 2	qualitativ, 6	качественный, 33
85	84	qualitative analysis, 8	qualitative Analyse, 8 f	качественный анализ, 19
86	85	approach, 70	Herangehensweise, 8 f, Methode, 24 f	подход, 60

Рисунок 3.4 – Блок лексически связанных компонентов

Терминологический блок. Второй блок словаря содержит всю терминологию, включая лексемы, вошедшие в лексически связанные компоненты; они выделены курсивом и формально представляют собой ссылки на соответствующие элементы первого блока (поиск осуществляется по нумерации) [52]. Фрагмент терминологического блока изображен на рисунке 3.5.

2772	2186	vulgar fraction, 2	gewöhnlicher Bruch, 2 m	простая дробь, 2
2773	2187	vulgar, 2	vulgär, 2	вульгарный, 2
2774		***W***		
2775	***540***	***waiting, 8***	***Warten, 8 n***	***ожидание, 8***
2776	***541***	***waiting time, 8***	***Wartezeit, 2 f***	***время ожидания, 6***
2777	2188	wave function, 2	Wellenfunktion, 2 f	волновая функция, 2
2778	2189	wave, 12	Welle, 7 f	волна, 9
2779	2190	welfare, 3	Wohl, 6 n	благополучие, 2
2780	2191	white noise, 8	weißes Rauschen, 2 n	белый шум, 3
2781	2192	white, 8	weiß, 2	белый, 3
2782	2193	whole, 15	ganz, 9	целое, 69
2783	2194	wholeness, 6	Ganzheit, 2 f	целостность, 23
2784	***48***	***word, 22***	***Wort, 3 n***	***слово, 22***
2785	***49***	***word address, 22***	***Wortadresse, 3 f***	***адрес слова, 2***
2786	2195	work file, 3	Arbeitsdatei, 3 f	рабочий файл, 2
2787	2196	work, 3	Arbeit, 24	работа, 24
2788	2197	working space, 3	Arbeitsraum des Speichers, 2 m	рабочая область памяти, 33
2789	***127***	***working, 9***	***Arbeiten, 9 n***	***работа, 24***
2790	***128***	***working cell, 3***	***Arbeitszelle, 6 f***	***рабочая ячейка, 4***
2791	2198	world, 7	Welt, 3 f	мир, 111
2792	2199	write operation, 4	Schreiboperation, 3 f	операция записи, 4
2793	***542***	***write, 12***	***schreiben, 12***	***написать, 12***
2794	***543***	***write time, 12***	***Schreibzeit, 4 f***	***время записи, 9***
2795		***X***		
2796	2200	X-axis, 4	Abszissenachse, 2 f	ось абсцисс, 2
2797		***Y***		
2798	2201	Y-axis, 4	Ordinatenachse, 2 f	ось ординат, 2
2799		***Z***		
2800	***275***	***zero, 65***	***Null, 4 f***	***нуль, 2***
2801	***276***	***zero level, 15***	***Nullniveau, 10 n***	***нулевой уровень, 14***
2802	2202	zonal, 15	zonal, 1	зональный, 1
2803	2203	zone, 41	Zone, 4 f	зона, 7
2804	2204	zoom, 12	zoomen, 20 n	масштабировать, 6

Рисунок 3.5 – Терминологический блок

Терминологический блок предназначен в качестве справочного материала для анализа уникальности текстов иностранных терминологий, позволяет реализовать мультилингвистический подход.

Отметим, что оба блока упорядочены по алфавиту, добавлены литеры. Группировка осуществлена средствами excel. Для первого блока группировка является трехуровневой (литера – основная лексема – связанные лексемы), для второго – двух (литера – основная лексема).

Базовый подхода к терминологическому анализу текста. Структура словаря делает возможным применение мультилингвистического подхода и методики анализа текста на уникальность на основе лексически связанных компонентов. Что в свою очередь обеспечивает систематизацию текстового материала и формирование строго организованных систем ассоциативных зависимостей как внутри одного языка, так и между различными языковыми аналогами непосредственно в процессе анализа уникальности текстов.

Мультилингвистический подход к анализу уникальности текстов. Что касается мультилингвистического подхода, то для его применения необходимо, чтобы частотные словари были синхронизированы для двух/трех и более языков одновременно. В этом случае дополнительный иностранный язык представляет собой множество терминов-аналогов, что необходимо для установления однозначного соответствия между языковыми аналогами представленных для анализа уникальности многоязычных текстов.

Методика на основе лексически связанных компонентов. Методика строится на терминологии первого блока словаря. Последовательность анализа текстового материала согласно данной методике такова:

– основная лексема – перевод, аналог на иностранном языке (в рамках МЛ-технологии);

– связанная лексема – перевод, аналоги на иностранном языке (в рамках МЛ-технологии);

– лексическое сочетание основной и связанной лексем – перевод сочетания, аналоги на иностранном языке (подчеркнем, что языковой аналог именно лексического сочетания, но не лексем по отдельности);

– переход к следующей связанной лексеме;

– переход к следующему лексически связанному компоненту.

Особенность подхода заключается в формировании ассоциативных связей внутри анализируемой терминологии, причем большинство связей приходится на основные лексемы, специально отобранные как наиболее значимые для изучения. Таким образом, уменьшается

скорость забывания наиболее значимой для изучения иностранной терминологии, что повышает качество обучения в целом.

Дополнительно. Навигация осуществляется средствами MS Excel. Словарь доступен для правки специалистом-переводчиком (для этого необходимо в свойствах файла снять атрибут "только чтение").

Относительно предыдущей версии словаря [33], терминологическая база расширена (около 500 специальных терминов), полностью изменена структура словаря (два блока и группировки элементов), добавлены функции навигации. Словарь поддерживает ЛСК-методику.

Реализация предложенной методики, алгоритмов и структур данных в интерактивных компьютерных системах анализа уникальности текстов основывается на следующих основных выводах и результатах:

1) область применения лексически связанных компонентов как способа организации информационно-терминологической базы включает в себя решение задач о построении электронных и составлении бумажных словарей, разработке интерактивных поисковых систем и систем анализа текста;

2) информационно-терминологический базис мультилингвистической адаптивно-обучающей технологии может быть дополнен и преобразован для использования в рамках методики анализа уникальности текстов на основе лексически связанных компонентов;

3) описана процедура и основные моменты построения мультилингвистического частотного словаря на основе лексически связанных компонентов;

4) разработанный двухблочный трехуровневый электронный англо-немецко-русский частотный словарь по информатике и системному анализу отражает предложенный подход и состоит из двух специальных блоков, что обеспечивает возможность применения ЛСК-методики.

4. Мультилингвистическая модель запросов ЛПР в распределенной информационной системе

В данном разделе представлены процедуры, позволяющие сформировать модель запросов ЛПР, а также профиль ЛПР, что имеет существенное значение при учете производственных факторов, влияющих на процесс принятия решения, анализе последствий, к которым может привести та или иная ситуация (то или иное решение) и т.д. Очевидно, что сформированный профиль ЛПР отражается на выборе рациональной стратегии поведения в сложных производственных ситуациях, зависит не столько от количественных соотношений, сколько от качественных формулировок. Выбор категории или профиля ЛПР зачастую определяется набором типовых ситуаций, которые возникают в системе управления только в том виде, в котором он был зафиксирован на практике.

При формировании профиля ЛПР учитывается ряд атрибутов, характеризующий различные производственные ситуации, возникающие в процессе функционирования промышленного предприятия:

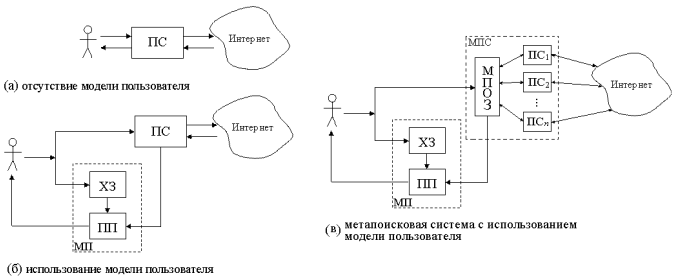
- наименование ситуации;
- описание причин ее возникновения;
- указание последствий, к которым она может привести;
- временные характеристики, определяющие время возникновения и окончания ситуации;
- наименование структурных подразделений предприятия, где возникла производственная ситуация;
- описание мероприятий, необходимых для ее успешного разрешения;
- указание производственных факторов, влияющих на процесс принятия решений;
- список документов, регламентирующих поведение ЛПР в сложившейся ситуации;
- перечень данных, поступающих от различных информационно-измерительных систем предприятия и используемых ЛПР в процессе выработки решения.

Необходимо отметить, что на ряде промышленных предприятий некоторые из указанных выше характеристик производственных ситуаций могут не использоваться управленческим персоналом в

процессе принятия решения. Кроме того, ряд характеристик производственных ситуаций может оставаться не определенным экспертами в течение значительного промежутка времени. В этих случаях в разрабатываемой поисковой системе при формировании профиля ЛПР будет учитываться только известные (базовые) характеристики и не будут рассматриваться остальные.

В работе моделирование ЛПР строится на основе его информационных потребностей в конкретной производственной ситуации, поскольку основной задачей управляющего персонала является поиск необходимой информации для принятия решения. Представлена методология латентно-семантического анализа (ЛСА), преимущества ее применения при поиске информации. Показан переход к вероятностному ЛСА, применение для моделирования ЛПР и формирования алгоритма непрерывной корректировки параметров модели при работе ЛПР с системой. Тем самым, достигается индивидуализация (или, другими словами, персонализация) интерфейса ЛПР в ИУС, что позволяет обеспечивать информационную поддержку принятия решения в контексте текущей производственной ситуации. То есть, еще на стадии обработки результатов поискового запроса отсеивается большая часть не релевантных документов и данных.

В работе рассматривается новая схема моделирования ЛПР, основанная на инициализации начального профиля и его последовательной корректировке в процессе работы (рисунок 4.1).



Примечания: ПС – поисковый сервис, МП – модель пользователя, ХЗ – хранилище запросов, ПП – профиль пользователя, МПС – метапоисковая система, МПОЗ – модуль предварительной обработки запросов.

Рис. 4.1 - Подходы к организации поисковых систем.

При этом документы могут быть представлены как векторы латентного семантического пространства. Итак, при создании

поисковых систем в ИУС важную роль играет модель ЛПР. От того насколько правильно система сможет судить о том, кто с ней в данный момент в данной производственной ситуации работает, зависит правильность работы (адаптации) всей системы в целом.

В работе [13] сущность и цель модели студента определяется как “представление некоторых характеристик обучающихся, которые (характеристики) полезны для достижения адекватного индивидуального взаимодействия между компьютерным приложением и студентом”. Заменив слово «студент» словом «пользователь», мы получим определение модели пользователя применимое к АГ системам.

Моделирование пользователя при организации адаптивных поисковых систем является достаточно сложной задачей. До настоящего времени не существует единых подходов, методов или стандартов разработки таких моделей.

- Моделью называется некий объект-заместитель, который в определенных условиях может заменять объект-оригинал, воспроизводя интересующие исследователя свойства и характеристики оригинала, причем существенные преимущества удобства [12].
- В контексте данной работы, объектом-заместителем является пользователь, который рассматривается в виде совокупности некоторых характеристик.
- Моделирование пользователя – это процесс, охватывающий весь жизненный цикл существования модели пользователя, от сбора и обработки информации о пользователе до непосредственного формирования модели, ее корректировки и подстройки.

Какую именно информацию будет содержать модель пользователя, зависит от множества прикладных задач, для решения которых проектируется данная гипермедиа-система. Как уже было неоднократно сказано, модель пользователя – это представление важных характеристик пользователя (в контексте формирующейся системы) в терминах данной системы. Первоначальные сведения в модель закладывает разработчик системы, который имеет некоторые представления как о будущей аудитории, так и о предметной области, для которой система разрабатывается.

Нужно понимать, что использование достаточно сложных моделей в адаптивных гипермедиа-системах не всегда целесообразно, так как связано с большими затратами на формирование, проверку адекватности и последующую эксплуатацию. При этом даже простые модели, являющиеся грубой оценкой некоторых характеристик пользователя, позволяют в значительной мере повысить функциональность и удобства использования ГС. “Создавая модель

пользователя нужно понимать, что она никогда не будет идеальным его отражением, а скорее некоторой грубой аппроксимацией его основных характеристик” [23].

Моделирование пользователя может помочь при решении следующих задач, связанных с гипермедиа-системами:

- Помощь пользователю при изучении определенной тематики.
- Предоставление важной для данного пользователя информации.
- Адаптация интерфейса под пользователя.
- Помощь пользователю в поиске информации.
- Поддержка совместной работы.
- Помощь пользователю при работе с самой системой.

4.1 Характеристики модели запросов ЛПР

В настоящее время не существует общепринятого определения понятия «модель запросов», «модель ЛПР» или «модель пользователя». Но можно выделить два основных подхода к построению такого класса моделей.

1. Под моделью пользователя (ЛПР) понимают набор характеристик (параметров) и совокупность правил, которые на основании значений этих характеристик управляют процессом общения системы с пользователем [17].
2. В других классах систем под моделью пользователя обычно понимают набор параметров, измеряемых во время работы системы с пользователем и определяющих степень усвоения им информации по рассматриваемой предметной области.

На самом деле, понимание модели пользователя как набора параметров – это слишком узкое понимание. Без учета методов, которые работают с данным набором характеристик, этот набор теряет смысл. Следовательно, модель пользователя можно определить как совокупность набора характеристик пользователя и методов (правил) обработки этого набора.

В гипермедиа-системах эти правила привязаны к конкретной предметной области, что позволяет более качественно имитировать общение пользователя со специалистом в данной предметной области. Адаптивная гипермедиа-система, в общем случае, не должна быть ориентирована на конкретную предметную область, поэтому и

назначение правил, входящих в модель пользователя, должно быть другим. В первую очередь эти правила должны проводить изменения самой модели пользователя по результатам его работы с системой. Это позволит разработчику системы управлять формированием этой модели, т.е. осуществлять алгоритмическую настройку без программирования.

Модель пользователя играет важную роль при создании адаптивных гипермедиа-систем. Информация, содержащаяся в модели пользователя, отражает следующие основные его характеристики [25]:

- Цели.
- Знания.
- Предпочтения.
- Уровень подготовки и имеющийся опыт.

Цель пользователя или задача пользователя – это параметр, зависящий, в большей степени, от самой природы работы пользователя в гипермедиа, нежели от пользователя как такового. В зависимости от типа системы, это может быть рабочая цель (в прикладных системах), цель поиска (в информационно-поисковых системах), и цель обучения или решения (в обучающих системах). Во всех этих случаях цель является ответом на вопрос «Почему пользователь использует гипермедиа-систему и чего пользователь хочет в итоге достичь?». Цель пользователя – это наиболее изменчивая характеристика пользователя. Она почти всегда будет другой при новом сеансе работы с системой, а иногда может изменяться и во время одного сеанса работы. В некоторых системах целесообразно различать локальные, или цели нижнего уровня, которые могут изменяться достаточно часто и общие, или цели (задачи) высокого уровня, являющиеся более стабильными. Например, в обучающих системах цель обучения – это цель высокого уровня, в то время как цель решения задачи – цель нижнего уровня, которая может изменяться от одной учебной задачи к другой несколько раз в течение одного сеанса работы. Цель пользователя может рассматриваться как очень важная характеристика пользователя в адаптивных гипермедиа-системах. Почти третья часть существующих технических приемов адаптации основывается на этом. Интересно, что почти все эти приемы относятся к техническим приемам адаптивной поддержки навигации.

Одной из важных характеристик пользователя, рассматриваемой системами адаптивной гипермедиа, является *набор (система) предпочтений пользователя*. По различным причинам пользователь может предпочитать некоторые узлы, ссылки и части страниц

гипермедиа другим. Эти предпочтения могут быть абсолютными или относительными, то есть, зависящими от текущего узла, цели или текущего контекста вообще. Предпочтения наиболее тяжело использовать в поисковой гипермедиа. В то же самое время, в наиболее адаптивных поисковых гипермедиа-системах предпочтения – единственная накапливаемая информация о пользователе.

Предпочтения отличаются от других компонентов модели пользователя по нескольким аспектам. В отличие от других компонентов, предпочтения не могут быть логически выведены системой. Пользователь должен непосредственно или косвенно (простая обратная связь) сообщить системе свои предпочтения. Еще одной специфической особенностью моделирования предпочтений является способ представления. В то время как другие компоненты модели пользователя обычно представляются в символьном виде, предпочтения зачастую представляются и вычисляются в численной форме с помощью особых методов [36]. Численная форма представления имеет некоторые преимущества перед символьной: она дает возможность комбинировать несколько моделей пользователя и накапливать *модель группы пользователей*. Модель группы отражает предпочтения определенной группы пользователей (таких, как научно-исследовательские лаборатории). Групповая модель – хорошая начальная модель для нового члена группы. Групповые модели имеют большое значение для совместной работы (тяжело наладить сотрудничество, когда сотрудники используют индивидуальные модели пользователя и, таким образом, имеют различные «точки зрения» на одну и ту же тему).

Следующие две характеристики пользователя имеют нечто общее с уровнем знаний пользователя, но функционально отличаются от него. Это – уровень подготовки пользователя и имеющийся опыт работы пользователя с данной гипермедиа. Под *уровнем подготовки пользователя* мы понимаем всю информацию, связанную с предыдущим опытом работы пользователя, не относящуюся к теме данной гипермедиа-системы. Подготовка включает в себя профессию пользователя, опыт работы в смежных областях, а также точку зрения пользователя и его перспективу. Некоторые системы включают уровень подготовки пользователя в модель пользователя и применяют этот параметр для адаптивного представления материала, другие применяет его для адаптивной поддержки навигации.

Опытом работы пользователя в данной гипермедиа мы называем характеристику, отражающую то, насколько хорошо пользователь знаком со структурой гипермедиа и то, насколько легко он может

осуществлять навигацию в данной гипермедиа. Это – не то же самое, что уровень знаний пользователя по теме, представленной в гипермедиа [41]. Иногда бывает так, что пользователь хорошо знаком с темой, но абсолютно не знаком со структурой гиперпространства. И наоборот, пользователь может быть хорошо знаком со структурой гиперпространства, но не иметь глубоких знаний по теме, представленной в нем. Еще одной причиной того, что мы должны различать опыт работы в гипермедиа и уровень знаний пользователя, является существование технического приема адаптивной поддержки навигации [39], который основан на этой характеристике пользователя.

4.2 Частотная терминологическая модель запросов ЛПР

Индивидуализация (или, другими словами, персонализация) интерфейса пользователя, благодаря алгоритмам идентификации пользователя, позволяет учитывать неявные интересы пользователя и использовать их в контексте текущего запроса. Тем самым еще на стадии обработки результатов запроса отсеивается большая часть не релевантных документов.

Из предыдущих параграфов становится ясно, что в настоящее время использованию моделей пользователя в гипермедиа-системах уделяется огромное внимание со стороны исследователей. Однако на текущий момент не предложено эффективных моделей, позволяющих описывать пользователя в режиме реального времени, а тем более производить корректировку такой модели в соответствии с новой информацией или изменением состояния окружения гипермедиа-системы.

Любая гипермедиа-система – это, прежде всего, информационная система, т.е. система, представляющая информацию по некоторой предметной области в удобном для пользователя виде. Удобство представления обеспечивается введением, механизма ссылок и, наряду с текстовой информацией, мультимедиа-элементов в узлы гипермедиа-системы. В гипермедиа-системах выделяют два основных способа поиска информации. Во-первых, это навигация по ссылкам, т.е. перемещение от одного узла системы к другому. Во-вторых, это поисковые запросы, т.е. описание необходимой информации в виде строки запроса и активация механизма поиска. В этом случае в ответ на запрос может быть выдана совокупность страниц.

В данном параграфе представлен алгоритм непрерывной корректировки модели пользователя на основе текущих запросов с использованием методологии PLSA [33]. Один из распространенных подходов к представлению документов (и запросов) при извлечении информации из сети основан на понятии модели векторного гиперпространства [39].Описанное векторное пространство при использовании методологии латентной семантической индексации заменяется представлением документа в латентном пространстве меньшей размерности [29]. В данной работе предлагается расширить понятие латентного семантического пространства с учетом текущих интересов пользователя. Поскольку со временем интересы могут меняться, то нужно предусмотреть возможность изменения (уменьшения или увеличения) важности интересов пользователя. В предлагаемой модели вводится понятие временного измерения в латентном семантическом пространстве. Результирующее пространство было названо временным латентным семантическим пространством. Оно служит для отслеживания динамики изменения интересов (профиля) пользователя с течением времени. Координаты документа и запроса в новом латентном семантическом пространстве рассчитываются аналогично схеме, предложенной Томасом Хофманом (T.Hofmann) в работе [33]. Различие заключается лишь в том, что запросы имеют временное измерение (текущий вес), начальное значение которого задается положительными величинами, убывающими с течением времени (таким образом, реализовано предположение об изменении интересов со временем).

Проблема извлечения информации в гипермедиа-системах получила новое развитие в связи с появлением всемирной сети Интернет. В настоящее время каждый пользователь, имеющий доступ к Интернет, имеет доступ ко всем источникам информации, представленным в ней. Казалось бы, что теперь со своевременным получением необходимой информации по интересующей тематике не должно возникнуть больших проблем: ввел запрос поисковой машине и жди ответа в виде набора ссылок на интересующие документы. Однако на деле оказалось, что качество поиска информации при всей ее доступности очень низкое. В поисковых машинах (иначе их называют поисковыми сервисами) отсутствуют эффективные алгоритмы поиска релевантной информации (набора релевантных документов, отражающих суть запроса). И в ответ на запрос такой сервис может выдать сколь угодно большое количество документов либо отдаленно отражающих сферу интересов пользователя, либо вовсе не имеющих никакой связи с сутью запроса.

Среди исследователей можно выделить приверженцев двух идей. С одной стороны – традиционная лингвистическая школа, пытающаяся научить компьютер естественному языку. С другой – сообщество, ориентированное на использование статистических методов. Подход PLSA относится ко второй группе.

Первоначально было введено понятие модели векторного пространства [29]. При этом любой документ представлялся как вектор частот появления определенных терминов в нем. В этом подходе отношения между документами и терминами представлялись в виде матрицы смежности $\boldsymbol{A}$, элементом w_{ij} которой является частота появления термина t_j в документе d_i. Обозначим через m количество проиндексированных терминов в коллекции документов d, а через n – количество самих документов. В общем случае элементом w_{ij} матрицы $\boldsymbol{A}$ является некоторый вес, поставленный в соответствие паре документ-термин (d_i, t_j). После того, как все веса заданы, матрица $\boldsymbol{A}$, становится отображением коллекции документов в векторном гиперпространстве. Таким образом, каждый документ можно представить, как вектор весов терминов:

$$\boldsymbol{A} = \begin{pmatrix} w_{11} & \bullet & \bullet & w_{1n} \\ \bullet & \bullet & & \bullet \\ \bullet & & \bullet & \bullet \\ w_{m1} & \bullet & \bullet & w_{mn} \end{pmatrix} \equiv \begin{pmatrix} d_1 & \bullet & \bullet & d_n \end{pmatrix} \equiv \begin{pmatrix} t_1 \\ \bullet \\ \bullet \\ t_m \end{pmatrix}$$

Подход LSA (Latent Semantic Analysis – Латентный Семантический Анализ), предложенный в 1990 в работе [29], заключается в отображении документа в латентное семантическое пространство. Основная цель – отразить скрытую (латентную) связь между терминами и документами. Это достигается использованием сингулярного разложения (***SVD***-разложение) матрицы $\boldsymbol{A}$. Предполагается, что такое пространство несет в себе основную смысловую нагрузку. Оценка схожести документов формируется по близости расположения точек латентного семантического пространства.

В основе методологии PLSA лежит идея, предложенная в LSA и описанная выше [33]. В PLSA на латентном семантическом пространстве вводится понятие латентного класса

$$z \in \boldsymbol{Z} = \{z_1, \ldots, z_k\}.$$

А также рассматривается условные вероятности среди документов

$$d \in \boldsymbol{D} = \{d_1, \ldots, d_k\},$$

и терминов

$$w \in \boldsymbol{W} = \{w_1, \ldots, w_k\}.$$

Далее предполагается, что распределение слов, принадлежащих данному классу, не зависит от документа и пары наблюдений документ-термин (d, w) – независимы.

Распределение терминов в документе $P(w|d)$ определяется выпуклой комбинацией факторов $P(w|z)$ и $P(z|d)$ и записывается следующим образом:

$$P(w \mid d) = \sum_{z \in Z} P(w \mid z) P(z \mid d)$$

Совместная вероятность документа и термина определяется следующим соотношением:

$$P(d, w) = P(d) P(w \mid d) = \sum_{z \in Z} P(z) P(d \mid z) P(w \mid z)$$

Используя алгоритм максимизации математического ожидания (Expectation-Maximization, EM algorithm), который состоит из двух этапов (E и M), оцениваются вероятности $P(w|z)$ и $P(z|d)$, максимизируя логарифмическую функцию правдоподобия:

$$L = \sum_{d \in D} \sum_{w \in W} n(d, w) \log P(d, w)$$, где

$n(d, w)$ – частота термина в документе (количество появлений термина w в документе d).

Вероятность того, что появление термина w в документе d объясняется принадлежностью их к классу z оценивается следующим образом на шаге E:

$$P(z \mid d, w) = \frac{P(z) P(d \mid z) P(w \mid z)}{\sum_{z' \in Z} P(z') P(d \mid z') P(w \mid z')}.$$

На этапе M происходит переоценка вероятностей:

$$P(w \mid z) = \frac{\sum_{d} n(d, w) P(z \mid d, w)}{\sum_{w'} \sum_{d} n(d, w') P(z \mid d, w')},$$

$$P(d \mid z) = \frac{\sum_{w} n(d, w) P(z \mid d, w)}{\sum_{d', w} n(d', w) P(z \mid d', w)},$$

$$P(z \mid d,w) = \frac{\sum_{d,w} n(d,w) P(z \mid d,w)}{\sum_{d,w} n(d,w)}.$$

Т. Хофман в работе [33] предложил обобщенную модель для оценивания условной вероятности, которую он назвал ослабленной процедурой максимизации математического ожидания (TEM – Tempered Expectation Maximization). При этом на этапе E в оценку условной вероятности вносится регуляризационный параметр β:

$$P_\beta(z \mid d,w) = \frac{P(z)\left[P(d \mid z)P(w \mid z)\right]^\beta}{\sum_{z' \in Z} P(z')\left[P(d \mid z')P(w \mid z')\right]^\beta}$$

Согласно (4.2) любая условная вероятность $P(w|d)$ может быть аппроксимирована полиномом, представляющим собой выпуклую комбинацию условных вероятностей $P(w|z)$. Геометрическая интерпретация весовых коэффициентов $P(z|d)$ – координаты документа в подпространстве, определяемом как латентное семантическое пространство [33].

Рассмотрим новую схему моделирования интересов пользователя, основанная на инициализации начального профиля и его последовательной корректировке в процессе работы.

Документы могут быть представлены как векторы латентного семантического пространства так, как это показано в предыдущем пункте, за более подробным описанием обращайтесь к [33, 34]. Для того чтобы следить и непрерывно анализировать возможные изменения интересов пользователя предлагается ввести понятие временного измерения в латентном семантическом пространстве, тем самым рассматривая уже не само латентное семантическое пространство, а его модификацию – временное латентное семантическое пространство. Каждое измерение (за исключением временного) такого векторного пространства, представляет собой условные вероятности при заданном классе $P(\bullet| z)$. Документы представляют собой векторы с весовыми коэффициентами (координатами) $P(z|d)$, временное измерение полагаем равным нулю.

Запросы, равно как и сами документы, могут быть представлены в виде векторов во временном латентном семантическом пространстве. Кроме весов $P(z| Q)$ у них есть дополнительное (временное) измерение (текущий вес), первоначально равный некоторой положительной величине, уменьшающейся с течением времени, исходя из

предположения о падении интереса пользователя к определенной тематике при отсутствии ее фигурирования в запросах продолжительное время. Если пользователь инициирует запрос, связанный с определенной категорией из его текущего профиля, то вес данной категории может быть либо стабилизирован на определенное время, либо увеличен.

Согласно геометрии латентного семантического пространства запрос, состоящий из терминов, проецируется в латентное семантическое пространство [33]. Таким образом, гиперповерхность S_i, образованная запросом Q_i является пересечением вероятностных поверхностей всех классов, введенных на латентном семантическом пространстве, в которых с определенной вероятностью фигурирует данный термин:

$$S_i = \bigcap_k H_{ki}$$

Алгоритм адаптивной коррекции профиля пользователя основан на неявной обратной связи с пользователем, которая реализуется на основе истории его запросов. На вход алгоритма поступает запрос пользователя, на выходе – одна или более троек (триплетов) вида (C_i, W_i, α_i), где C_i – категория интересов, W_i – текущий вес, α_i – уровень изменчивости (смысл данной величины состоит в том, чтобы отразить насколько изменяются интересы пользователя в рамках текущего запроса по отношению к прошлым запросам).

Итак, профиль пользователя представляет собой набор троек. При этом он организован таким образом, что интересы пользователя разделены на два типа: краткосрочные (*краткосрочный профиль*) и долгосрочные (*долгосрочный профиль*). Как правило, емкость долгосрочного профиля больше емкости краткосрочного. Структуру профиля можно представить таблицей, так как это показано на рисунке 4.2.

кино	музыка	квантовая физика	спорт
95	85	35	70
0.60	0.45	0.20	0.15

категория

текущий вес

уровень изменчивости

Рисунок 4.2 - Краткосрочный профиль пользователя.

При этом считается, что тройки, в которых величина текущего веса положительная, относятся к краткосрочному профилю, если вес отрицательный – то к долгосрочному профилю. При этом для троек, находящихся в краткосрочном профиле, текущий вес уменьшается линейно, тогда как для троек, находящихся в долгосрочном профиле снижение весов – экспоненциально.

Формально профиль в текущий момент i описывается следующим образом

$$Pr_i = \{(C_j, W_j, \alpha_j)_i,\ j=1, k\}$$

При этом,

$$Pr_i = PrR_i \cup PrL_i,$$

где

$PrR_i = \{(C_j, W_j, \alpha_j)_i \mid \forall\ W_j \geq 0,\ j=1, k\}$ – краткосрочный профиль,

$PrL_i = \{(C_j, W_j, \alpha_j)_i \mid \forall\ W_j < 0,\ j=1, k\}$ – долгосрочный профиль.

Уровень изменчивости (α_i) рассчитывается как близость двух последовательных запросов Q_i и Q_{i-1} представленных в пространстве частот их терминов:

$$\alpha_i = \frac{\sum_w \tilde{n}(Q_i, w)\tilde{n}(Q_{i-1}, w)}{\sqrt{\sum_{w'} \tilde{n}(Q_i, w)^2 \sum_d \tilde{n}(Q_{i-1}, w)^2}},$$

где $\tilde{n}(Q_i, w)$ – взвешенные частоты терминов.

При использовании данного алгоритма предполагается, что существует некоторое хранилище предыдущих запросов пользователя. В текущий момент времени i пользователь вводит новый запрос, который после соответствующей обработки помещается в хранилище запросов. Обновленное (или дополненное) в момент времени i текущим запросом хранилище запросов будем обозначать Q_i .

Перед тем, как передать запрос для работы алгоритму производится его обработка на предмет выделения ключевых терминов.

Далее производится пересчет взвешенных частот терминов в хранилище запросов Q_i с учетом нового запроса. Когда пользователь вводит очередной запрос, ключевым словам (терминам) данного запроса назначаются наибольшие веса. При поступлении запроса в хранилище запросов происходит проверка на наличие в хранилище терминов, присущих текущему запросу. Если термин встречается впервые, то при его занесении в хранилище вес остается без изменений, если же такой термин уже существует в хранилище (это означает, что пользователь уже когда-то использовал запрос, включающий данный термин), то производится пересчет весового коэффициента данного

термина. В конечном счете, происходит нормирование весовых коэффициентов.

Категории интересов C_i для включения в текущий профиль извлекаются из хранилища посредством использования методологии PLSA.

Далее представлен пошаговый алгоритм непрерывной корректировки профиля пользователя.

1. Инициализировать хранилище запросов $Q_i = \{w_{1i}, w_{2i}, \dots, w_{ki}\}$, где w_{ki} – термины хранилища запросов, $k = 1 \dots M$.
2. Выделить набор ключевых терминов текущего запроса.
3. Скорректировать весовые коэффициенты терминов и произвести их нормировку с учетом нового запроса.
4. Рассчитать уровень изменчивости α_i.
5. Рассчитать условные вероятности классов, используя процедуру TEM следующим образом

$$P(z \mid Q_i) = P(w_{ki}) P_\beta(z \mid Q_i, w_{ki}) = \sum_{w_{ki}} P(w_{ki}) \frac{P(z)\left[P(Q_i \mid z) P(w_{ki} \mid z)\right]^\beta}{\sum_{z'} P(z')\left[P(Q_i \mid z') P(w_{ki} \mid z')\right]^\beta}$$

6. Рассчитать вероятность категории C_i для заданного класса латентного семантического пространства,

$$P(C_i \mid z) = \frac{\sum_{Q_i} n(Q_i, C_i,) P_\beta(z \mid Q_i, C_i)}{\sum_{C_i', Q_i'} n(C_i', Q_i) P_\beta(z \mid C_i', Q_i)}.$$

7. Рассчитать вероятность включения категории C_i для текущего состояния хранилища запросов Q_i ,
8. Занести категорию в профиль пользователя. Для этого включить соответствующую тройку (C_i, W_i, α_i) в профиль, согласно схеме представленной на рисунке 2.1.
9. Если уровень изменчивости $\alpha_i > \alpha_0$ (где α_0 заданная величина), то увеличить текущий вес категории C_i на величину

$$\Delta W_i : W_i = W_i + \Delta W_i.$$

10. Отсортировать последовательность троек (C_i, W_i, α_i) в профиле по порядку убывания веса W_i.
11. Сохранить получившийся профиль как текущий.

4.3 Экспериментальные исследования

Одним из широко распространенных поводов для критических замечаний для исследований в области информационного поиска является использование относительно небольших тестовых наборов данных, не отражающих ситуацию в реальных системах. Более того, очень сложно сравнивать эффективность разных методов решения задач информационного поиска из-за того, что, для их экспериментальной проверки используются разные тестовые наборы данных [55].

Для решения этой проблемы в течение последнего десятка лет был создан ряд стандартных тестовых наборов данных, которые в настоящее время повсеместно используются для проведения экспериментов в области информационного поиска. Для исследования описанного алгоритма использовалось 4 набора документов:

- Doc_a001 – 1033 документа;
- Doc_a002 – 1400 документа;
- Doc_a003 – 3204 документа;
- Doc_a004 – 1460 документа.

В качестве критериев качества подходов к решению задач информационного поиска используются точность (*Precision*) и полнота (*Recall*) ответа. Обозначим через ***C*** – коллекцию документов, в которой осуществляется поиск, ***A*** – множество документов-ответов на запрос, ***R*** – множество истинно релевантных документов. Тогда обозначенные выше критерии могут быть записаны в соответствии с рисунком 4.3.

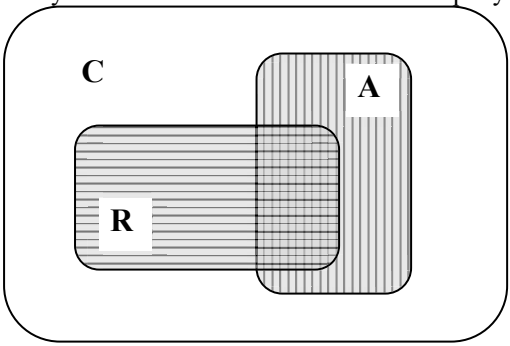


Рис. 4.3 - Критерии качества подходов к решению задач информационного поиска.

Результаты экспериментальных исследований представлены на рисунке 4.4.

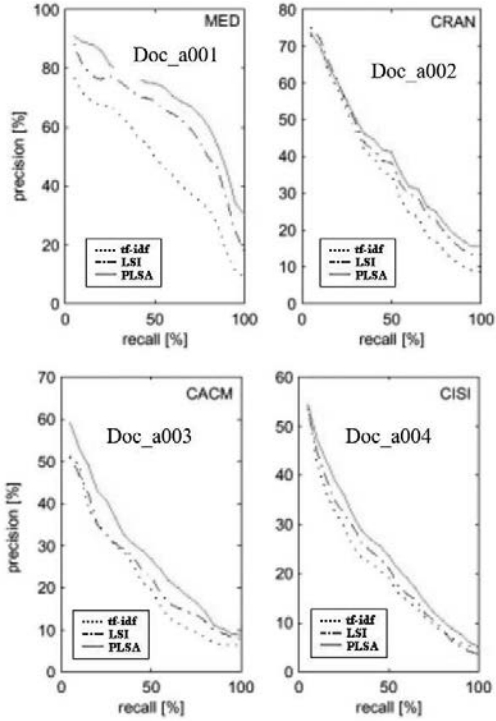


Рис. 4.4 - Результаты сравнения подходов.

Проведенные экспериментальные исследование показали, что разработанный подход дает увеличение качества поиска на всех рассмотренных коллекциях документов (рисунок 4.5).

	PLSA vs tf-idf MED	**PLSA vs LSI**
Doc_a001	10%	5%
Doc_a002	6%	2%
Doc_a003	8%	6%
Doc_a004	3%	3%

Рис. 4.5 - Результаты эксперимента.

Таким образом, при организации адаптивной поисковой системы в среде распределенных информационных ресурсов КИУС, важную роль

играет модель запросов ЛПР. Однако в настоящее время отсутствуют эффективные способы описания ЛПР (пользователя) таких систем, во многом это касается неоднозначности в определении важности характеристик ЛПР, которые должны быть отражены в модели.

По своей сути совокупность распределенных информационных ресурсов КИУС является системой предоставления информации ЛПР по различным производственным ситуациям и предметным областям деятельности ЛПР. Основной целью ЛПР при этом является отыскание «нужной» информации. Предложено учитывать именно “информационные” интересы в модели ЛПР. Предложенная схема организации профиля пользователя представляется в виде множества троек вида: (Категория интересов, Текущий вес категории, Уровень изменчивости). При этом профиль делится на две группы (два подпрофиля): краткосрочный и долгосрочный для учета краткосрочных и долгосрочных интересов пользователя. Представление модели пользователя в таком виде позволило применить для оценки релевантности документов хорошо зарекомендовавшую себя в информационном поиске методологию вероятностного латентно-семантического анализа (PLSA).

Предложенный способ организации модели запросов ЛПР позволил разработать алгоритм непрерывной корректировки модели с учетом новой информации, получаемой при взаимодействии ЛПР с КИУС. Разработанный алгоритм адаптивной подстройки модели на основе новых данных текущего запроса с использованием методологии вероятностного латентного семантического анализа (PLSA) позволяет корректировать параметры модели на основе семантической близости терминов нового запроса и текущего терминологического базиса. При этом возможно два способа применения модели. В первом случае на начальном этапе уточняется запрос на основе информации в модели. Во втором случае данные из модели запросов ЛПР используются для оценки релевантности документов на конечном этапе поискового процесса по этому запросу ЛПР.

Разработанный алгоритм формирования и непрерывной корректировки параметров модели ЛПР реализован на основе методологии вероятностного латентно-семантического анализа, что позволило учесть семантическую близость терминологического состава последовательных запросов ЛПР при реализации поддержки принятия решений в ИУС современных предприятий и корпораций.

5 Система информационно-лингвистического обеспечения принятия решений в распределенных системах

5.1 Функциональная структура систем

В предыдущих разделах были разработаны модели формирования информационного обеспечения для распределенной СППР, имеющей тезаурус с весами и без весов, а также разработана модель синтеза структуры информационного обеспечения распределенной системы поддержки принятия решений. Для апробации предложенных моделей и алгоритмов, учитывая цели данной работы, было решено реализовать две логически связанные программные подсистемы, показанные на рисунке 5.1.

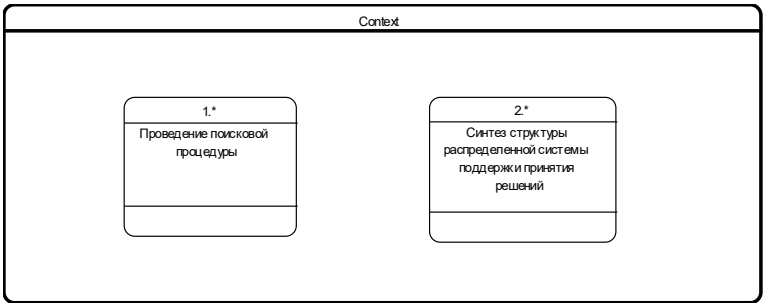


Рисунок 5.1 – Структура системы информационно-лингвистическиого обеспечения распределенных систем поддержки принятия решений.

Первая подсистема отвечает за поиск качественной информации в корпоративной информационной сети (с возможностью подключения глобальной сети Internet). Так же данная подсистема выполняет определение релевантности документов, полученных в процессе проведения поисковой процедуры, и производит ранжирование полученных документов. Кроме того, необходимо указать, что в результате проведения анализа предметных областей был сделан следующий вывод. В классическом понимании характеристика

релевантности является булевой (документ принадлежит запросу или нет). Однако после анализа текстов можно прийти к выводу, что это не совсем так, а именно, применяя алгоритмы определения релевантности можно доказать, что некоторые документы являются «как бы более релевантными запросу», а некоторые «менее». Именно этим и занимаются алгоритмы определения релевантности и ранжирования. Так же выявлены условно релевантные документы; это документы из смежных предметных областей. Следовательно, при поиске необходимо учитывать возможность попадания в результирующую выборку данных предметных областей. Зная о близости данных областей, необходимо учитывать выбор пользователя системы на их включение или исключение в результирующую выборку. Кроме того, в выбранном тексте релевантным предметной области может быть не весь документ, а только часть его, например учебники общего назначения, сборники статей и т.д.

Вторая подсистема отвечает за формирование синтеза структуры информационного обеспечения распределенной системы поддержки принятия решений. Данная подсистема необходима при формировании или модификации структуры распределенной системы. Это связано с тем, что необходимо учитывать возможности узлов хранения и обработки информации, а также затраты на трафик как внутренний, так и внешний.

Перейдем к рассмотрению решения поставленных задач.

5.2 Алгоритм работы поисковой подсистемы

Обобщенный алгоритм работы поисковой системы можно представить в соответствии с рисунком 5.2. Необходимо отметить, что система работает в четырех режимах:

- Произвести поиск без использования тезаурусов.
- Произвести поиск с использованием тезауруса.
- Произвести поиск с использованием частотного тезауруса.
- Произвести поиск по уже анализируемому ранее запросу.

Каждый из возможных режимов работ будет рассмотрен более подробно ниже.

Как видно из рисунка, первый процесс при выполнении поисковой процедуры – это процесс первичного анализа проблемы (более подробно он будет описан далее). На данном этапе происходит задание

начальных параметров поиска и инициализация поисковой процедуры. Также на данном этапе производится выбор одного из четырех возможных путей выполнения процедуры поиска.

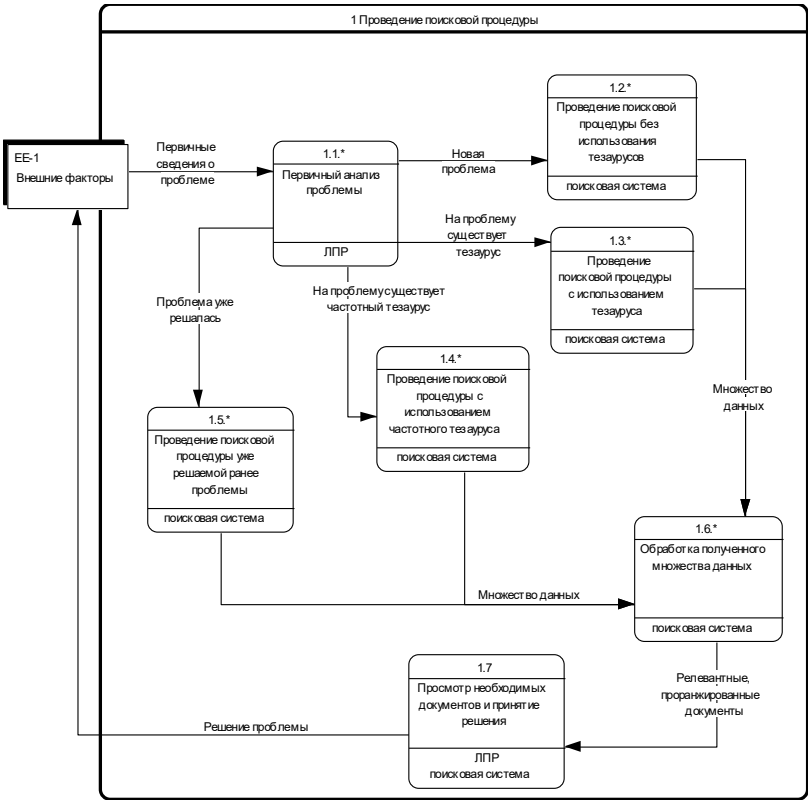


Рисунок 5.2 – Алгоритм работы поисковой подсистемы

После того как будет произведена поисковая процедура по одному из возможных вариантов, необходимо обработать полученные данные на предмет определения релевантности и их ранжирования. Следующим шагом предлагаемого алгоритма является просмотр полученных и обработанных данных лицом, принимающим решение, для выработки окончательной стратегии решения возникшей проблемы.

Рассмотрим каждый из процессов более подробно.

Первый процесс в рамках предлагаемого алгоритма – это процесс первичного анализа проблемы. Данный процесс представлен на рисунке 5.3.

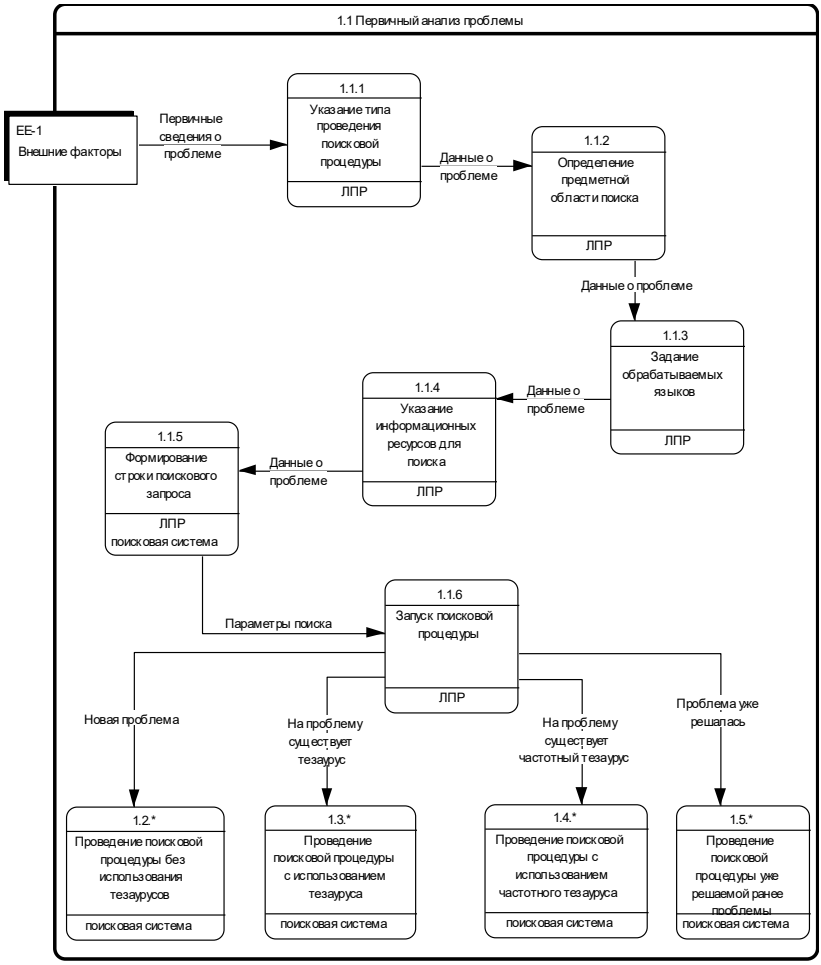


Рисунок 5.3 – Процесс первичного анализа проблемы

Как видно из рисунка, вначале необходимо указать тип проведения поисковой процедуры. Здесь можно выделить следующие возможные типы проведения поиска:

- Поиск по заданному информационному ресурсу корпоративной сети.
- Поиск по информационному ресурсу в сети Internet.
- Метапоиск по заданным информационным ресурсам корпоративной сети, либо по всем ресурсам корпоративной сети.

- Метапоиск по заданным информационным ресурсам сети Internet.
- Метапоиск по заданным поисковым ресурсам сети Internet.
- Смешанный метапоисковый алгоритм как по ресурсам корпоративной сети, так и в сети Internet.

В зависимости от выбора того или иного типа поиска информации в пункте выбора информационных ресурсов автоматически изменяется список ресурсов, например, если не производится поиск в сети Internet, то данные ресурсы не войдут в список выбора.

Следующий шаг выполнения данной функции – это определение предметной области поиска. На данном этапе необходимо указать тезаурус или ключевые слова предметной области (необязательно имеющие отношение к конкретной проблеме из предметной области), в рамках которой необходимо будет проводить поисковую процедуру.

Далее происходит указание обрабатываемых языков. Это очень важная характеристика, так как она связана с обеспечением мультилингвистичности как поисковой строки, так и ответа системы. По умолчанию система работает в одноязычном варианте. Так же возможность мультилингвистичности может быть ограничена наличием/отсутствием мультилингвистических тезаурусов.

Далее необходимо указать информационные ресурсы, в рамках которых будет проведена поисковая процедура. Как говорилось выше, данная функция генерирует список ресурсов динамически в зависимости от выбора типа поиска.

Следующий шаг – это формирование поисковой строки. Данный шаг может быть выполнен или с помощью тезаурусов, или «вручную».

И наконец, после того как лицо, принимающее решение, определилось с проблемой и задало все начальные характеристики необходимо выполнить поисковую процедуру по всем четырем алгоритмам. Рассмотрим каждый из них подробно.

Покажем обобщенный алгоритм работы проведения поисковых процедур (см. рисунок 5.4) на примере возникновения новой ситуации. Остальные типы поиска опишем словесно.

Первый алгоритм поисковой процедуры возникает в случае возникновения новой ситуации, по которой ранее поиск не проводился и, следовательно, отсутствует как обычный тезаурус, так и частотный.

Данный режим работы поисковой системы наиболее прост в исполнении, однако дает наихудшие результаты. Здесь производится поиск по сформированной вручную поисковой строке. Качество выполнения данной процедуры очень сильно зависит от качества поисковой строки. Так же процессы определения релевантности и

ранжирования зависят от того, каким образом данные обрабатывались на информационном ресурсе, на что, в общем случае, мы повлиять не можем.

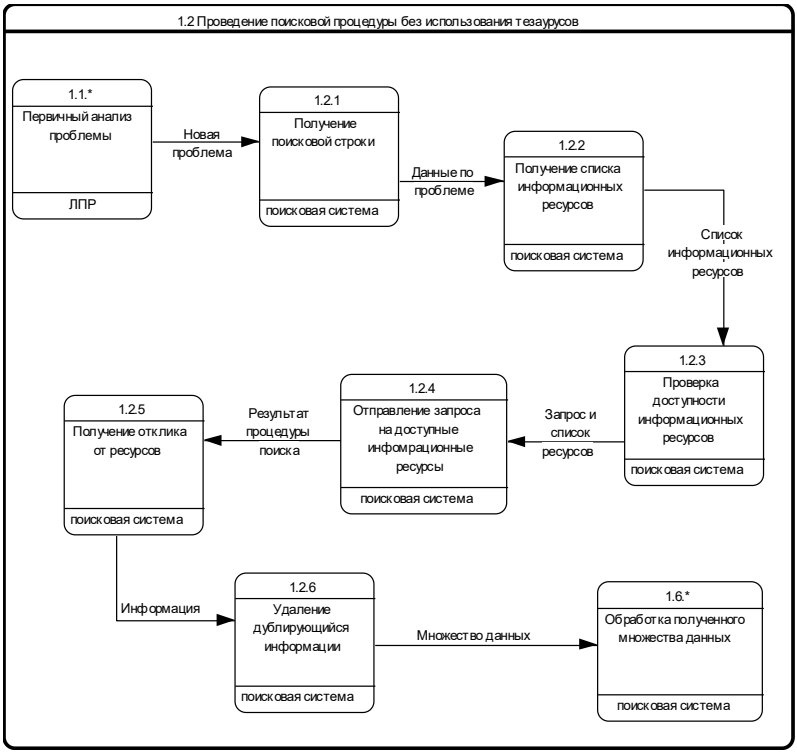


Рисунок 5.4 – Проведение поисковой процедуры без тезаурусов

При рассмотрении процесса можно выделить функции проверки доступности информационного ресурса и удаление дублирующейся информации. Первая несколько ускоряет процесс поиска за счет того, что автоматически удаляет недоступные ресурсы из поиска, вторая отвечает за увеличение коэффициентов релевантности и ранжирования данного документа.

Второй и третий возможный алгоритм работы поисковой процедуры схожи с исходным. Отличие состоит в том, что первый использует обычные тезаурусы, а второй частотные. Также важной частью работы алгоритма является мультилингвистичность тезаурусов. Данные особенности возникают на этапах получения поисковой строки. Здесь необходимо отметить, что если пользователь задал одноязычную

поисковую строку и выбрал мультилингвистический поиск, то система автоматически переведет одноязычную строку поиска в мультилингвистическую, используя тезаурус (при обычном мультилингвистическом тезаурусе), или при формировании поисковой строки добавит новые мультилингвистические термы, используя частотные характеристики термов из частотного мультилингвистического тезауруса. Так же система проверит наличие мультилингвистичности термов в заданной поисковой строке и выбранных информационных ресурсах.

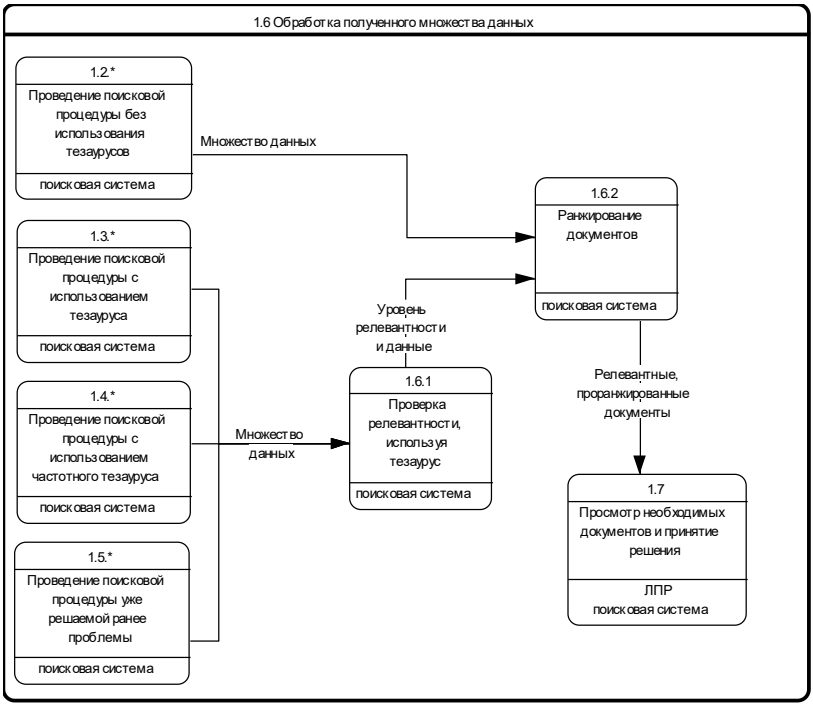


Рисунок 5.5 - Обработка полученного множества данных.

И, наконец, последний режим работы поисковой процедуры - проведение поисковой процедуры для уже решаемой ранее проблемы. Здесь так же нет принципиальных отличий от приведенного выше рисунка, однако необходимо указать, что в результате проведения проверки релевантности и ранжирования информации сыграют существенную роль уже просмотренные ранее документы. То есть, если лицо, принимающее решение, ранее просматривало документ,

представленный системой, то его релевантность будет выше, и при ранжировании данные документы будут на первом месте.

После проведения поисковой процедуры необходимо провести обработку полученного множества данных (см. рисунок 5.5).

Как видно из рисунка, проведение поисковой процедуры без использования мультилингвистических тезаурусов происходит без дополнительной проверки уровня релевантности полученных документов, что значительно снижает качество дальнейшего процесса ранжирования. В этом случае ранжирование происходит, исходя из ранга документа, который присваивает ему ресурс, с которого данный документ был получен, а также учитывается частота встречаемости каждого документа на разных информационных ресурсах.

Для оставшихся трех вариантов необходимо произвести дополнительную проверку уровня релевантности каждого документа. Исходя из данного уровня релевантности, происходит ранжирование документа, однако при ранжировании в случае поиска по решаемой ранее проблеме документы, которые просматривались ранее, получают наивысший ранг, а затем идут все остальные документы.

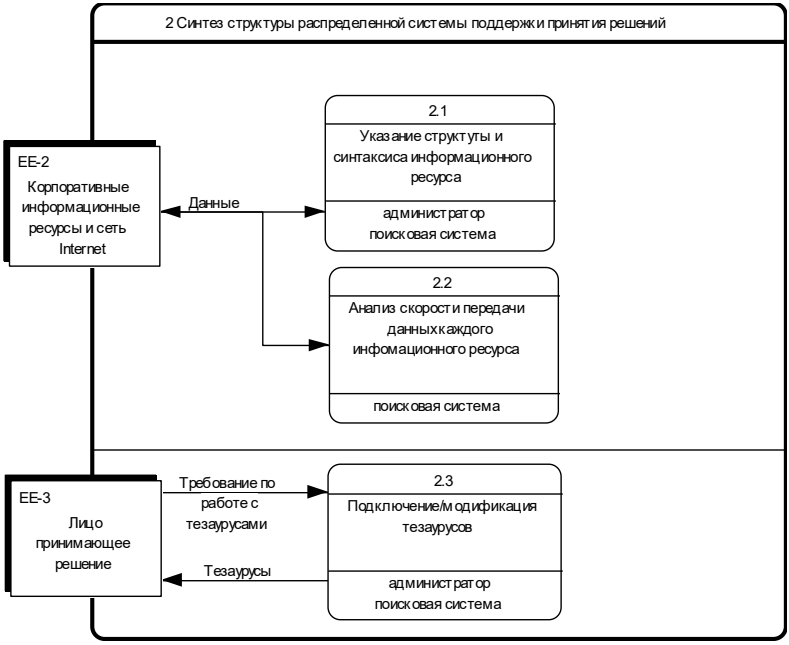


Рисунок 5.6 – Синтез распределенной системы поддержки принятия решений.

Далее рассмотрим следующую подсистему.

Рассмотрим подробнее подсистему синтеза структуры информационного обеспечения распределенной системы поддержки принятия решений (см. рисунок 5.6). Кратко, данную подсистему можно охарактеризовать как администраторскую. Ее назначение - администрирование не только системы поиска, но и всей корпоративной сети в целом.

Как видно из рисунка данная подсистема решает три важные задачи:

- Учет структуры и синтаксиса информационного ресурса.
- Анализ скорости передачи данных по каждому информационному ресурсу.
- Подключение и/или модификация мультилингвистических тезаурусов.

Учет структуры является важной функцией для работы корпоративной информационно-управляющей системы, в которую входит не только информационное обеспечение, но и характер сетевой структурной организации КИУС, а также учет характеристик оборудования на узлах сети.

Следующая функция - анализ скорости передачи данных по каждому информационному ресурсу. Данная функция необходима для того, чтобы оценить затраты на обработку информации и загруженность сетевого трафика. Имея подобные данные, можно принимать решение о целесообразности обновления как коммуникационного оборудования, так и корпоративных серверов.

И последняя функция предназначена для управления информационно-лингвистическим базисом распределенных систем поддержки принятия решений. Основными задачами данной функции являются:

- Подключение новых тезаурусов.
- Замена простых тезаурусов на частотные (по мере их формирования).
- Замена простых тезаурусов и/или частотных тезаурусов на мультилингвистические (по мере их формирования).
- Формирование требований на составление новых тезаурусов по заданным предметным областям.

После того как рассмотрены алгоритмы и функции подсистем, необходимо перейти к их программной реализации.

5.3 Программная реализация поисковой подсистемы

Рассматривается подсистема поиска информации в корпоративной сети с возможностью использования информационных ресурсов глобальной сети Internet. Основное назначение данной подсистемы состоит в формировании информационно-лингвистическиого обеспечения распределенных систем поддержки принятия решений.

Главное окно данной подсистемы приведено на рисунке 5.7. Для удобства пользователя в данном окне основные функции продублированы как кнопками, так и пунктами главного меню.

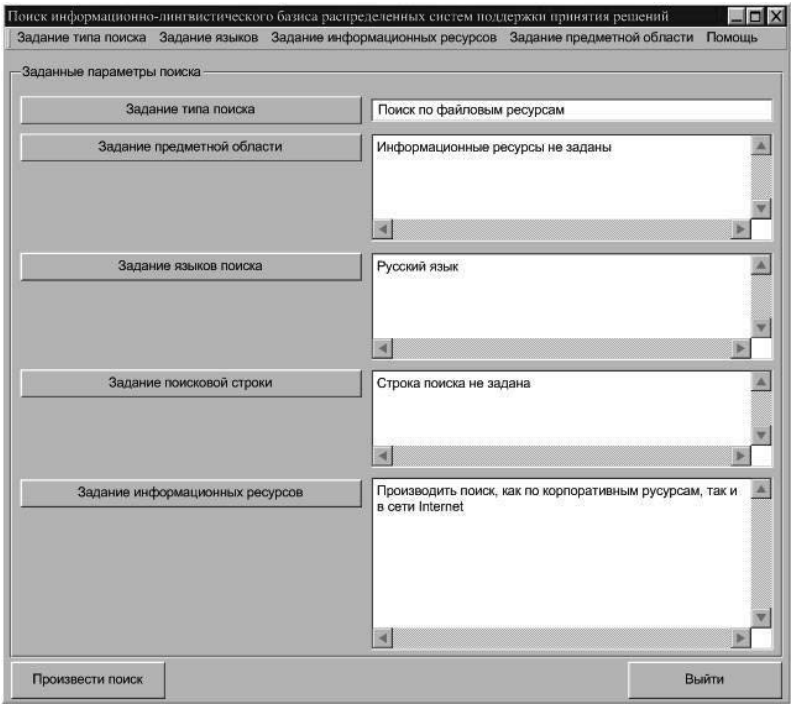


Рисунок 5.7 – Главное окно подсистемы поиска информационно-лингвистического базиса распределенных систем поддержки принятия решений

Данное окно предназначено для задания начальных параметров

поисковой системы:

- Выбор типа поиска – по умолчанию поиск по файловым ресурсам сети.
- Задание предметной области.
- Задание мультилингвистичности поиска (выбор языков поиска) – по умолчанию поиск только на русском языке.
- Задание поисковой строки.
- Задание информационных ресурсов – по умолчанию расширенный поиск (как в корпоративной сети, так и в Internet).

Также данное окно инициализирует начало поисковой процедуры.

Рассмотрим более подробно каждую функция данного окна.

Первая функция - задание типа поиска. Как говорилось выше, поиск может производиться в шести вариантах (см. рисунок 5.8):

- Поиск по заданному информационному ресурсу корпоративной сети.
- Поиск по информационному ресурсу в сети Internet.
- Метапоиск по заданным информационным ресурсам корпоративной сети, либо по всем ресурсам корпоративной сети.
- Метапоиск по заданным информационным ресурсам сети Internet.
- Метапоиск по заданным поисковым ресурсам сети Internet.
- Смешанный метапоисковый алгоритм, как по ресурсам корпоративной сети, так и в сети Internet.

Необходимо отметить, что первые пять пунктов требуют указания ресурсов, по которым будет производиться поиск (метапоиск).

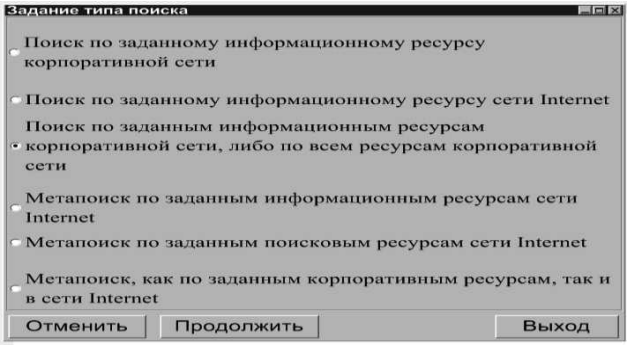


Рисунок 5.8 – Задание типа поиска.

При нажатии кнопки «отменить» происходит закрытие окна без учета изменений. При нажатии кнопки «продолжить» все изменения

будут приняты.

Следующая функция главного окна – задание предметной области - представлена на рисунке 5.9.

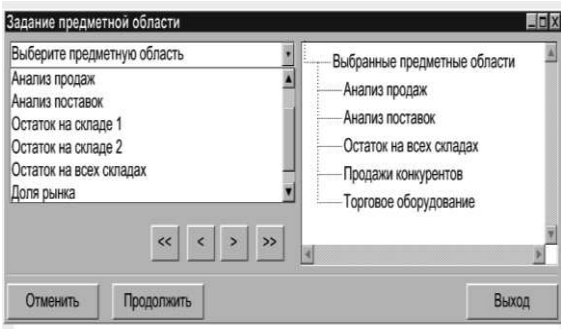


Рисунок 5.9 – Задание предметной области.

Как видно из рисунка, в левой области поле с раскрывающимся списком, в котором необходимо выбирать интересующие ЛПР предметные области. Далее нажатием кнопки”>” выбранная предметная область отобразиться в правой части и будет задана для поиска. Для отмены выбранной предметной области необходимо нажать кнопку ” < ”. Для выбора всех и отмены всех предметных областей необходимо нажать кнопки ” >>” ” <<”, соответственно.

Следующие функции: задание языков поиска и выбор информационных ресурсов выполнены аналогично функции задания предметной области.

Рассмотрим последнюю функцию задания начальных параметров поисковой процедуры – задание поисковой строки. Данная функция выполнена в соответствии с рисунком 5.10.

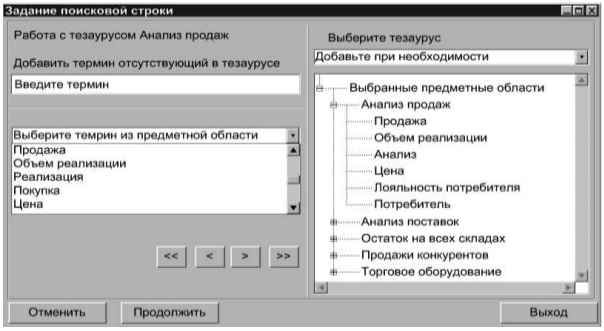


Рисунок 5.10 – Задание поисковой строки.

В левой части окна производится работа с терминами из тезаурусов. По умолчанию, автоматически, подключаются тезаурусы, относящиеся к предметным областям. Так же для расширения охвата поиска можно добавить еще некоторые тезаурусы из смежных предметных областей, кроме того, можно добавить термин, не содержащийся в тезаурусе.

Таким образом, задание начальных параметров закончено, и можно приступать к автоматической поисковой процедуре. Результат выполнения данной функции показан на рисунке 5.11.

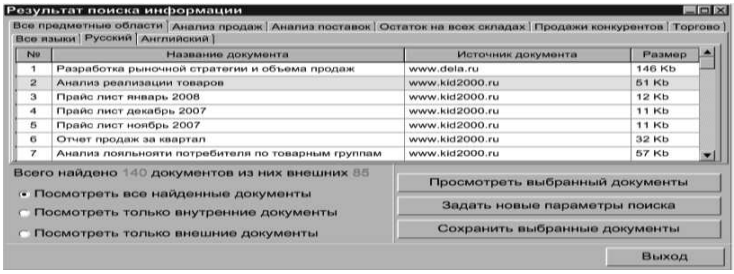


Рисунок 5.11 – Результат поиска информации.

Как видно из рисунка, разработанная система позволяет просматривать как каждую предметную область отдельно, так и совместно. Также при просмотре результата поисковой процедуры можно выбирать просмотр по языкам или «мультилингвистично». Так же система позволяет просматривать найденные документы по принадлежности к собственным ресурсам компании, внешние ресурсы (ресурсы сети Internet), а также и в «объединенном» режиме. Так же можно сохранить выбранные документы на локальном рабочем месте лица, принимающего решение.

Таким образом, показана программная реализация проблемы формирования информационно-лингвистического базиса распределенных систем поддержки принятия решений.

5.4 Модельный эксперимент

При решении практической задачи определения параметров тезауруса с применением информетрических процедур использована текстовая коллекция по направлению «организация производства», данная коллекция была получена из документов в русскоязычном

секторе Интернет, а также подборки статей на английском языке. Эти тексты использованы для получения статистически достоверной выборки. Для формирования тезауруса произведен выбор инструментальных средства формирования словника, средств частотного анализа текстовых массивов, средств семантического анализа текста.

В ходе исследования была сформирована совокупность терминов по предметным областям организации производства, учитывающая частоту встречаемости терминов в разных текстах. Необходимо отметить, что для построения тезауруса наличие полного перечня ключевых слов предметной области является необходимым условием. Но в качестве ключевых терминов могут выступать не только отдельные слова, но и устойчивые словосочетания. Для определения словосочетаний проведен семантический анализ текстов. Каждый терм характеризуется частотой встречаемости и связью с подобным термом из других языковых множеств.

В настоящее время в тезаурус включено более 250 понятий. Ведется постоянная работа по его дополнению и уточнению. Сопоставление подготовленного тезауруса с подобными словарями, существующими в Интернет, показывает, что практически все слова этих словарей представлены в тезаурусе.

Целью проведения эксперимента является сравнительный анализ предложенных решений с существующими поисковыми алгоритмами и существующими решениями.

В ходе проведения эксперимента были введены некоторые ограничения:

- ограничение, связанное с включением в поиск информации предлагаемой системой только трех поисковых сервисов рунета;
- ограничение, связанное с уходом от мультилингвистичности (поиск только в русскоязычном сегменте Internet);
- ограничение, связанное с ограниченностью экспертной комиссии по проверке релевантности и ранжирования отклика системы на запросы;
- ограничение, связанное с выбором только первых 100 ссылок от каждой поисковой системы из сети Internet при проверке уровня релевантности документов и качества ранжирования отклика.

Однако, даже несмотря на данные ограничения, можно принимать первичное решение о качестве разработанных алгоритмов и моделей поиска.

Для проведения эксперимента была реализована метапоисковая система, которая обращается к 3 поисковым системам рунета: Yandex,

Google и Rambler. В результате опроса Internet, используя одну и ту же поисковую строку («системный анализ»), было получено с системы Yandex – 849 198, Google – 1 561 782, Rambler – 884 437. В совокупности получается, что предлагаемая система получила – около 3 300 000 ссылок на страницы в сети Internet документов с трех поисковых сервисов. Однако необходимо оговориться, что после удаления ссылок на общие документы (по прогнозам) осталось – около 2 200 000 документов.

Таким образом, видно, что по сравнению с одним поисковым сервисом метапоисковая технология активного поиска позволяет обработать информацию с большим охватом сети Internet. Значит, предлагаемая методика поиска значительно лучше, чем стандартные поисковые системы с принудительной индексацией информационных ресурсов.

Следующий шаг эксперимента посвящен проверке уровня релевантности полученных документов предлагаемой системой и системой Yandex. Yandex выбрана в связи с тем, что на текущий момент времени (по рекомендациям экспертов) данная система является наиболее предпочтительной при поиске информации в рунете. Ей отдают предпочтение порядка 50% пользователей поисковых сервисов.

После проведения проверки полученного множества документов из сети Internet с помощью предлагаемой системы (при поиске с технологией частотного тезауруса) общее число полученных документов сократилось с 652 753 до 25 326, что составляет примерно 96.2 % нерелевантных документов и 3,8% релевантных документов. При поиске с использованием простого тезауруса были получены следующие результаты: число документов сокращено до 45 782, что составляет примерно 93,1 % нерелевантных документов и 6,9 % релевантных документов. Также была проверка анализа релевантности документов, полученных с поисковой системы каталожного типа. Анализ показал, что предлагаемая методика не уступает по релевантности каталожным системам. Реально релевантных документов в выборке по оценке примерно 2 %, что составляет 13 000 документа. Результат проведения эксперимента можно представить графически (см. рисунок 5.12 и табл. 5.1).

Таблица 5.1. Данные эксперимента.

Используемый подход	Частотные словари	Частотный тезаурус	Тезаурус	Yandex

Количество нерелевантных документов	5 000	12 326	32 782	639 753
Количество релевантных документов	13 000	13 000	13 000	13 000

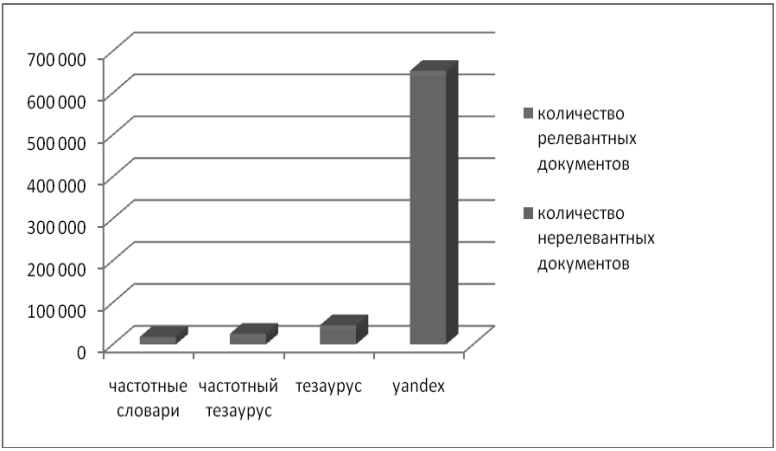


Рисунок 5.12 – Соотношение релевантных документов к общему количеству

Таким образом, мы видим, что предлагаемая технология позволяет получить гораздо лучший результат в задачах определения релевантности узкоспециализированной информации по сравнению с классическими поисковыми системами. Кроме того, необходимо отметить, что предлагаемые решения дают несколько худший результат по сравнению с использованием модели, основанной на частотных мультилингвистических словарях, составленных по предметной области. Так же при анализе релевантности предлагаемая технология сопоставима с каталожными системами.

Последний эксперимент нацелен на ранжирование отклика сети Internet на пользовательский запрос.

Группа экспертов пришла к единогласному выводу, что предлагаемая технология расчета рангов документов значительно превосходит методику, разработанную крупными поисковыми сервисами, но нацеленную не на работу с узкоспециализированной литературой, а для удовлетворения широких нужд пользователей. При сравнении с тематическими каталогами сети Internet, например, с

сайтом, посвященным математической литературе – www.math.ru , можно сделать вывод, что предлагаемая методика интересна в связи с тем, что ранжирование происходит не по алфавиту, а в соответствии с требованиями пользователя.

Подведя итог эксперимента, можно сказать следующее:

- предлагаемая технология позволяет производить поиск исходной информации для тематических коллекций с охватом значительно большего сегмента сети Internet (за счет использования метапоисковой технологии активного поиска);
- уровень релевантности предлагаемой технологии значительно выше при поиске узкоспециализированной информации, чем у классической поисковой технологии, и сопоставим с поисковыми системами каталожного типа, за счет использования проверки уровня релевантности на основе узкоспециализированных частотных словарей;
- по качеству ранжирования информации предлагаемая технология работает значительно лучше, чем:
- классические поисковые технологии сети Internet;
- метапоисковые сервисы сети Internet;
- не уступает, а зачастую, и значительно превосходит каталожные системы.

Итак, в работе предложена функциональная структура системы формирования и обработки информационно-лингвистического базиса распределенных систем поддержки принятия решений. Система содержит два модуля: информационно-лингвистическая поисковая подсистема и подсистема анализа/синтеза структуры. В рамках разработанных программных систем реализованы предложенные модели и алгоритмы обработки информации.

Представлена программная реализация предложенных решений. Основной акцент при программной реализации уделен поисковым решениям в корпоративной сети и сети Internet. Показана возможность проведения поисковой процедуры без использования тезаурусов вообще, с использованием простого тезауруса и с использованием частотного тезауруса. Также продемонстрирована возможность проведения мультилингвистического поиска и обработки информации в рамках разработанных алгоритмов и решений.

В ходе эксперимента получены результаты, которые демонстрируют существенно лучшее качество анализа документов на релевантность по сравнению с классическими поисковыми системами. Также показано, что по сравнению с реализацией подхода, основанного на частотных мультилингвистических словарях, степень релевантности определяется незначительно хуже, однако это не снижает значение полученного

результата, так как составление тезаурусов значительно менее трудоемкая задача по сравнению с составлением мультилингвистических тематических словарей.

При анализе уровня качества ранжирования информации получены не менее интересные результаты. Так, качество ранжирования предлагаемых алгоритмов значительно превосходит качество ранжирования информации классическими поисковыми сервисами. Также качество ранжирования предлагаемых решений не уступает поисковым системам каталожного типа и системе, построенной на использовании тематических частных словарей.

6. Мультилингвистическая система поиска и анализа текстов для принятия решений

В данном разделе приводится описание реализованной системы поиск мультилингвистической информации для поддержки принятия решения в корпоративных информационно-управляющих системах (рис. 6.1). Следует отметить, что для практической реализации рассмотренных процедур поиска данных и документов предлагается оригинальная методика определения расстояния между отдельными точками метрического пространства производственных ситуаций. При этом для определения расстояния между двумя произвольными производственными ситуациями $S_1(t)$, $S_2(t) \in \{S(t)\}$ необходимо предварительно определить функции, устанавливающие степень сходства, существующего между соответствующими характеристиками $S_1(t)$, $S_2\{t)$.

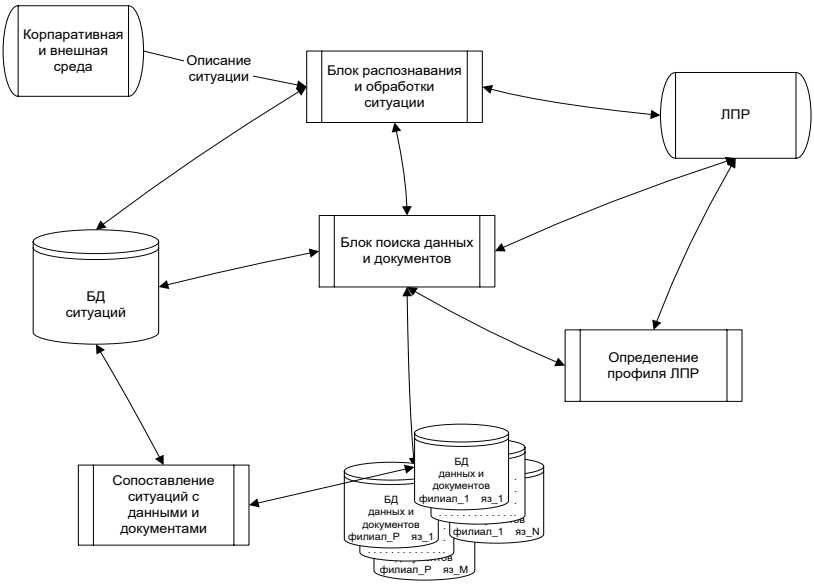


Рис. 6.1 – Структурная схема реализации мультилингвистической технологии поиска данных для подготовки и принятия решения в ИУС.

Характеристики разбиты на четыре группы. В первой группе

характеристик содержится текстовая информация, во второй группе хранится информация в числовой форме, в третьей группе — ориентированные графы, используемые ЛПР в процессе подготовки и принятия решения. При этом граф G_1 характеризует взаимосвязи, существующие между параметрами объекта управления, системы управления и среды, влияющими на процесс принятия решения в ситуации S, а граф G_2 учитывает взаимосвязи между административными мероприятиями, которые рекомендуется осуществить в сложившейся ситуации. В характеристиках четвертой группы хранится название производственных ситуаций.

Информационные элементы рассмотренной модели хранятся в базе данных реляционного типа в виде совокупности отношений различной структуры. Поиск этих элементов осуществляется по запросу СУБД, сформированному на языке SQL и выдаваемому при обработке каждой производственной ситуации.

В работе рассмотрены процедуры формирования функций сходства для каждой из указанных групп характеристик, затем на их основе выбрана метрика, определяющая расстояние между отдельными производственными ситуациями. Функции Q_i^j, $j = 1,...,4$ используются в составе метрики p_s, задающей расстояние между $S_1(t)$, $S_2(t)$ в пространстве производственных ситуаций. Принято во внимание, что при t = const метрика $p_s(S_1(t), S_2(t))$ должна быть действительной числовой функцией, для которой выполняются известные аксиомы метрики. В качестве функции, заведомо обладающей данными свойствами, выбрано евклидово расстояние, определяемое по формуле

$$p_s(S_1(t), S_2(t)) = [\sum^4_{i=1} Q_i^2]^{1/2}.$$

На основе предложенной процедуры определения расстояния между различными производственными ситуациями, возникающими в процессе функционирования объекта управления, реализованы новые алгоритмы оперативного поиска мультилингвистических данных и документов. Компоненты инструментальных средств, разработанных в ходе работы, успешно внедрены в составе информационно-управляющих систем ряда промышленных предприятий и организаций.

Итак, необходимо еще раз обратиться к проблемам современных информационно-управляющих систем. Как уже говорилось выше, существует несколько основных характеристик подобных систем:

- большие объемы информации. Данная проблема связана с тем, что на текущем этапе развития корпораций документооборот значительно увеличивается с ростом предприятия. Так же данная проблема становится более актуально при открытии предприятием новых филиалов и представительств.

- проблема устаревания и дублирования информации. Проблема дублирования информации становится наиболее актуальной при появлении новых филиалов и представительств. Проблема устаревания и модификации информации наиболее актуальной при расширении корпорации. Это связанно с тем, что в рамках одного представительства можно несложными приемами организовать отслеживание и структуризацию информационных источников и хранилищ, однако при росте предприятия, уже необходимо организовывать взаимодействие между данными предприятиями, что в свою очередь ведет к модификации и усложнению методов обеспечения целостности и непротиворечивости информации.

- проблема мультилингвистичности. За последнее десятилетие данная проблема стала актуальна и для российских компаний. В первую очередь она связанна с появлением международных контактов, как с поставщиками, так и потребителями товаров и услуг. Так же данная проблема актуальна при работе с технической документацией на оборудование и т.п.

6.1 Алгоритм работы системы

Общую схему работы предлагаемой подсистемы поддержки принятия решений, в корпоративных информационно-управляющих системах, можно представить в соответствии с рисунком 6.2.

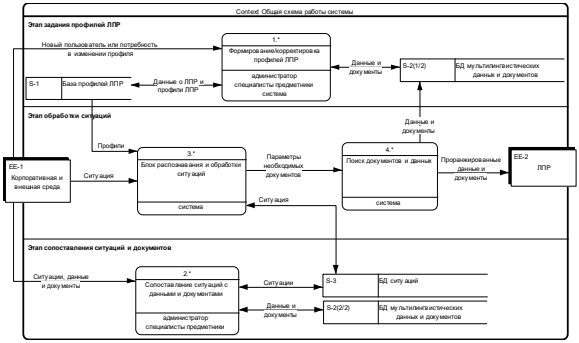


Рисунок 6.2 – Общая схема работы системы

Как видно из рисунка, при работе системы выделяют четыре

основных процесса:

- формирование/корректировка профилей ЛПР;
- сопоставление ситуаций с данными и документами;
- блок распознавания и обработки ситуаций;
- поиск документов и данных.

Данные процессы можно сгруппировать по этапам, как показано на данном рисунке:

- этап задания профилей ЛПР и этап сопоставления ситуаций и документов. Данные этапы можно отнести к предварительным или этапам настройки параметров системы;
- этап обработки ситуаций. Данный этап является основным, в котором реализованы предложенные выше алгоритмы и методы.

Далее необходимо рассмотреть каждый из процессов работы системы.

Начнем с подготовительного этапа работы: задание профилей ЛПР и сопоставления ситуаций и документов.

Первым рассмотрим процесс формирование/корректировки профилей ЛПР. Данный процесс можно показать в соответствии с рисунком 6.3.

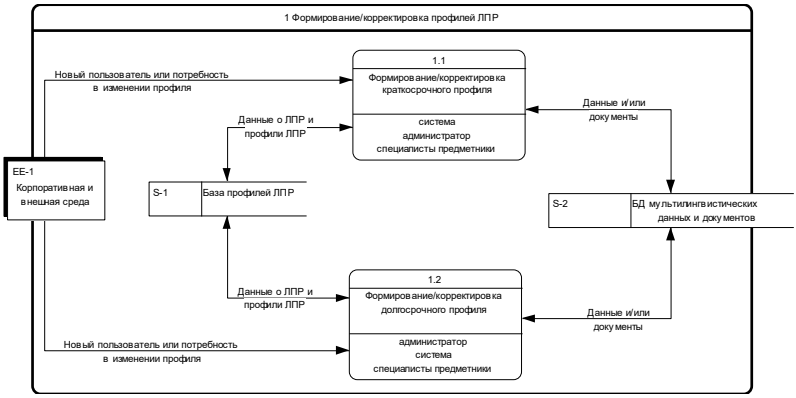


Рисунок 6.3 – Процесс формирования и корректировки профилей ЛПР

Как видно из рисунка выделяется два профиля пользователя: краткосрочный и долгосрочный. Основное отличие между данными профилями можно показать следующим образом: емкость долгосрочного профиля больше емкости краткосрочного. При этом значения переменных, находящихся в краткосрочном профиле,

изменяются линейно, тогда как для переменных, находящихся в долгосрочном профиле снижение весов – экспоненциально.

Следующая функция системы — это сопоставление ситуаций и документов. Процесс выполнения данной функции можно показать в соответствии с рисунком 6.4.

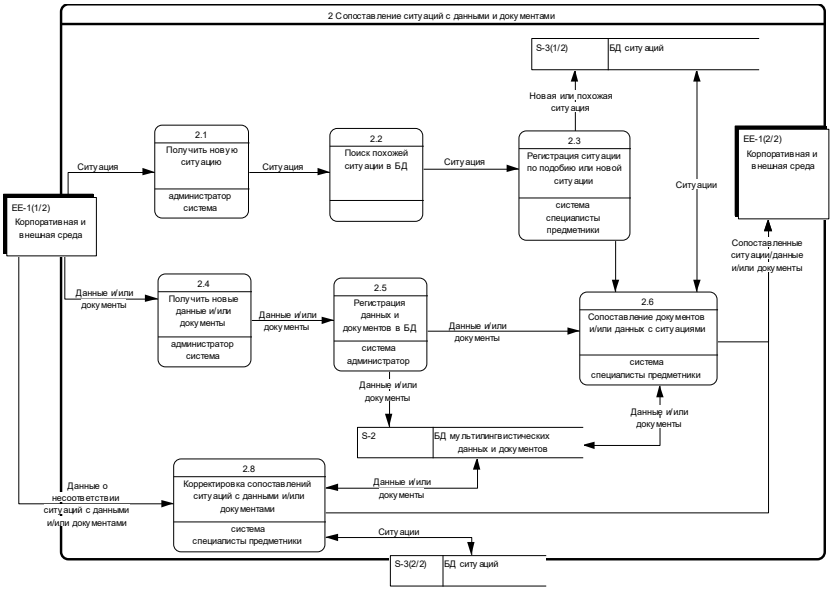


Рисунок 6.4 – Сопоставление ситуаций с данными и документами

Как видно из рисунка, у нас может происходить работа как по получению новой ситуации, новых данных и/или документов, так и по корректировке существующих значений.

В случае возникновения новой ситуации необходимо произвести поиск подобных ситуаций, для упрощения процесса сопоставления данных и/или документов данной ситуации. Если ситуаций не найдено, то процесс сопоставления происходит в полном режиме. Если найдены подобные ситуации, то необходимо проанализировать данные и/или документы, необходимые для решения возникшей проблемы, и, возможно, принять их без изменения или внести корректировку в соответствии с требованиями данной новой ситуации.

В случае получения новых данных и/или документов необходимо произвести анализ, к каким ситуациям эти данные и/или документы

относятся. Далее необходимо произвести сопоставление данных и/или документов с выбранными ситуациями.

И третий режим работы возникает, в случае если выявляются некорректности в сопоставлении ситуаций и/или данных. Для исправления подобных несоответствий необходимо произвести анализ сопоставленных ситуаций с данными и/или документами и исправить проблемные области.

Далее необходимо показать алгоритм работы системы на третьем этапе. Данный этап посвящен анализу и обработке возникших проблем и состоит из двух процессов:

- распознавание и обработка проблем;
- поиск документов и данных.

Рассмотрим процесс распознавания и обработки проблемы, который можно показать в соответствии с рисунком 6.5.

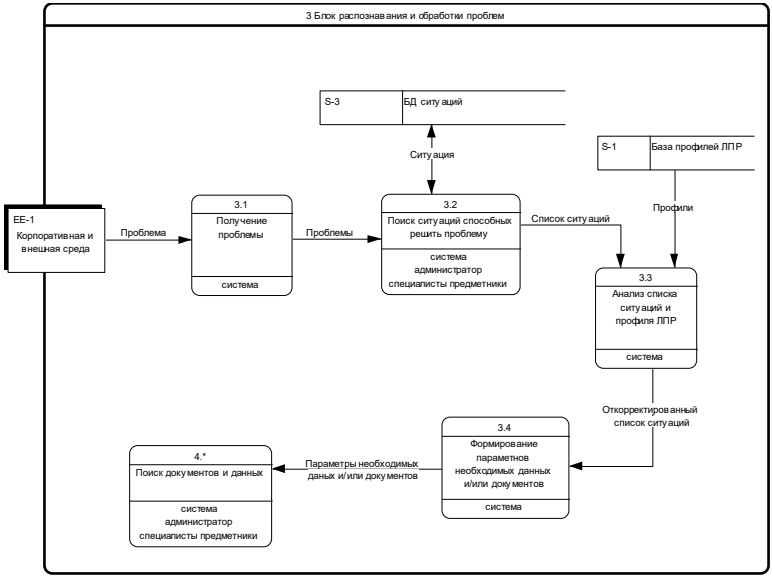


Рисунок 6.5. Распознавание и обработка проблем

Необходимо отметить, что понятие проблема шире, чем понятие ситуация. То есть для решения одной проблемы, возможно, необходимо выбрать одну или более ситуаций.

Очень важно корректно производить разбор проблемы на входящие в них ситуации. Так же не менее важно рассмотреть профиль ЛПР,

который пытается решить возникшую ситуацию. От текущих параметров ЛПР зависит подбор данных и/или необходимых документов, а также уровень полномочий данного сотрудника и что не менее важно, это уровень и степень знания иностранных языков, что связанно с мультилингвистичностью представленной информации.

В результате выполнения данной процедуры будут получены параметры необходимых данных и/или документов. На данном этапе наиболее важной информацией являются профили ЛПР. Так же здесь происходит перерасчет текущих значений профиля каждого ЛПР, которые участвуют в решении проблемы. В параметры входит информация из профиля ЛПР, которая относится к предпочтениям тех или иных корпоративных хранилищ информации, баз данных и предпочтений по корпоративным серверам.

После получения параметров необходимой информации можно приступать к произведению поисковой процедуры.

Поисковая процедура показана в соответствии с рисунком 6.6.

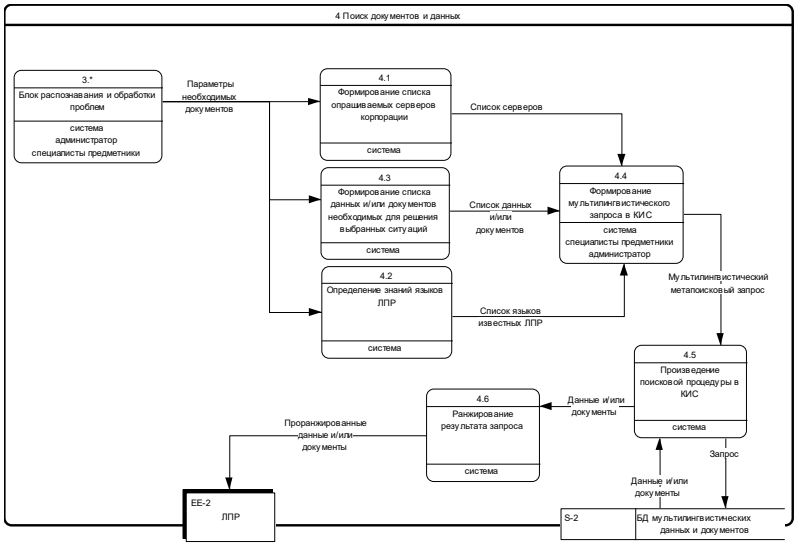


Рис. 6.6 – Процедура поиска данных и/или документов.

Выполнения данной процедуры начинается с анализа параметров необходимых данных и/или документов необходимых для решения возникшей проблемы. Анализ состоит из трех функций:

- формирование списка опрашиваемых серверов. Данный список

зависит от профиля пользователя и включает в себя те сервера, которые имеют наибольший приоритет для каждого ЛПР. Необходимо отметить, что в результате выполнения данной функции список серверов для разных ЛПР будет разный.

- формирование списка данных и/или документов необходимых для решения выбранных ситуаций входящих в проблему. Данный список зависит от профиля пользователя и включает в себя те данные и документы, которые имеют наибольший приоритет для каждого ЛПР. Необходимо отметить, что в результате выполнения данной функции список данных и/или документов для разных ЛПР будет разный.
- определение знаний языка ЛПР. Данная функция отвечает за мультилингвистический аспект КИС. Для каждого ЛПР данный список так же уникален.

После того как параметры запроса в КИС были определены, необходимо еще раз проанализировать данную информацию и сформировать запрос или несколько запросов в информационное пространство корпорации.

Следующий этап посвящен произведению поисковой процедуры в КИС. Данный запрос будет производиться по модели описанной выше. Необходимо отметить, что при формировании запроса не менее важную роль играет информация не только о качестве и достоверности того или иного хранилища данных, но и профиль ЛПР в который входят его предпочтения и прошлый опыт работы с системой.

После получения отклика системы необходимо приступать к обработке полученных данных и/или информации. На данном этапе применяются описанные выше алгоритмы определения релевантности и ранжирования информации. В результате выполнения данной процедуры ЛПР получит необходимую ему информацию, для принятия решения по разрешению возникшей проблемы.

Далее необходимо переходить к программной реализации предлагаемых алгоритмов и их апробации.

6.2. Реализация моделей и алгоритмов

При работе с учетом профиля пользователя необходимо произвести настройку параметров системы и настройку самого профиля всех ЛПР. Поэтому работа с системой начинается с утилиты «настройка системы ЛПР для принятия решения». Главное окно данной утилиты выглидит в

соответствии с рисунком 6.7.

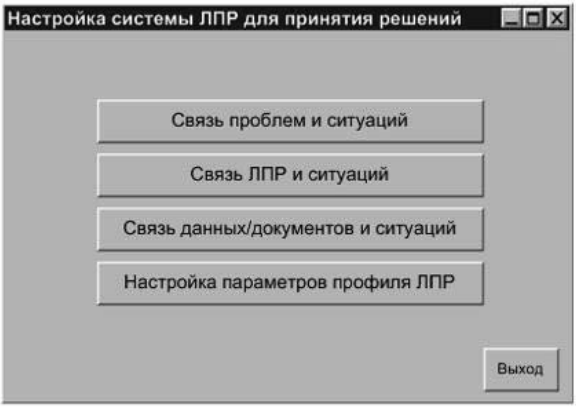


Рисунок 6.7 – Главное окно утилиты.

Как видно из рисунка утилита позволяет настраивать 4 параметра системы:

- связь проблем и ситуаций;
- связь ЛПР и ситуаций;
- связь данных/документов и ситуаций;
- настройка параметров профиля ЛПР.

Рассмотрим каждую из данных возможностей отдельно.

Первая возможность – Связь проблем и ситуаций. Окно настройки данной функции выглядит в соответствии с рисунком 6.8.

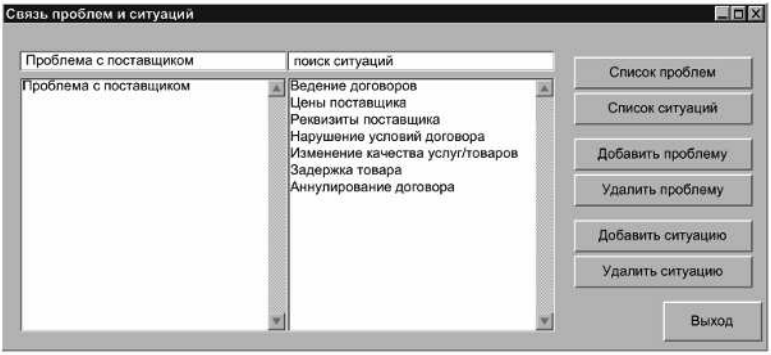


Рисунок 6.8 – Связь проблем и ситуаций

Как уже говорилось, ЛПР встречается с проблемой, которая может

быть разбита на ряд примитивных блоков (ситуаций). На рисунке приведен пример проблемы «Проблемы с поставщиками» и показаны какие ситуации могут возникать в рамках данной проблемы: «Ведение договоров», «Цены поставщиков» и т.д. кроме того, данное окно позволяет управлять списками проблем и ситуаций. Для просмотра списка проблем и ситуаций необходимо нажать на кнопки «Список проблем» и «Список ситуаций». Так же можно производить действия на добавление и удаления проблем и ситуаций, то есть их связи между собой.

Списки проблем и ситуаций можно показать в соответствии с рисунками 6.9 и 6.10.

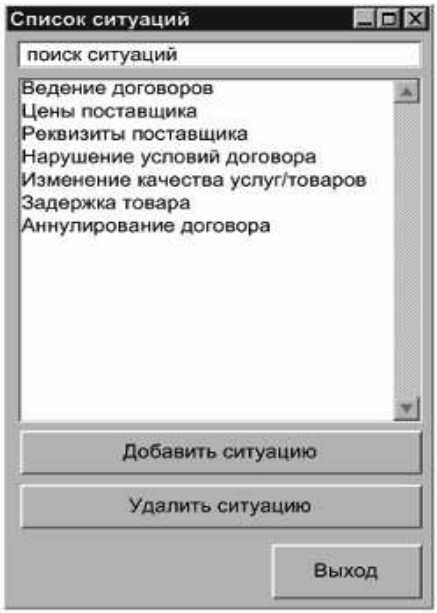


Рисунок 6.9 – Список ситуаций.

Данное окно предназначено для управления перечнем ситуаций и позволяет производить поиск ситуаций, добавление и удаление ситуаций, а также корректировку корректности названия ситуации. По аналогии реализована возможность работы со списком проблем, программное окно данного списка показано ниже.

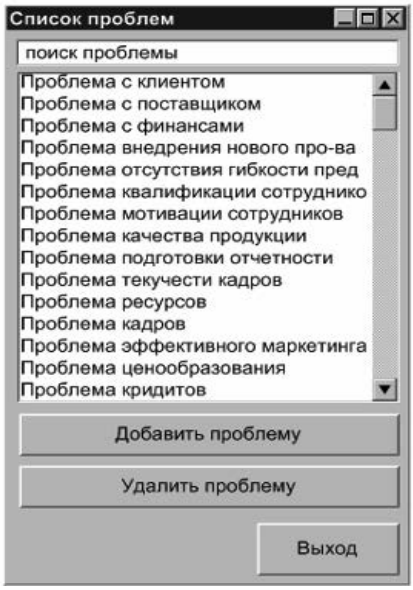


Рисунок 6.10 – Список проблем.

Следующая важная функция, которую можно решать с помощью разработанной утилиты – Связь ЛПР и ситуаций. Окно, предназначенное для реализации данной функции, можно показать в соответствии с рисунком 6.11.

Данное окно разделено на три области:

- данные о ЛПР;
- данные о ситуациях и проблемах, которые может решать ЛПР;
- и область управления.

Рассмотрим каждую из данных областей отдельно.

Данные о ЛПР содержит информацию о должности ЛПР и его персональных данных: фамилия, имя и отчество. Напротив данной области в области управления находятся кнопки для добавления должностей и ЛПР, а также для их удаления.

Необходимо оговориться, что добавление ЛПР подразумевает не внесение ЛПР с корпоративную информационную систему, а добавление в список для управления проблемами и ситуациями, которые данный ЛПР может разрешать.

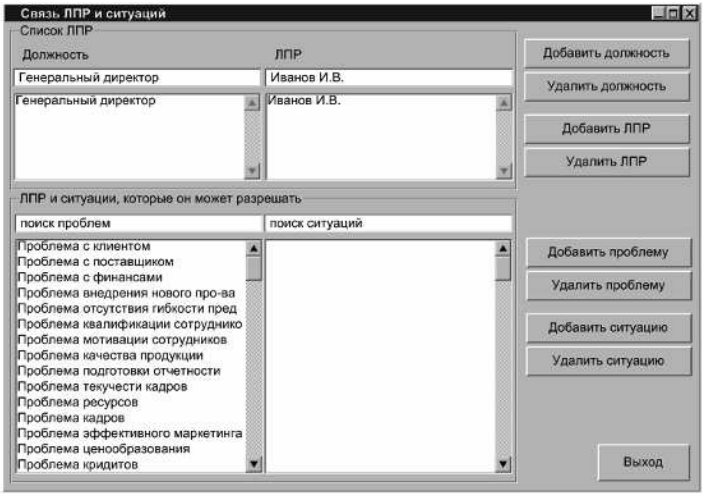


Рисунок 6.11 – Связь ЛПР и ситуаций.

Следующая область окна относится к Проблемам и ситуациям, на которые ЛПР может влиять. Напротив данной области находятся кнопки для расширения полномочий ЛПР. А именно управления проблемами и ситуациями, которые предназначены для данного ЛПР.

Необходимо отметить, что даже в рамках одной проблемы список ситуаций для разных ЛПР будет разный.

Также необходимо отметить, что чем выше должность и полномочия ЛПР, тем больше списки проблем и ситуаций, на которые он может влиять.

В окне также реализована возможность быстрого поиска для всех полей из областей данных о ЛПР и данных о ситуациях и проблемах, которые может решать ЛПР.

Следующая крупная возможность, которую предоставляет данная утилита – это связь данных/документов и ситуаций. Данную возможность можно показать в соответствии с рисунком 6.12.

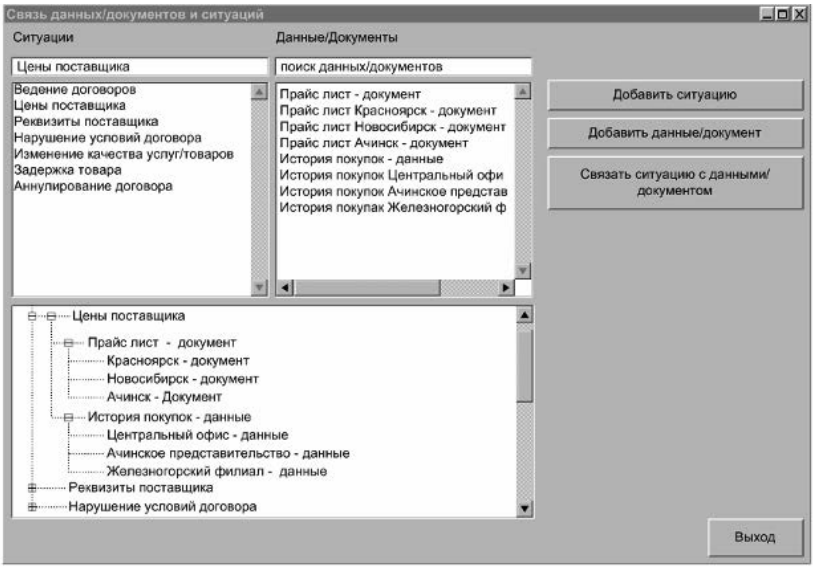


Рисунок 6.12 – Связь данных/документов и ситуаций.

Данная возможность является ключевой при работе с системой поддержки принятия решений, т.к. на этом этапе необходимо указать какая информация должна быть предоставлена ЛПР при возникновении той или иной проблемы. Необходимо понимать, что понятия данные и документы — это совершенно разные информационные источники. Данные — это информация, которая хранится в корпоративной базе данных, как на главном сервере корпорации, так и в филиалах и представительствах. Тогда как документы это и есть электронные корпоративные документы, которые физически находятся в неструктурированном информационном пространстве корпоративной сети.

Рассмотрим более подробно работу с данным окном.

На окне можно выделить четыре области:

- ситуации;
- данные/документы;
- визуализация связей между ситуациями и данными/документами;
- управление ситуациями и данными/документами.

В области ситуаций, при выборе конкретной ситуации отображается список данных/документов, которые связаны с данной ситуацией. На

этом этапе можно корректировать список ситуаций и данных/документов влияющих на данную ситуацию.

В области визуализации связей между ситуациями и данными/документами можно увидеть древовидную структуру, которая помогает быстрому просмотру связей.

Следующая крупная функция – «настройка параметров профиля ЛПР» может быть показана в соответствии с рисунком 6.13.

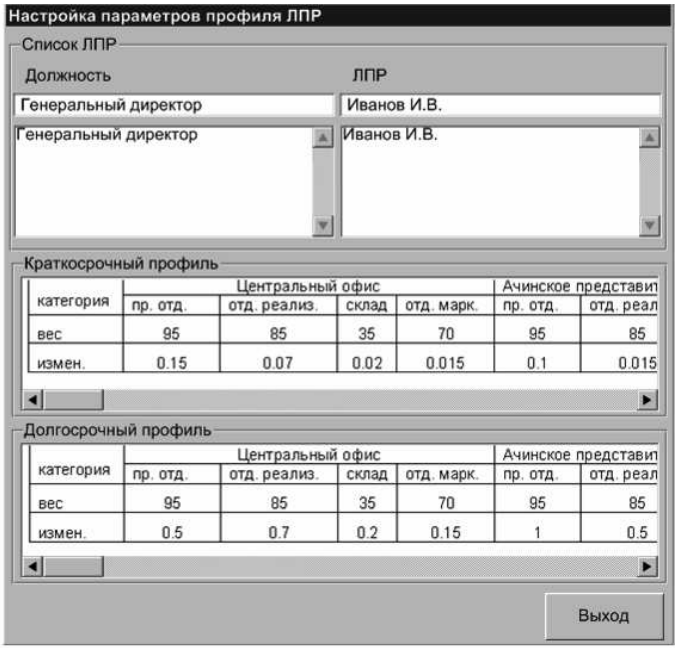


Рисунок 6.13 – Настройка параметров профиля ЛПР.

Данное окно позволяет просматривать и корректировать значения текущего веса и коэффициента изменчивости в профилях ЛПР. Необходимо отметить, что реализованы два профиля: краткосрочный и долгосрочный.

После того как параметры системы настроены, можно переходить к получению и обработки проблем ЛПР.

При возникновении проблем ЛПР производит настройку возможных ситуаций с помощью, которых данную проблему можно решить. Данную функцию можно показать в соответствии с рисунком 6.14.

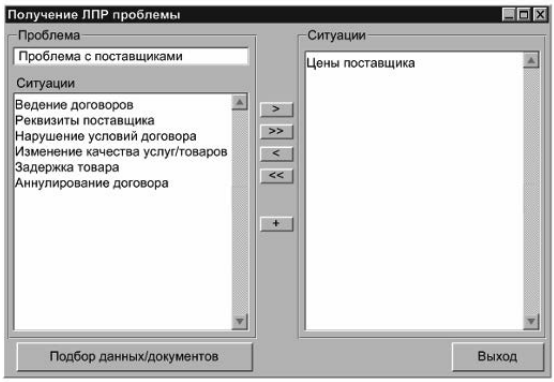


Рисунок 6.14 – Получение ЛПР проблемы.

При получении проблемы она отображается в левой верхней части окна, а ниже перечень ситуаций, которые данный ЛПР может обрабатывать. При обработке проблемы ЛПР может выбрать необходимые ситуации из списка, для того чтобы более корректно произвести подбор необходимых данных/документов. Можно выбирать как по одной ситуации, так и перенести все ситуации, связанные с данной проблемой. Также возможна ситуация, при которой при решении возникшей проблемы необходимо принимать в рассмотрение ситуацию, которая не указана в соответствие с данной проблемой.

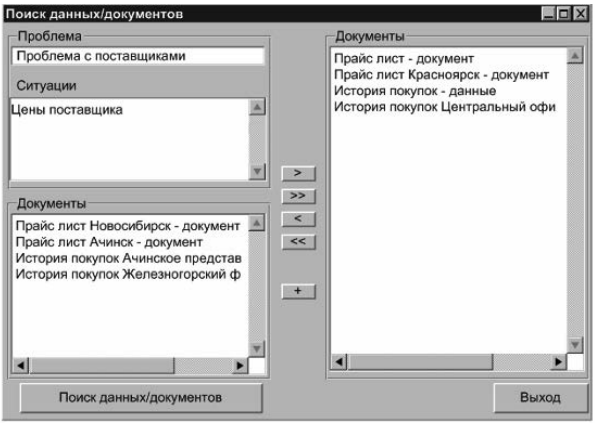


Рисунок 6.15 – Поиск данных/документов.

После того как ЛПР выбрал необходимые ситуации для обработки необходимо нажать на кнопку подбора данных/документов. После этого откроется окно поиска данных/документов, которое можно показать в соответствии с рисунком 6.15.

Данное окно предназначено для настройки параметров поиска необходимой информации. В нем можно увидеть возникшую проблему, ситуации, которые помогут решить данную проблему и перечень документов, которые связанны с выбранными документами. Для решения возникшей проблемы необходимо выбрать необходимые ЛПР данные из списка и возможно добавить данные/документы, не вошедшие в список документов.

После того как ЛПР произведет выбор данных/документов из списка, необходимо наддать на кнопку «поиск данных/документов». В результате будет произведен поиск необходимой информации в мультилингвистической корпоративной информационной среде и произведено ранжирование необходимой информации в соответствии с предлагаемыми алгоритмами. Результат выполнения данного действия можно показать в соответствии с рисунком 6.16.

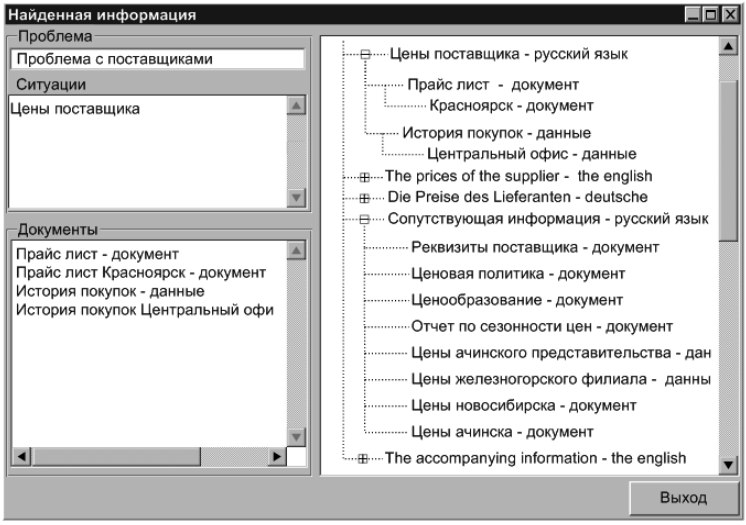


Рисунок 6.16 – Найденная информация.

В данном окне представлен результат выполнения поисковой процедуры системой поддержки принятия решения в корпоративных информационных системах. Как видно левая часть окна предоставляет

вспомогательную информацию о проблеме, ситуации и документах, которые выбрал ЛПР. В правой части виден результат в виде ссылок на документы и данные. Двойной клик по каждому листу дерева документов открывает необходимые документы или выполняет запрос на получение данных из корпоративных баз данных и хранилищ информации. Так же в данном дереве видно, что информация по интересующему вопросу разделена на принадлежность к языковому множеству и по требованиям запросу и документам, относящимся к предметной области возникшей проблемы. На данном этапе ключевую роль играет профиль пользователя и мультилингвистический частотный словарь по предметным областям, относящимся к возникающим проблемам.

Разработанная система поддержки принятия решения в корпоративных информационно управляющих системах создана на базе современных информационных технологий, и в полной мере реализует модели и методики поиска и ранжировании информации. Это позволяет перейти на новый этап развития технологии поддержки принятия решения за счет существенного сокращения трудозатрат при обработке необходимой информации, а также при возникновении проблем и ситуаций, требующих принятия решения на корпоративном уровне.

Реализован алгоритм ранжирования и определения релевантности документов производственной ситуации. Полученные результаты можно применять как при работе с системами поддержки принятия решения, так и в поисковых системах, направленных на обработку результатов поиска как одного, так и нескольких поисковых сервисов (метопоисковые технологии).

Реализован алгоритм текстового анализа документа для определения уровня релевантности этого документа заданной производственной (корпоративной) проблеме/ситуации. Данный алгоритм позволяет анализировать тексты, принадлежащие к любой языковой группе. Возможен анализ письменности, основанной на иероглифах.

Предложенная поисковая методика в полной мере охватывает технологию работы с мультилингвистическими корпоративными базами данных и информационными хранилищами. Данная методика позволяет перейти на новый уровень развития технологий корпоративных информационных систем, учитывающей всестороннюю поддержку ЛПР (профиль пользователя, его требования к необходимой информации).

Предложенный способ организации модели запросов ЛПР в ИУС позволил разработать новый алгоритм формирования и непрерывной

корректировки данной модели с учетом новой информации, получаемой при взаимодействии ЛПР с системой в рамках конкретных производственных ситуаций.

ЗАКЛЮЧЕНИЕ

На основе общих тенденций развития компьютерного моделирования в диссертации решена актуальная проблема, заключающаяся в разработке программно-алгоритмических и информационных средств поддержки мультилингвистической системы проверки уникальности текстов, ориентированной на формирование внутриязыковых и межъязыковых ассоциативных связей. Ее решение базируется на следующих основных результатах:

- выполнен теоретико-информационный анализ структуры информационной базы при компьютерном моделировании процесса мультилингвистического анализа уникальности текстов и показано, что при компьютерном моделировании электронные частотные словари являются основой информационного обеспечения мультилингвистической технологии анализа уникальности текстов на основе лексически связанных компонентов;
- разработаны алгоритмические и информационные средства поддержки мультилингвистического анализа уникальности текстов на базе лексически связанных компонентов с формированием межъязыковых/внутриязыковых ассоциативных связей;
- разработаны структуры частотных лексически связанных словарей как средства информационной поддержки мультилингвистического анализа уникальности текстов и показано, что лексически связанные компоненты обеспечивают целостность восприятия базисных компонентов ИТБ при анализе текста на уникальность;
- разработаны алгоритмы структурно-параметрического синтеза информационно-терминологического базиса как совокупности лексически связанных компонентов и выполнено экспериментальное сопоставление нисходящего и восходящего алгоритмов, что позволило выявить их преимущества и недостатки при выборе эффективного алгоритма с учетом пользовательских требований, предъявляемых к базису при мультилингвистическом анализе уникальности текстов;
- сформирован информационно-терминологический базис в виде электронного двухблочного многоязычного лексически связанного словаря на основе разработанных алгоритмов и компьютерных методов обработки информации. МЛЧС откорректирован и расширен по сравнению с предыдущей версией. Специально разработанная

двухблочная структура словаря позволяет эффективно применять его как в «классическом» виде, так и посредством разработанной автором методики компьютерного анализа уникальности текстов.

В данной работе также решена задача повышения эффективности формирования информационного обеспечения распределенных СППР на базе частотных мультилингвистических тезаурусов как с использованием ресурсов корпоративной сети, так и сети Internet, что имеет существенное значение для развития теории и практики решения задач анализа и обработки мультилингвистической информации. При решении данной задачи получены следующие основные результаты:

- Проведен анализ типовых структур и возможностей СППР, в результате которого можно сделать вывод о том, что в настоящее время недостаточно уделяется внимания проблеме мультилингвистичности представления информации в корпоративных информационных системах. Так же можно указать на недостаточную разработку подходов к оптимизации структуры информационного обеспечения распределенной СППР корпоративного типа.
- Показано, что существующие модели поиска и обработки информации в корпоративных информационно-управляющих системах, как правило, не позволяют эффективно решать задачи формирования многоязычного информационно-лингвистического базиса распределенных систем поддержки принятия решений.
- Предложены новые модели представления и обработки мультилингвистической информации в распределенных информационных системах, обеспечивающие эффективное формирование информационно-лингвистического базиса распределенных систем поддержки принятия решений. Модифицирована модель информационного обеспечения распределенной СППР, основанной на тезаурусе, а также учитывающая весовые коэффициенты терминов используемого мультилингвистического тезауруса. Это позволяет ускорить работу как метопоисковых технологий, так и совершенствовать процесс разработки новых информационных коллекций для частотного информационно-лингвистического базиса СППР.
- Впервые предложена модель синтеза структуры информационного обеспечения распределенной СППР на основе мультилингвистичного тезауруса. Значимость данного подхода в современных условиях очевидна, так как большинство крупных как российских, так и зарубежных компаний работают на международном рынке и, следовательно, при подготовке и принятии решений основываться только на одноязычных данных уже нецелесообразно.

- Предложена новая схема работы поисковой подсистемы распределенной СППР, основанной на мультилингвистических частотных тезаурусах, и использующей несколько режимов: поиск как с частотными, так и без частотных характеристик, метапоиск с использованием тезаурусов для проверки релевантности и ранжирования, а так же проведение поисковой процедуры без использования тезаурусов вообще, что существенно увеличивает возможности предлагаемого решения по сравнению с аналогами.
- Разработана функциональная структура системы формирования и обработки информационно-лингвистического базиса распределенных систем поддержки принятия решений. Система содержит два модуля: информационно-лингвистическая поисковая подсистема и подсистема анализа/синтеза структуры. В рамках разработанных программных систем реализованы предложенные модели и алгоритмы обработки информации.
- Предложен способ организации модели запросов ЛПР в ИУС, который позволил разработать новый алгоритм формирования и непрерывной корректировки данной модели с учетом новой информации, получаемой при взаимодействии ЛПР с системой в рамках конкретных производственных ситуаций.
- Разработан алгоритм формирования и непрерывной корректировки параметров модели ЛПР, который реализован на основе методологии вероятностного латентно-семантического анализа, что позволило учесть семантическую близость терминологического состава последовательных запросов ЛПР при реализации поддержки принятия решений в ИУС современных предприятий и корпораций.
- В ходе эксперимента получены результаты, которые демонстрируют существенно лучшее качество анализа документов на релевантность по сравнению с классическими поисковыми системами. Также показано, что по сравнению с реализацией подхода, основанного на частотных мультилингвистических словарях, степень релевантности определяется незначительно хуже, однако это не снижает значение полученного результата, так как составление тезаурусов значительно менее трудоемкая задача по сравнению с составлением мультилингвистических тематических словарей. Качество ранжирования предлагаемых алгоритмов значительно превосходит качество ранжирования информации классическими поисковыми сервисами, не уступает поисковым системам каталожного типа и системе, построенной на базе тематических частных словарей.

БИБЛИОГРАФИЧЕСКИЕ ССЫЛКИ

1. Авдеева Н.В., Блинова Т.А., Груздев И.А., Ледовская В.М., Лобанова Г.А., Сусь И.В. Допустимый объем заимствований как проблема организации научной деятельности и контроля ее качества. *Открытое образование*. 2018;22(5):74-83. https://doi.org/10.21686/1818-4243-2018-5-74-83
2. Азимов Э.Г. К типологии обучающих компьютерных программ по русскому языку как иностранному //Современные технические средства в обучении русскому языку как иностранному. – М., 1990. – С. 136-143.
3. Алексеев П.М. Частотные словари: Учебное пособие. СПб.: Изд-во С.-Петерб. Ун-та, 2001. 56 с.
4. Алексеев П.М., Турыгина Л.А. Частотный англо-русский словарь-минимум газетной лексики. – М.: Воениздат, 1984.
5. Алексеев П.М. Частотный англо-русский словарь-минимум по электронике. – М.: Воениздат, 1971.
6. Андрианов И.А., Григорьева А.Н. Эффективный поиск плагиата в программном коде для системы дистанционного практикума по программированию // В сборнике: Информатизация инженерного образования. Труды Международной научно-практической конференции ИНФОРИНО-2016. 2016. С. 485-488.
7. Апальков В.Г. Использование современных средств обучения и интернет-технологий в обучении иностранным языкам // Открытое образование. 2013. № 6(101). С. 71-75.
8. Артемова Г.О., Гусарова Н.Ф., Коцюба И.Ю. Автоматизация поддержки принятия решений при разработке онтологий в сфере образования на основе промежуточных моделей. Открытое образование. 2015;(5(112):4-10. https://doi.org/10.21686/1818-4243-2015-5(112-4-10
9. Атанов Г.А. Моделирование учебной предметной области, или предметная модель обучаемого. //Educational Technology & Society, 2001 № 4(1), http://ifets.ieee.org/russian/depository/v4_il/html/4.html
10. Атанов Г.А., Локтюшин В.В. Организация вводно-мотивационного этапа деятельности в компьютерной обучающей системе. //Educational Technology& Society, 2000 №1(1), http://ifets.ieee.org/periodical/vol_1_00/informal_summary_katy_luchini.html
11. Атанов Г.А., Пустынникова И.Н. Структурирование понятий предметной области с помощью методов представления знаний. //Искусственный интеллект, №2, 1997. - с.29-40.
12. Башмаков А.И., Башмаков И.А. Технология и инструментальные средства проектирования тренажерно-обучающих комплексов для

профессиональной подготовки и повышения квалификации. Часть 1. //Информационные технологии. - №6, 1999. – С. 40-45.

13. Башмаков А.И., Башмаков И.А. Технология и инструментальные средства проектирования тренажерно-обучающих комплексов для профессиональной подготовки и повышения квалификации. Часть 2. //Информационные технологии. - №7, 1999. – С. 39-45.

14. Беляков В.М. Разработка функциональной модели автоматизированной обучающей системы по русскому языку как иностранному. – Автореф. дис. … канд.филол.наук. – М., 1996.

15. Бовтенко М.А. Компьютерная лингводидактика: Учеб. пособие. – Новосибирск: Изд-во НГТУ, 2000.

16. Брусиловский П.Л. Интеллектуальные обучающие системы. //Информатика. Информационные технологии. Средства и системы, 1990. №2. – С. 3-22.

17. Брусиловский П.Л. Модели обучения. //Техническая кибернетика, №5, 1992. – С. 97-119.

18. Василенко Н.В., Ковалев И.В., Суздалева Е.А. Информационно-алгоритмическое обеспечение мультилингвистической обучающей технологии //Телематика'2001: Труды международной научно-методической конференции. - Санкт-Петербург, 2001. – С. 68-69.

19. Власов Е.А., Юдина Т.Ф. и др. Компьютеры в обучении языку: проблемы и решения. – М., 2009.

20. Гаврилова Т.А., Зудилова Е.В. Адаптивный диалог и модель пользователя //Диалог-95: материалы международного семинара по компьютерной лингвистике и ее приложениям. – Казань, 1995. – С. 88-97.

21. Галеев И.Х., Н.Х.Галеева Н.Х. Опыт анализа публикационной и издательской активности с использованием РИНЦ//Международный электронный журнал "Образовательные технологии и общество (EducationalTechnology&Society)" -2012. -V.15. -№1. -С.594-608.

22. Галеев И.Х. Практика применения баз данных научного цитирования при оценке публикационной активности КНИТУ//Международный электронный журнал "Образовательные технологии и общество (EducationalTechnology&Society)" -2013. -V.16. -№4. -С.387-402.

23. Галеев И.Х. Анализ юзабилити системы SCIENCE INDEX* //Международный электронный журнал "Образовательные технологии и общество (Educational Technology & Society)" -2015. -V.18. -№4. -С.767-779.

24. Галеев И.Х. Модели и методы построения автоматизированных обучающих систем (обзор). //Информатика. Научно-технический сборник. Серия Кадровое обеспечение. - Вып.1. – М.: ВМНУЦ ВТИ, 1990. – С. 64-72.

25. Гак В.Г. О концепции учебного словаря //Иностранные языки в школе. – №4. - М., 1991. – С. 15-21.

26. Гутгарц Р.Д., Чебышева Б.П. Компьютерная технология обучения //Информатика и образование. - №5, 2000. – С. 44-45.

27. Демушкин А.С. Компьютерные обучающие программы //Информатика и образование. - №3, 1995. – С. 72-76.

28. Домрачев В.Г., Ретинская И.В. О классификации компьютерных образовательных информационных технологий. //Информационные технологии, 1996. №2. – С. 10-14.

29. Драгальчук Е. English-Russian dictionary of computer terms. Англо-Русский словарь компьютерных терминов. [Электрон. ресурс] 2012. Режим доступа: https://slovar-vocab.com/englishrussian/computer-terms-vocab.htm

30. Дудура К.А., Лесков В.О., Огнерубов С.С. Информационная модель динамического формирования электронных терминологических словарей. // Недра Кузбасса. Инновации: Труды V Всероссийской научно-практической конференции. – Кемерово: ИНТ, 2006. – С. 75-76

31. Евсеева Ю.И., Бождай А.С., Гудков А.А. Применение методов интеллектуального анализа данных для реализации рефлексивной адаптации в системах E-LEARNING. *Открытое образование*. 2016;(4). https://doi.org/10.21686/1818-4243-2016-4-

32. Елисеева О. Е. Создание частотного словаря словоформ с помощью инструментов Microsoft Word и Excel. [Электрон. ресурс]. Режим доступа: http://it.lang-study.com/sozdaniechastotnogo-slovarya-word-ex

33. Калянов Г.Н. CASE структурный и системный анализ. Автоматизация и применение. – М.: Изд-во «ЛОРИ», 1996.

34. Калянов Г.Н., Козлинский А.В., Лебедев В.Н. Сравнение и проблема выбора методов структурного системного анализа //PC WEEK/RE. - № 34, 2006.

35. Карасева М.В. Англо-русский частотный словарь по системному анализу. - Красноярск: САА. 1994.

36. Карасева М.В., Ковалев И.В., Лесков В.О. Алгоритмизация процедур включения связанных лексем в структуру информационно-терминологического базиса // Программные продукты и системы, 3, 2009. – С. 35-38.

37. Карасева М.В., Ковалев И.В., Лесков В.О. Внутриязыковые ассоциативные поля в мультилингвистической адаптивно-обучающей технологии. // Системы управления и информационные технологии 3.1(33), 2008. – С. 157-160

38. Карасева М.В., Ковалев И.В., Лесков В.О. Двухблочный трехуровневый электронный англо-немецко-русский частотный словарь по информатике и системному анализу – М.:ВНТИЦ, 2009 - № 50200900111.

39. Карасева М.В., Ковалев И.В., Лесков В.О. Компоненты информационной поддержки мультилингвистической адаптивно-обучающей технологии. //Системы управления и информационные технологии, 1.3(35), 2009. – С. 360-363.

40. Карасева М.В., Лесков В.О. Автоматизация формирования информационной базы мультилингвистической адаптивно-обучающей технологии // Вестник СибГАУ 4(17), 2007. – С. 117-124

41. Карасева М.В., Лесков В.О., Ковалев И.В. Адаптивный алгоритм обучения иностранной лексике на основе лексически связанных компонентов // Системы управления и информационные технологии, 4(34), 2009. – С. 78-82

42. Ковалев И. В., Лесков В. О., Шукшина Е. Е. Процедура построения частотного словаря на основе лексически связанных компонентов // Вестник СибГАУ 2(23). 2009. С. 119-122.

43. Ковалев И. В., Карасева М.В., Лесков В. О. Алгоритмизация процедур включения связанных лексем в структуру информационно-терминологического базиса // Программные продукты и системы 4. 2009. С. 91-94.

44. Ковалев И.В., Лесков В.О. Программа анализа и формирования информационного мультилингвистического терминологического базиса, на основе реляционной модели оптимизации TuMLas v.1.0 – М.: ВНТИЦ, 2008 - № 50200701283.

45. Ковалев И.В., Лесков В.О. Программа контекстного анализа методом «Скрытых Марковских цепей» – М.: ВНТИЦ, 2008 - № 50200501669.

46. Комлева Н.В. Моделирование процесса создания открытых электронных образовательных ресурсов. *Открытое образование*. 2013;(6(101):24-30. https://doi.org/10.21686/1818-4243-2013-6(101-24-30

47. Краснощекова Г.А. Обучение иностранному языку студентов инженерных специальностей на основе реализации контекстного подхода. *Открытое образование*. 2014;(3(104):81-84. https://doi.org/10.21686/1818-4243-2014-3(104-81-84

48. Кривошеев А.О., Голомидов Г.С., Таран А.Н. Перспективные internet-технологии информационного обеспечения образовательных услуг Часть 1. // Информационные технологии. - №7, 1998. – С. 38-44.

49. Кромер В.В. Подпорно-экспоненциальная модель генеральной лексической совокупности английского языка // Деп. в ИНИОН РАН - 18.12.97, № 53134. – Новосибирск: НГПУ, 1997.

50. Кромер В.В. Ядерно-веерная модель вертикального распределения слов в русском языке // Деп. в ИНИОН РАН. – 31.03.97, № 52458. – Новосибирск: НГПУ, 1997.

51. Кручинин В.В. Разработка компьютерных учебных программ. – Томск, 1998.

52. Кудрявцева Е.В. Компьютерные технологии обучения. – http://www.bytic.ttk.ru/cue99M/cu849jujj0.html

53. Лагун И.М., Лукашин О.В. Частотный словарь как инструмент повышения эффективности обучения иностранных студентов // Вестник Тульского государственного университета. Серия Современные образовательные технологии в преподавании естественнонаучных дисциплин. 2012. № 11. С. 115-118.

54. Лесков В.О. Адаптивно-обучающий алгоритм ЛСК-методики – М.:ВНТИЦ, 2009 - № 50200900256

55. Лесков В.О. Восходящий алгоритм формирования ЛС-компонентов – М.:ВНТИЦ, 2008 - № 50200802427.

56. Лесков В.О. Два блока частотного словаря: значение и организация// Успехи современного естествознания №4, 2009. – С. 31-32

57. Лесков В.О. Два подхода к формированию лексически связанных компонентов//Современные наукоёмкие технологии №12, 2008. – С. 29-30

58. Лесков В.О. Комплекс программного моделирования КПМ v.1.0 – М.: ВНТИЦ, 2008 - № 50200802242.

59. Лесков В.О. Комплекс программного моделирования КПМ v.2.0 – М.:ВНТИЦ, 2009 - № 50200900124.

60. Лесков В.О. Нисходящий алгоритм формирования ЛС-компонентов – М.:ВНТИЦ, 2008 - № 50200802428.

61. Лесков В.О. О путях искуственного формирования сложных ассоциативных связей в процессе обучения иностранной лексике. //Современные наукоёмкие технологии №4, 2008. – С. 78-79

62. Лесков В.О. Область применения лексически связанных компонентов// Современные наукоёмкие технологии №2, 2009. – С. 63-64.

63. Лесков В.О. Специфика модели обучения на основе лексически связанных компонентов// Современные наукоёмкие технологии № 4, 2009. – С. 53-54.

64. Лесков В.О. Формирование лексически связанных компонентов информационно-терминологического базиса // Вестник СибГАУ 2(23) 2009. – С. 133-136.

65. Лесков В.О., Огнерубов С.С. Реляционная модель и алгоритмы оптимизации модульной структуры мультилингвистического информационно-терминологического базиса // Вестник Университетского Комплекса, 7(21), 2006. – С. 116-133.

66. Лычагин М.В., Лычагин А.М., Бекарева С.В., Шевцов А.С. Опыт создания частотного словаря современной английской экономической лексики // Вестник Новосибирского государственного университета. Серия: Лингвистика и межкультурная коммуникация. 2006. Т. 4. № 1. С. 103-106.

67. Мерхелевич Г.В. Русско-английский словарь фраз и словосочетаний для обучающихся на английском языке/Г.В. Мерхелевич -Донецк: ЧП АРПИ, 2009. -912 с.

68. Мерхелевич Г. В., Захарова О. А. Обработка иноязычной информации с использованием информационных технологий при ее содержательно-смысловом восприятии и передаче средствами перевода на родной язык // Образовательные технологии и общество. Вып.2, Том 22. 2019. С. 38-43.

69. Позднеев Б.М., Сутягин М.В. Развитие международных стандартов по информационным технологиям в образовании, обучении и подготовке. *Открытое образование*. 2015;(1(108):4-11.

70. Попов К.А. Использование частотных словарей при изучении иностранных языков URL: http://www.openclass.ru/node/6373. 7. 15. 15.

71. Растригин Л.А. Адаптация сложных систем. - Рига: Зинатне, 1981.

72. Растригин Л.А., Эренштейн М.Х. Адаптивное обучение с моделью обучаемого. - Рига: Зинатне, 1988.

73. Ржеуцкая С.Ю., Мальцева Т.В. Программное обеспечение для формирования частотных словарей в процессе об учения английскому языку в техническом вузе // В сборнике: Современные тенденции технических наук материалы II Международной научной конференции. Сер. “Молодой ученый”. 2013. С. 14-16.

74. Русско-английский словарь фраз и словосочетаний для университетов с преподаванием на английском языке/Под редакцией Г.В. Мерхелевича. -Донецк: Норд-Пресс -ЧП АРПИ, 2008. -872с.

75. Скакунова В.А. Информационно-деятельностный подход в обучении иностранным языкам и в повышении качества профессионально-ориентированного образования. *Открытое образование*. 2013;(5(100):24-27. https://doi.org/10.21686/1818-4243-2013-5(100-24-27

76. Смирнова В.Н. Обучение научно-техническому переводу в условиях электронной информационно-образовательной среды в строительном университете. *Открытое образование*. 2019;23(2):4-13. https://doi.org/10.21686/1818-4243-2019-2-4-13

77. Сосновский С.А. ByHeart: система для самообучения иностранным языкам (обзор программного обеспечения). //Educational Technology & Society, 2000 № 3(3), http://ifets.ieee.org/russian/depository/v3_i3/html/6.html

78. Суздалева Е.А. Компьютерные технологии изучения лексики иностранного языка //Решетневские чтения: Материалы Всерос. научно-практ. конф. студентов, аспирантов и молодых специалистов. – Вып. 3. – Красноярск: САА, 1999. – С. 151-152.

79. Суздалева Е.А. Мультилингвистическая информационная технология и основные аспекты ее обеспечения //Решетневские чтения: Материалы всероссийской научно-практической конференции студентов, аспирантов и молодых ученых. – Вып. 5. - Красноярск, 2001. – С. 121-122.

80. Суздалева Е.А. Оптимизация процесса изучения лексики иностранного языка: мультилингвистический подход //Решетневские чтения: Материалы Всерос. научно-практ. конф. студентов, аспирантов и молодых специалистов. – Вып. 4. – Красноярск: САА, 2000. – С. 226-228.

81. Суздалева Е.А. Программно-технические средства поддержки мультилингвистической образовательной технологии //Качество образования: достижения, проблемы. EQ-2001: Материалы четвертой международной научно-методической конференции. - Новосибирск, 2001. – с. 286.

82. Суздалева Е.А. Статистические методы в исследовании профессионально-ориентированной языковой коммуникации //Социально-психологические проблемы развития личности: Материалы первой

всероссийской научной internet-конференции. - Вып. 5. - Тамбов, 2001. – с. 29.

83. Сычев О.А., Мамонтов Д.П. Автоматическое определение ошибок в порядке расположения лексем в ответах на вопросы с открытым ответом в СДО Moodle. *Открытое образование*. 2014;(2(103):79-88. https://doi.org/10.21686/1818-4243-2014-2(103-79-88

84. Сычев О.А., Терехов Г.В. Инструменты помощи автору регулярных выражений для тестовых вопросов в СДО Moodle. *Открытое образование*. 2016;(3):43-50. https://doi.org/10.21686/1818-4243-2016-3-43-50

85. Терентьева И.А. Частотные словари при обучении иностранным языкам. Актуальные проблемы гуманитарных и естественных наук. 2014. № 11-2. С. 185-187.

86. Трембач В.М. Электронные обучающие системы с использованием интеллектуальных технологий. *Открытое образование*. 2013;(4(99)):52-62.

87. Усачев А. В. Нейросетевая мультилингвистическая система адаптивного обучения терминологической лексике : Дис. ... канд. техн. наук : 05.13.01 : Красноярск, 2004 154 с. РГБ ОД, 61:04-5/3244.

88. Филатова Н.Н., Ахремчик О.Л. Центр «Компьютерные технологии образования»: его место в учебном процессе технического университета. //Educational Technology & Society, 2000 № 3(2), http://ifets.ieee.org/russian/depository/v3_i2/html/5.html

89. Харина М.В. Применение частотных словарей в процессе развития англоязычной лексической компетенции студентов ИТ-направлений. *Открытое образование*. 2018;22(5):65-73. https://doi.org/10.21686/1818-4243-2018-5-65-73

90. Чепегин В.И. Компьютерные технологии обучения – проблемы классификации. //Educational Technology & Society, 2000 № 3(4), http://ifets.ieee.org/russian/depository/v3_i4/html/1.html

91. Яблочкин Л.Б., Андриянова М.А., Каргин А.В. Гипертекст – новая технология для создания систем тестирования знаний //Информационные технологии. - №3, 2001. – С. 25-27.

92. Carroll J.B., Davies P., Davies M., Gardner D. A Frequency Dictionary of Contemporary American English: Word Sketches, Collocates, and Thematic Lists. Routledge, 2010. P. 368.

93. Eastment D. Technology-Enhanced Language Learning: Hype or Gold Mine? – http://www.kyoto_su.ac.jp

94. Eisenreich, Günter und Sube, Ralf. Langenscheidts Fachwörterbuch Mathematik: Englisch-Deutsch-Französisch-Russisch. - Langenscheidt, 1996.

95. Ferretti, Vittorio. Wörterbuch der Datentechnik: Englisch-Deutsch, Deutsch-Englisch; 9200 Einträge aller Gebiete der Informatik, mit Hauptbegriffen der angrenzenden Fachgebiete sowie des allgemeinen technischen Sprachgebrauchs (insgesamt 94 Fachgebiete). - Berlin: Springer, 1996.

96. Ferretti, Vittorio. Wörterbuch der Elektronik, Datentechnik und der Telekommunikation: Deutsch-Englisch. - Berlin: Springer, 1992.

97. Goodfellow R. CALL Programs for Vocabulary Instruction //Computer Assisted Language Learning Journal. – Vol.8. – №2, 1995. – Pp.205-226.

98. Haberfeller K., Daenzer W. Systems engineering: Methodik und Praxis. - Betr.Inst.d.ETH. - Zürich: Verl.Industrielle Organisation, 2007.

99. Herdan G. Quantitative Linguistics. - London, 1964.

100. Hoffmann L. Anwendungsmoeglichkeiten und bisherige Anwendung von statistischen Methoden in der Fachsprachenforschung//In: Languages for Special Purposes (Edited by L. Hoffmann). - Vol. 1. - Berlin, 1999. – Pp. 241-249.

101. Jones G. Computer Simulations in Language Teaching – the Kingdom Experiment //System. – Iss.17, 1986. – Pp.35-47.

102. Kotik, Michail. Wörterbuch Luft- und Raumfahrttechnik: Russisch-Englisch-Deutsch. - Düsseldorf, 1986.

103. Leech G., P. Rayson, A. Wilson. Word Frequencies in Written and Spoken English: based on the British National Corpus. London: Longman, 2001. P. 320.

104. Legenhausen L., Wolff D. Call in Use – Use of CALL: Evaluating CALL Software // System. – Iss.18, №1, 2000. – Pp.1-13.

105. Levy M. Computer Assisted Language Learning: Context and Conceptualization. – Oxford: Clarendon Press, 2007.

106. Mueller Ch. Initiation a la statistique linguistique. - Paris, 1978.

107. Peterson M. Creating Hypermedia learning Environments: Guidelines for Designers. – http://www.jaist.ac.jp

108. Souzdaleva E. Effective Information Training Technology Based on the Learner's Memory State Model //In: AMSE Periodicals, 2000 – Vol.21, № 3. - Pp.11-26.

Научное издание

Ковалев Игорь Владимирович

МУЛЬТИЛИНГВИСТИЧЕСКИЙ АНАЛИЗ УНИКАЛЬНОСТИ ТЕКСТОВ НА БАЗЕ ЛЕКСИЧЕСКИ СВЯЗАННЫХ ИНФОРМАЦИОННЫХ КОМПОНЕНТОВ

Монография

В авторской редакции

Подп. в печать 20.05.2020. Формат 60х84/16. Бумага офсетная.
Печать цифровая. Усл. печ. л. 6,9. Уч.-изд. л. 8,7.
Тираж 100 экз. Заказ 2180.

Издательство ОУ «Красноярский краевой Дом науки и техники Российского Союза научных и инженерных общественных объединений» (ОУ «ККДНиТ»)
660049, г. Красноярск, улица Урицкого, д. 61.
Тел. (391) 227-85-00, info@domnit.ru
www.domnit.ru

Отпечатано в типографии «Крафт»,
ООО «Палитра», 660050, Красноярский край, Красноярск,
улица Кутузова, дом № 1, строение, 37;
тел. +7 (391) 209-68-28, e-mail: 2096828@mail.ru;
www.kraft-pt.ru, www.4uprint.ru