Федеральное государственное образовательное учреждение высшего профессионального образования – «Национальный исследовательский технологический университет «МИСиС»

На правах рукописи

Буштедт Владислав Андреевич

МОДЕЛЬ ПРИНЯТИЯ РЕШЕНИЯ НА ОСНОВЕ СИНТАКСИЧЕСКОГО АНАЛИЗА В ЗАДАЧАХ ОБРАБОТКИ ПАТЕНТНОЙ ИНФОРМАЦИИ

Специальность 05.13.01

«Системный анализ, управление и обработка информации (в производственной сфере)»

АВТОРЕФЕРАТ

диссертации на соискание ученой степени кандидата технических наук

Москва 2011

Работа выполнена на кафедре АСУ Федерального государственного образовательного учреждения высшего профессионального образования - «Национального исследовательского технологического университета «МИСиС».

Научный руководитель:

к.т.н., доцент Поляков В. Н.

Официальные оппоненты:

Прошин Иван Александрович, д.т.н., профессор.

Филиппович Андрей Юрьевич, к.т.н., доцент.

Ведущая организация:

Федеральное государственное автономное образовательное учреждение высшего профессионального образования «Казанский (Приволжский) федеральный университет»

Защита состоится «16» ноября 2011 г. в 14 часов на заседании Диссертационного совета Д.212.132.07 при Национальном исследовательском технологическом университете «МИСиС» по адресу: 119049, Москва, ул. Крымский Вал, 3, ауд. К-325.

С диссертацией можно ознакомиться в библиотеке МИСИС.

Автореферат разослан « » 2011 г.

Ученый секретарь Диссертационного Совета — к.т.н., профессор Калашников Е.А.

ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

Последнее десятилетие ознаменовано феноменальным прогрессом в области телекоммуникаций, электронного документооборота и автоматизации работы с информацией. Это, в свою очередь, вызвало бурный рост объемов информации в сети Интернет, в системах документооборота и архивах организаций, в том числе на предприятиях металлургического комплекса. Стало возможной организация удаленного доступа к различным библиотечным ресурсам: различным электронным библиотекам, подборкам статей, базам данных патентных документов и т. д.

Такой рост объема информации, происходящий одновременно с ростом информационных потребностей пользователей и общей тенденцией к понижению требований к их квалификации в области организации поискового процесса, ставит старую проблему эффективного информационного поиска остро как никогда ранее. Необходимо отметить, что, несмотря на непрерывно идущие исследования в данной области и совершенствование поисковых технологий (о чем косвенно может свидетельствовать постоянное появление новых информационно-поисковых систем в сети Интернет), нельзя сказать, что поставленная проблема близка к своему решению.

Так, например, в настоящее время большинство запросов к поисковой системе состоят из более, чем одного слова, и этот показатель растет со временем. Для поисковой системы Яндекс на момент написание данной работы в среднем каждый запрос состоял из трех слов. При этом за последний год этот показатель вырос на 0,5 слова и продолжит расти. Использование словосочетаний позволяет во многих случаях снять лексическую многозначность запросов. Словосочетание же является минимальной синтаксической конструкцией. Это подтверждает актуальность исследования и использования моделей синтаксического анализа в поисковых технологиях.

В настоящий момент для информационно-поисковых систем очевидны следующие области применения:

- патентный поиск;
- библиотечный поиск;
- поиск в системах документооборота предприятий;
- поиск в хранилищах текстовой информации (новости, научные ресурсы);
- поиск в Интернет и др.

Синтаксический анализ является частью задачи автоматического анализа текста на естественном языке в информационно-поисковых системах.

Попытки создания синтаксического анализатора для русского языка велись еще в конце 1960-х годов. Но быстродействие ЭВМ того времени явилось основным препятствием для реализации сложных алгоритмов анализа в полном объеме. Исследователям того времени приходилось упрощать алгоритмы, например, отказываясь от перебора всех омонимичных вариантов в тексте, что в свою очередь приводило к малой точности синтаксического анализа предложения.

Данная задача не решена полностью и в настоящее время. Одной из причин этого является сложность описания семантических моделей, влияющих на сочетаемость лексических единиц, а также то, что до недавнего времени большое число исследователей сходились во мнении о нецелесообразности введения модуля синтаксического разбора в системы автоматического анализа текста. Однако оказалось, что, несмотря на ограниченную

точность синтаксических анализаторов, их использование способно заметно повысить качество таких систем в случае комбинирования с известными статистическими методами. Современным исследователям также приходится искать компромисс между следующими параметрами при синтаксическом анализе:

- полнота анализа – степень описания при помощи синтаксических связей любого предложения;
- точность анализа – доля ошибок в созданных анализатором структурах предложения;
- быстродействие – скорость работы анализаторов текста: несмотря на революционное развитие компьютерной техники за последние 50 лет, в области лингвистики существуют такие прикладные задачи, которые не могут быть решены в приемлемое для конкретных прикладных задач время.

В настоящее время в России проводятся мероприятия, направленные на развития 4 основных направлений модернизации: институты, инфраструктура, инновации и инвестиции. Данная концепция развития была предложена Президентом РФ Дмитрием Медведевым. Для развития инновационного подхода необходимо увеличение интенсивности развития наукоемких производств, что невозможно без разработки новых эффективных методов обработки информации.

Сейчас положение дел в России обстоит таким образом, что проблема с соблюдением авторских прав на изобретения и другие виды интеллектуальной собственности стоит наиболее остро в научной среде. Большое количество полезных изобретений так и не выходят за пределы лабораторий, так как недобросовестные конкуренты, незаконно воспользовавшись идеями изобретателя, пока тот пытается в течение более года оформить патент, налаживают выпуск своих собственных продуктов.

Очень важным для любого изобретения является правильное и быстрое оформление права на него. Для этой цели существует патент. Он необходим для того, чтобы защитить рынок, исключить возможность незаконного использования товара третьим лицом. Патент дает исключительное право на изобретение. Использование изобретения третьим лицом без согласования с владельцем патента преследуется по закону.

Поисковые технологии с использованием моделей синтаксического анализа способны дать существенный выигрыш по времени при проведении патентного поиска.

Поэтому задача разработки системы качественного и быстрого патентного поиска с использованием современных поисковых технологий на основе синтаксического анализа в настоящее время является актуальной.

Таким, образом, **актуальность работы** определяется следующим:

- Необходимостью создания информационных систем патентного поиска с использованием моделей синтаксического анализа;
- Возросшей вычислительной мощностью современных компьютеров, что позволяет решать задачи синтаксического анализа с использованием подходов, требующих больших вычислительных ресурсов, но обеспечивающих более высокое качество анализа;
- Накопленным опытом создания подобных систем, позволяющим предложить новые решения на основе блочного подхода к синтаксическому анализу, проводить частичный синтаксический анализ с использованием ограниченного количества правил.

Цель работы заключается в исследовании особенностей документооборота в области патентного поиска, а также в моделировании процессов синтаксического разбора и создании моделей принятия решения при выборе патентов аналогов.

Для достижения поставленной цели были рассмотрены и решены следующие **задачи**:

- Изучены информационные потоки и особенности патентного поиска.
- Изучены различные грамматики, позволяющие описать синтаксическую структуру предложения.
- Созданы методы и алгоритмы частичного синтаксического анализа текста на русском языке.
- Создан программный комплекс, выполняющий поиск и выделение чанков с именами существительными из предложения.
- Разработаны эвристики, улучшающие точность работы программного комплекса.
- В целях апробации разработанного метода создан прототип системы принятия решения при выборе патентов аналогов.

Научная новизна диссертационного исследования заключается в следующем:

- Разработана модель и алгоритм принятия решения патентным поверенным в области патентного поиска при выборе патентов-аналогов.
- Усовершенствована модель частичного синтаксического анализа, основанная на блочном подходе.
- Предложены и формально описаны эвристики, улучшающие качество синтаксического анализа.
- Описана математическая постановка задачи частичного синтаксического анализа в логико-математической нотации.
- Предложена расширенная нотация математической постановки задачи частичного синтаксического анализа с использованием инструментария модификаторов грамматических категорий.
- Экспериментально выявлен вклад каждой эвристики в результаты синтаксического анализа.
- Решена задача принятия решения в процессе патентного поиска с использованием модели частичного синтаксического анализа.

Теоретическая значимость заключается в следующем:

- Выполнена формальная постановка задачи для разработки системы принятия решения в области патентного поиска, позволяющая выделять наиболее релевантные заданному условию поиска патенты-аналоги.
- Построена интегральная модель синтаксического анализа, основанная на последовательной системе фильтров.
- Предложены и формализованы эвристики, повышающие точность работы частичного синтаксического анализатора.

Практическая значимость заключается в следующем.

- Результаты работы нашли применение в области патентного поиска и могут быть использованы в различных системах электронного документооборота, предполагающих поиск документов на основе сложных синтаксических конструкций.
- В рамках данной диссертационной работы создан прототип системы поддержки принятия решений.
- Проведена апробация частичного синтаксического анализатора в рамках задачи патентного поиска.

Методы исследования

При разработке программного комплекса использовались:

- Методы системного анализа и принятия решений.
- Методы математической логики.
- Элементы теории множеств.
- Методы реляционной алгебры и методы статистического анализа для формирования базы правил в рамках системы принятия решений.
- Методы дистрибутивного анализа, в частности метод формирования новых эвристик, основанный на группировании омонимичных чанков.
- Методы алгоритмического моделирования и методы объектно-ориентированного программирования для построения опытного образца системы «Find-chunk».
- Методы оценки качества работы предложенной модели с использованием меры F_1.

Результаты работы были практически реализованы в виде программного комплекса, включающего в себя набор инструментов для частичного синтаксического анализа текста, анализатора омонимичных чанков и поиска патентов аналогов на основе запрашиваемого текста, представляющего собой сформулированную на естественном языке примерную формулу изобретения.

Результаты работы нашли применение в образовательном процессе при подготовке лабораторного практикума по курсу «Лингвистические основы информатики» для специальностей «Прикладная информатика» и «Автоматизированные системы управления» в НИТУ «МИСиС».

Результаты работы были приняты к внедрению в рамках проекта по созданию Базы знаний по тематическому направлению деятельности национальной нанотехнологической сети «Конструкционные наноматериалы» для целей анализа методов и технологий, а также сравнения научно-технических решений в указанной области. Работа ведется в рамках Федеральной целевой программы «Развитие инфраструктуры наноиндустрии в Российской Федерации на 2008—2011 годы» (ФЦПНано, Госконтракт № 16.647.11.2024).

Апробация работы

Результаты работы докладывались на следующих научных конференциях:

- VIII Казанская школа-семинар по компьютерной и когнитивной лингвистике (TEL-2006), Казань, Россия, 2006 г.
- Международная конференция «Когнитивное моделирование в лингвистике» (CML-2007), София, Болгария, 2007 г.
- X Казанская школа-семинар по компьютерной и когнитивной лингвистике (TEL-2008), Казань, Россия, 2008 г.

- Международная конференция «Когнитивное моделирование в лингвистике». CML-2008, Бечичи, Черногория, 2008 г.
- 64-е Дни науки в МИСиС, Москва, МИСиС, 2009 г.
- XI Казанская школа-семинар по компьютерной и когнитивной лингвистике (TEL-2009), Казань, Россия, 2009 г.
- 65-е Дни науки в МИСиС, Москва, НИТУ «МИСиС», 2010 г.
- Международная научная конференция «Перспективные технологи, оборудование и аналитические системы для материаловедения и наноматериалов», Волгоград, 2009 г.
- 66-е Дни науки в МИСиС, Москва, НИТУ «МИСиС», 2011 г.

Работа дважды проходила экспертизу Российского фонда фундаментальных исследований (РФФИ) и выполнялась при финансовой поддержке Фонда в рамках проектов:

- Грант № 05-07-90339-в, Тема «Система онтологического типа для поиска и обработки текстовой информации», 2005 -2007;
- Грант № 09-07-97007-р_поволжье_а, Тема «Модель извлечения информации из текстов на основе онтологии энциклопедических знаний», 2009 -2011.

Работа соответствует паспорту специальности 05.13.01, и выполнена в следующих областях исследования:

- Формализация и постановка задач системного анализа, оптимизации, управления, принятия решений и обработки информации.
- Разработка методов и алгоритмов решения задач системного анализа, оптимизации, управления, принятия решений и обработки информации.
- Разработка специального математического и программного обеспечения систем анализа, оптимизации, управления, принятия решений и обработки информации.
- Теоретико-множественный и теоретико-информационный анализ сложных систем.
- Методы и алгоритмы интеллектуальной поддержки при принятии управленческих решений в технических, медицинских и социальных системах.
- Визуализация, трансформация и анализ информации на основе компьютерных методов обработки информации.

Структура диссертации

Работа состоит из введения, 3 глав, заключения, списка литературы и двенадцати приложений.

На защиту выносится:

- Модель и алгоритм принятия решения патентным поверенным в области патентного поиска при выборе патентов-аналогов.
- Модель частичного синтаксического анализа, основанная на блочном подходе.
- Множество эвристик, улучшающих качество синтаксического анализа.
- Математическая постановка задачи частичного синтаксического анализа в логико-математической нотации.
- Расширенная нотация математической постановки задачи частичного синтаксического анализа с использованием инструментария модификаторов грамматических категорий.

- Программный комплекс «Find-Chunk», разработанный в рамках диссертационной работы для решения задач, связанных с областью патентного поиска с использованием частичного синтаксического анализа.

ЗАДАЧА СИНТАКСИЧЕСКОГО АНАЛИЗА ТЕКСТОВ В СИСТЕМАХ ДОКУМЕНТООБОРОТА И ПРИНЯТИЯ РЕШЕНИЙ

В первой главе рассматриваются различные технологии обработки естественного языка для поддержки принятия решений применительно к АСУ и АСУ ТП металлургических предприятий и патентного поиска.

В настоящее время патентный поиск производится вручную патентоведом с минимальным применением автоматических средств анализа. В среднем для проведения качественного предварительного патентного поиска по одному изобретению необходимо около 10-15 рабочих дней. Стоимость патентного поиска при этом составляет от одной до нескольких десятков тысяч рублей. Также необходимо заметить, что один человек не в состоянии просмотреть все имеющийся патенты в некоторой предметной области, поэтому ему приходится существенно сузить количество патентов для анализа для того, чтобы иметь возможность произвести его за некоторое приемлемое время с приемлемыми затратами ресурсов. Для проведения же полной экспертизы, как было указано выше, необходимо около 6 месяцев. В последнее время сфера патентного поиска попадает в фокус интересов исследователей поисковых технологий.

На основе всестороннего анализа существующей литературы в первой главе делаются следующие выводы:

1. Технологии обработки текстов на естественном языке в совокупности с теорией принятия решений имеют высокий потенциал для повышения эффективности патентного поиска.

2. Задача обработки текстовой информации даже не в полном объеме (например, получения неполных деревьев зависимости в предложениях) стоит в настоящее время очень остро в силу того, что разработать в обозримом будущем системы для исчерпывающего синтаксического анализа текста будет, вероятнее всего, невозможно.

3. Среди большого многообразия грамматик и формализмов для описания и обработки естественного языка наиболее предпочтительными для обработки текстов на русском языке представляются грамматики зависимости. Это объясняется, во-первых, естественным характером представления синтаксических связей в виде дерева зависимости, и, во-вторых, тем, что грамматики зависимостей больше чем другие подходят для языков со свободным порядком слов, к которым относится русский язык.

4. Несмотря на наличие общих подходов к синтаксическому анализу (СА), многие проблемы, возникающие в процессе СА (неоднозначности различного рода, непроективность, свободный порядок слов), не имеют пока общего теоретического решения.

5. Применение трибанков для повышения качества синтаксического анализа имеет существенный потенциал, однако корпусные исследования очень трудоемкие и рассчитаны на многолетний период.

6. Из-за явлений омонимии, синтаксической неоднозначности и непроективности пока не предложено алгоритма, гарантирующего полный и корректный парсинг.

7. Среди предложенных, обычно более быстрые алгоритмы дают больший процент ошибки, и наоборот - менее быстрые дают меньший процент ошибок.

8. Наличие головы (главного узла) – один из важнейших элементов разбора зависимостей.

9. Эффективность алгоритма (с точки зрения минимизации ошибок) может быть существенно улучшена за счет различных эвристик.

10. Примеры работы синтаксических анализаторов показывают актуальность проблемы совершенствования модели синтаксического анализа.

11. Чанкинг может выступать хорошей альтернативой полному СА в случаях, когда условия решения задачи синтаксического анализа не требуют построения полного дерева. В задачах поиска патентной информации, когда поисковый образ представляется довольно громоздкой синтаксической конструкцией формулы изобретения, дополнительные синтаксические сведения в виде набора чанков могут существенно улучшить качество поиска.

В главе приводится обзор существующих синтаксических анализаторов:

- Синтаксический анализатор «Syntax»;
- Синтаксический анализатор «Dictum»;
- Система ЭТАП-3;
- Система ПРОМТ;
- Синтаксический анализатор «Treevial».

ТЕОРЕТИЧЕСКИЕ ОСНОВЫ ПАТЕНТНОГО ПОИСКА С ИСПОЛЬЗОВАНИЕМ ЧАСТИЧНОГО СИНТАКСИЧЕСКОГО АНАЛИЗА

Во второй главе даётся подробное описание теоретических основ патентного поиска с использованием частичного синтаксического анализа.

Математическая постановка задачи принятия решения в процессе патентного поиска

Задача принятия решения при выборе документов-аналогов в области патентного поиска решается с использованием системы синтаксического анализатора. Под выбором документа-аналога понимается выбор релевантного запросу пользователя патента-аналога.

Под релевантностью патента R_i понимается численная оценка программой этого патента с точки зрения степени его удовлетворения условиям запроса.

Необходимо найти такой набор чанков H*, обнаруженных в поисковом запросе, который обеспечит выполнение следующих условий (1, 2, 3):

$$\max_{h \subset H} R(h), \quad \max_{h \subset H}(-|M(h)|), \quad H = \{h : H \in \psi, R(h) > 0, M(h) > 0\}, \qquad (1)$$

где R(h) – суммарная релевантность множества найденных патентных документов, M(h) – множество результатов (найденных патентов), H – множество результативных наборов чанков, ψ – множество всех обнаруженных в поисковом запросе чанков.

$$R(h) = \sum_{i=1}^{m} R_i, \qquad (2)$$

где $m = M : R_i > R_0$.

R_0 – пороговый уровень релевантности для патентов, при значениях релевантности ниже которого патенты считаются нерелевантными поисковому запросу.

При сортировке патентов по релевантности используется индекс релевантности, который рассчитывается следующим образом:

$$R_i = (k_1*(N_ч/N_{ч\_max})+k_2*(N_{сл}/N_{сл\_max}))/(k_1+k_2), \qquad (3)$$

где k_1 – весовой коэффициент для чанков;

$N_ч$ –число чанков, которые встретились в документе I;

$N_{ч\_max}$ – максимальное число чанков во всем пуле документов;

k_2 – весовой коэффициент для ключевых слов;

$N_{сл}$ – число ключевых слов, которые встретились в документе I;

$N_{сл\_max}$ – максимальное число ключевых слов во всем пуле документов.

На рисунке 1 представлен обобщенный алгоритм сценария поиска патентов-аналогов.

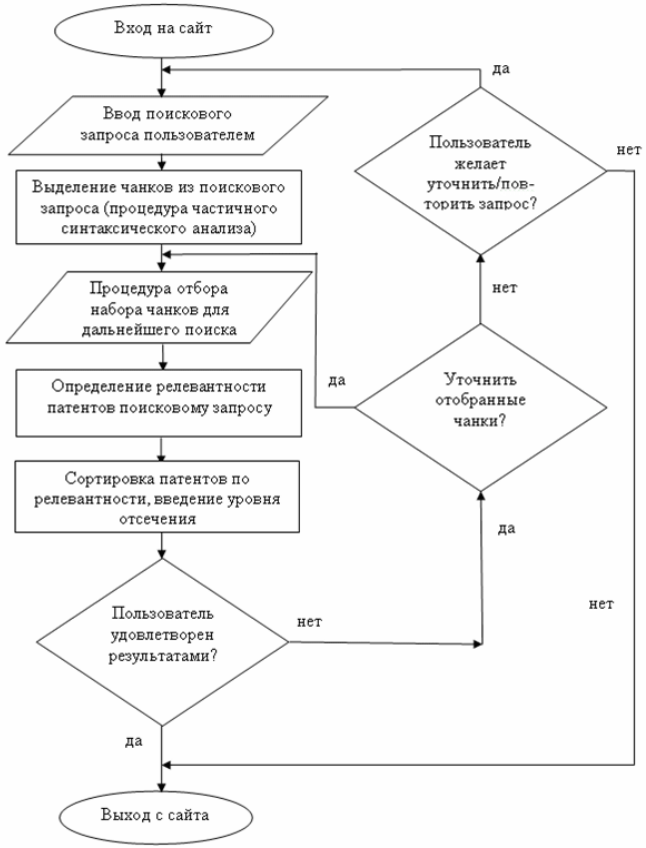

Рисунок 1 – Обобщенный алгоритм сценария патентного поиска

Задача синтаксического анализа текстов

Синтаксический анализ является частью системы полного автоматического анализа текстов на естественном языке.

Как отмечалось выше, систему полного синтаксического анализа текстов на естественном языке построить достаточно сложно из-за ряда причин, главной из которых является неоднозначность. Для многих прикладных задач, в том числе для задач патентного поиска и извлечения знания из текстов, оказывается достаточно частичного синтаксического анализа. Более того, частичный синтаксический анализ позволяет сократить время обработки текстов и, следовательно, принятия решения.

Было принято решение о проведении исследований, связанных с работой частичного синтаксического анализа, так называемого «Чанкера» (от англ. слова «chunk» - глыба, ломоть – то есть нечто грубое и общее, в смысле частичного синтаксического анализа по сравнению полным). Преимущества данного подхода заключаются в том, что для описания процесса синтаксического анализа требуется минимум грамматических правил и словарей.

Рассмотрим пример предложения: ***С*** *помощью электронного микроскопа можно изучать структуру наноматериалов.*

В данном предложении можно выделить следующие чанки: с *помощью микроскопа, электронного микроскопа, можно изучать, изучать структуру, структуру наноматериалов.*

Для улучшения качества работы частичного синтаксического анализатора было принято решение ввести в модель три группы эвристик.

Первая группа («А») работает на этапе поиска кандидатов в чанки, анализируя возможность существования каждого конкретного чанка в зависимости от окружения слов, входящих в него. Они базируются на анализе грамматической информации и носят лингвистический характер.

Эвристики из второй группы («В») являются по своей сути фильтрами. Они начинают свою работу после того, как для всего сегмента сформирован набор кандидатов в чанки. Эвристики из этой группы также носят лингвистический характер и принимают решение о возможности существования каждого отдельного кандидата в чанки, опираясь на информацию обо всех остальных кандидатах в чанки для анализируемого сегмента. Также эти эвристики обнаруживают сложные чанки, например, такие, в состав которых входит составной глагол.

Часть эвристик этой группы имеет ряд ограничений. Например, работа одной из эвристик основана на использовании информации о наличии подлежащего и сказуемого в анализируемом сегменте. Подлежащее и сказуемое же в сегменте определяется только в том случае, если подлежащим является существительное, а сказуемым – глагол при их одновременном присутствии в одном сегменте. Логика функционирования данной эвристики допускает такое упрощение.

Эвристики из третьей группы («С») также по своей сути являются фильтром. Эти эвристики основаны на математических свойствах дерева синтаксического подчинения. Они работают после эвристик второй группы.

В работе дается вербальное описание алгоритма частичного синтаксического анализа.

Используется блочный алгоритм для проведения синтаксического анализа. В настоящей работе рассматриваются применение всех блоков из приведенного алгоритма за исключением ролевых и контекстных фильтров.

В работе сформулированы условия проверки сочетаемости слов для построения чанков. Подробное описание эвристик с примерами представлено в диссертации.

Математическая постановка задачи частичного синтаксического анализа (чанкинга)

Рассмотрим математическую постановку задачи частичного синтаксического анализа (чанкинга). Адаптированная математическая постановка задачи в упрощенной нотации приведена в Приложении К диссертационной работы:

а) Сегмент можно представить в виде упорядоченного множества слов (словоформ)

$$S = \{w_1, w_2, w_3, ..., w_n\} \quad (4)$$

и заданного на этом множестве отношения порядка

$$N_1 < N_2 < N_3 < ... < N_n, \quad (5)$$

где N_i – место слова w_i в сегменте;

n – количество слов в сегменте.

б) Этап морфоанализа можно представить как [1]

$$(w_i^0, G_i) = MA(w_i), \quad (6)$$

где w_i^0 – нормальная форма слова;

G_i – кортеж грамматических характеристик:

$G_i =< PS_i, Gender_i, Case_i, Number_i, General_i, Subject_i, Predicate_i >$;

$MA(w_i)$ – функция морфоанализа.

Здесь:

- PS – *признак_части_речи*: $PS \in \{"noun", "verb", "article", "adjective", "participle", "gerund", "pronoun", "numeral", "adverb", "preposition", "conjunction"\}$.
- $Gender$ – *признак_рода*: $Gender \in \{"f", "m", "n"\}$.
- $Case$ – *признак_падежа*: $Case \in \{"nominative", "genitive ", "dative", "accusative", "instrumental", "prepositional"\}$.

[1] Расшифровка всех обозначений и переменных приведена в Приложении Б диссертации.

- *Number – признак _ числа*:
  $Number \in \{"singular","plural"\}$.
- *General – признак _ главного _ слова _ в _ чанке*:
  $General \in \{"true","false"\}$ [1].
- *Subject – признак _ подлежащего*:
  $Subject \in \{"true","false"\}$ [2].
- *Predicate – признак _ сказуемого*:
  $Predicate \in \{"true","false"\}$ [2].
- *Infinitive – признак _ инфинитива _ глагола*:
  $Infinitive \in \{"true","false"\}$.

в) Теперь сегмент (выражение (4)) может быть представлено в виде множества пар

$$T = \{(w_1^0, G_1),(w_2^0, G_2),...,(w_n^0, G_n)\} \quad (7)$$

и заданного на этом множестве отношения порядка (2).

г) Расстояние между словами в сегменте определяется как

$$Z = |i - j| \quad (8)$$

где i, j – порядковый номер в предложении слов w_i, w_j, которые анализируются в каждый момент;

д) Поиск чанка (связанного словосочетания) сводится к перебору всех комбинаций пар в сегменте и проверке выполнения условий.

$$\underset{Z \le Z_0\_для\_u_1,u_2,u_4,u_5,u_6}{Comp}(U_{ij},A_{ij},B_{ij},C_{ij}) = \begin{cases} \text{True, если } (U_{ij}=True) \wedge (A_{ij}=True) \wedge (B_{ij}=True) \wedge (C_{ij}=True), \\ \text{False, если } (U_{ij}=False) \vee (A_{ij}=False) \vee (B_{ij}=False) \vee (C_{ij}=False), \end{cases} \quad (9)$$

где $\underset{Z \le Z_0\_для\_u_1,u_2,u_4,u_5,u_6}{Comp}(U_{ij},A_{ij},B_{ij},C_{ij})$ – логическая функция сравнения;

Z_0 – область поиска чанков в сегменте;

U_{ij} – условия для первоначального поиска чанков;

A_{ij} – условия, описывающие эвристики, работающие на этапе поиска каждого чанка. Базируются на грамматических категориях.

B_{ij} – условия, описывающие эвристики, работающие после окончания поиска всех чанков в сегменте. Имеют в своей основе лингвистические правила;

[1] Главное слово в чанке – то слово, от которого производится процесс поиска возможного зависимого от него слова в синтаксическом сегменте.

[2] Подлежащее и сказуемое в предложении определяются только в случае одновременного присутствия в синтаксическом сегменте существительного в именительном падеже и глагола, которые образуют чанк. При наличии нескольких кандидатов на подлежащее и сказуемое в составе одного синтаксического сегмента соответствующие метки получают только слова, входящие в состав первого чанка.

C_{ij} – условия, описывающее эвристику, работающую после окончания поиска всех чанков в сегменте. Имеет в своей основе математические правила.

Также данная модель может интерпретироваться в терминах логических модусов вида «Модус Поненс».

$$\frac{\begin{array}{l} Если\_\Phi, то\_\Psi \\ \Phi, \end{array}}{\Psi,} \quad (10)$$

где Φ, Ψ – произвольные высказывания, являющиеся соответственно основанием и следствием импликативного высказывания вида *Если Ф, то* Ψ.

В терминах Модуса Поненса высказывание будет выглядеть следующим образом:

$$\frac{\begin{array}{l} Если\_(U_{ij}) \wedge (A_{ij}) \wedge (B_{ij}) \wedge (C_{ij}), \\ то\_\underset{Z \le Z_0\_для\_u_1,u_2,u_4,u_5,u_6}{Comp}(U_{ij},A_{ij},B_{ij},C_{ij}) \\ (U_{ij}) \wedge (A_{ij}) \wedge (B_{ij}) \wedge (C_{ij}) \end{array}}{\underset{Z \le Z_0\_для\_u_1,u_2,u_4,u_5,u_6}{Comp}(U_{ij},A_{ij},B_{ij},C_{ij})} \quad (11)$$

е) Рассмотрим условия, необходимые для первоначального поиска чанков:

$$\begin{array}{l} U_{ij}(u_{1i,j},u_{2i,j},u_{3i,j},u_{4i,j},u_{5i,j},u_{6i,j}) = True, \text{ если} \\ (u_{1i,j}=True) \vee (u_{2i,j}=True) \vee (u_{3i,j}=True) \vee \\ \vee (u_{4i,j}=True) \vee (u_{5i,j}=True) \vee (u_{6i,j}=True) \end{array} \quad (12)$$

В терминах Модуса Поненса высказывание будет выглядеть следующим образом:

$$\frac{\begin{array}{l} Если\_(u_{1i,j}) \vee (u_{2i,j}) \vee (u_{3i,j}) \vee \\ \vee (u_{4i,j}) \vee (u_{5i,j}) \vee (u_{6i,j}), \\ то\_U_{ij}(u_{1i,j},u_{2i,j},u_{3i,j},u_{4i,j},u_{5i,j},u_{6i,j}) \\ (u_{1i,j}) \vee (u_{2i,j}) \vee (u_{3i,j}) \vee \\ \vee (u_{4i,j}) \vee (u_{5i,j}) \vee (u_{6i,j}) \end{array}}{U_{ij}(u_{1i,j},u_{2i,j},u_{3i,j},u_{4i,j},u_{5i,j},u_{6i,j})} \quad (13)$$

Пусть изначально для каждого кандидата в чанки:

$$u_{1i,j} = False, \forall i,j \in 1..n; \quad (14)$$

$$u_{2i,j} = False, \forall i,j \in 1..n; \quad (15)$$

$$u_{3i,j} = False, \forall i,j \in 1..n; \quad (16)$$

$$u_{4i,j} = False, \forall i,j \in 1..n; \quad (17)$$

$$u_{5i,j} = False, \forall i,j \in 1..n; \quad (18)$$

$$u_{6i,j} = False, \forall i,j \in 1..n; \quad (19)$$

Сформулируем условия истинности.

«U.1». Имя Существительное – Имя Прилагательное. Совпадает род, число и падеж.

$$
\begin{aligned}
&((PS_i = "noun") \wedge \\
&\wedge (PS_j = "adjective") \wedge \\
&\wedge (Gen_i = Gen_j) \wedge \\
&\wedge (Num_i = Num_j) \wedge \\
&\wedge (Case_i = Case_j)) \Rightarrow (u_{1i,j}) \wedge P_{chunk}(w_i, w_j)
\end{aligned} \tag{20}
$$

где $P_{chunk}(w_i, w_j)$ - предикат, устанавливающий истинность синтаксических отношений между w_i и w_j.

Условия «U.2» … «U.6» представлены на стр. 109-111 диссертации.

ж) Рассмотрим условия, описывающие эвристики, работающие на этапе поиска каждого чанка:

$$A_{ij}(a_{1i,j}, a_{2i,j}, a_{3i,j}, a_{4i,j}) = True, \text{ если}$$

$$(a_{1i,j} = True) \wedge (a_{2i,j} = True) \wedge (a_{3i,j} = True) \wedge (a_{4i,j} = True) \tag{21}$$

В терминах Модуса Поненса высказывание будет выглядеть следующим образом:

$$
\begin{aligned}
&Если\ \_(a_{1i,j}) \wedge (a_{2i,j}) \wedge (a_{3i,j}) \wedge (a_{4i,j}), \\
&mo\ \_A_{ij}(a_{1i,j}, a_{2i,j}, a_{3i,j}, a_{4i,j}) = True \\
&\frac{(a_{1i,j}) \wedge (a_{2i,j}) \wedge (a_{3i,j}) \wedge (a_{4i,j})}{A_{ij}(a_{1i,j}, a_{2i,j}, a_{3i,j}, a_{4i,j})}
\end{aligned} \tag{22}
$$

Пусть изначально для каждого кандидата в чанки:

$$a_{1i,j} = True, \forall i, j \in 1..n; \tag{23}$$

$$a_{2i,j} = True, \forall i, j \in 1..n; \tag{24}$$

$$a_{3i,j} = True, \forall i, j \in 1..n; \tag{25}$$

$$a_{4i,j} = True, \forall i, j \in 1..n; \tag{26}$$

«A.1». Чанк неверный, если он вида «любое существительное (в главной позиции) + существительное в именительном падеже».

$$
\begin{aligned}
&((PS_i = "noun") \wedge \\
&\wedge (General_i = "true") \wedge \\
&\wedge (PS_j = "noun") \wedge \\
&\wedge (Case_j = "nominative")) \Rightarrow (a_{1i,j} = false)
\end{aligned} \tag{27}
$$

Эвристики «A.2» … «A.4» описаны на стр. 113 диссертации.

ж) Рассмотрим условия, описывающие эвристики, работающие после окончания поиска всех чанков в сегменте (эвристики основаны на лингвистических правилах согласования слов в предложении):

$$B_{ij}(b_{1i,j}, b_{2i,j}, b_{3i,j}, b_{4i,j}, b_{5i,j}, b_{6i,j}) = True, \text{ если}$$

$$
\begin{aligned}
&(b_{1i,j} = True) \wedge (b_{2i,j} = True) \wedge (b_{3i,j} = True) \wedge \\
&\wedge (b_{4i,j} = True) \wedge (b_{5i,j} = True) \wedge (b_{6i,j} = True)
\end{aligned} \tag{28}
$$

В терминах Модуса Поненса высказывание будет выглядеть следующим образом:

$$
\begin{aligned}
&Если\ \_(b_{1i,j}) \wedge (b_{2i,j}) \wedge (b_{3i,j}) \wedge \\
&\wedge (b_{4i,j}) \wedge (b_{5i,j}) \wedge (b_{6i,j}), \\
&mo\ \_B_{ij}(b_{1i,j}, b_{2i,j}, b_{3i,j}, b_{4i,j}, b_{5i,j}, b_{6i,j}) \\
&\frac{\begin{array}{c}(b_{1i,j}) \wedge (b_{2i,j}) \wedge (b_{3i,j}) \wedge \\ \wedge (b_{4i,j}) \wedge (b_{5i,j}) \wedge (b_{6i,j})\end{array}}{B_{ij}(b_{1i,j}, b_{2i,j}, b_{3i,j}, b_{4i,j}, b_{5i,j}, b_{6i,j})}
\end{aligned} \tag{29}
$$

Пусть изначально для каждого кандидата в чанки:

$$b_{1i,j} = True, \forall i, j \in 1..n; \tag{30}$$

$$b_{2i,j} = True, \forall i, j \in 1..n; \tag{31}$$

$$b_{3i,j} = True, \forall i, j \in 1..n. \tag{32}$$

$$b_{4i,j} = True, \forall i, j \in 1..n; \tag{33}$$

$$b_{5i,j} = True, \forall i, j \in 1..n; \tag{34}$$

$$b_{6i,j} = True, \forall i, j \in 1..n. \tag{35}$$

«B.1». Удаление из набора обнаруженных в сегменте чанков тех, которые являются полисемичными относительно входящих слов друг к другу за исключением первого из них.

$b_{1i,j} = True$, если

$$(w_i^0, C_i) \neq (w_j^0, C_j); \tag{36}$$

$b_{1i,j} = False$, если

$$(w_i^0, C_i) = (w_j^0, C_j). \tag{37}$$

Введем формальное описание явления полисемии (лексической многозначности).

Имеется набор концептов $C = \{c_i\}$.

Имеется набор лексических единиц $L = \{l_j\}$.

Имеется отношение лексикализации концептов R онтологической сети $R^{CL} = \{c_i, l_j\}$.

Явление лексической многозначности можно задать следующим условием:

$$
\begin{cases}
\exists c_i, c_k, l_j, l_e, что \\
p(c_i, l_j) \wedge p(c_k, l_e) \wedge (c_i \neq c_k) \wedge (l_j = l_e)
\end{cases}. \tag{38}
$$

Другими словами, существуют пары различных концептов, у которых совпадают лексические единицы, их означивающие в сегменте.

Описания эвристик «В.2» … «В.6» представлены на стр. 115-116 диссертации.

В итоге:

Если $Comp(U_{ij}, A_{ij}, B_{ij}, C_{ij})_{z \le z_0} = True$, то пара слов (w_i, w_j) является чанком, в противном случае пара слов (w_i, w_j) не является чанком.

Структурная схема процедуры частичного синтаксического анализа приведена в Приложении Г диссертационной работы. Выдержки из листинга программы, написанной на языке программирования Делфи приведены в Приложении Д диссертационной работы.

Расширенная нотация для задачи синтаксического анализа

Предложенная модель имеет ряд ограничений, не играющих существенной роли в задачах поиска. Эти ограничения связаны с сознательным упрощением алгоритма частичного синтаксического анализатора с целью получения частных результатов, необходимых для исследования в кратчайшие сроки. При этом для получения более общих результатов анализа любого предложения возможно использовать расширенную нотацию для работы синтаксического анализатора. В данном разделе будет показано, как можно путем модификации нотации правил и небольшой модернизации алгоритма чанкинга учитывать следующие явления в русском языке:

- сослагательное наклонение у глаголов, возникающее в русском языке при использовании указателя сослагательного наклонения – частицы «бы»;
- присутствие отрицания в предложении на основе проверки наличия отрицательных частиц «не/ни», которые могут быть перед существительным, прилагательным, причастием, глаголом, деепричастием и наречием;
- присутствие предлога как падежной характеристики существительного;
- наличие союзов «и/или» в предложении с однородными членами;
- наличие составных глаголов в предложении.

Для этой цели используются идея введения модификаторов, которые данные явления представляют как способы модификации базовой грамматической характеристики слова, описанной в формуле (6).

Применения такого подхода стало возможным вследствие того, что в русском языке часть грамматических характеристик передаются не только морфологическими, но и лексическими средствами (частицы, предлоги).

Например, при нахождении частицы «бы» в предложении наклонение глагола, стоящего перед или после этой частицы, меняется (модифицируется) на сослагательное и слово «бы» исключается из дальнейшего рассмотрения в качестве кандидата на главное или подчиненное слово, составляющего чанк.

При обнаружении частиц «не/ни», которые также являются одним из видов семантических модификаторов для слова, следующего за этой частицей, в картеж грамматических характеристик добавляется помета о том, что это слово употребляется в данном случае с отрицанием. А сама частица как слово исключается из дальнейшего рассмотрения.

Предлог в этом случае станет выступать как падежная характеристика существительных. То есть для каждого существительного к имеющимся морфологическим характеристикам будет добавлена еще одна, определяющая, стоит ли перед данным существительным предлог и какой он. Предлог как слово при дальнейшем анализе исключается из рассмотрения.

В случае с однородными членами будет использоваться составной элемент чанка, когда несколько чанков, в составе которых меняются только однородные члены, будут объединены в один чанк, где вместо слова $\mathbf{w_i}$ будет присутствовать группа объединенных элементов. Для выделения однородных слов используется еще одна дополнительная характеристика, в которой для каждого слова указывается номер группы однородных членов предложения. Каждое слово из каждой группы однородных членов по отдельности не рассматривается при дальнейшем анализе.

Составные глаголы в расширенной нотации также будут объединяться в составные элементы чанков. Для этого будет использоваться очередная дополнительная характеристика. Каждое слово из каждой группы составных глаголов по отдельности не рассматривается при дальнейшем анализе.

Введение расширенной нотации приводит к усложнению алгоритма анализа, но одновременно позволяет сделать набор эвристик более простым и универсальным.

<u>В результате проведенной работы были получены следующие научные результаты:</u>

1. Разработаны модель и алгоритм принятия решения для системы поддержки принятия решения в области патентного поиска при выборе патентов-аналогов. Модель основана на комбинировании методов оценки релевантности патентов по чанкам и по словам.

2. Усовершенствована модель частичного синтаксического анализа, основанная на блочном подходе. Важным результатом является тот факт, что данный подход дает возможность применять неограниченное количество «блоков» правил и фильтров. Каждый «блок» при этом выполняется последовательно, то есть информация, полученная в результате работы одного «блока», является входящей информацией для следующего «блока». Это предотвращает потери информации, наблюдаемые ранее при синтетическом подходе к построению дерева синтаксического подчинения. Кроме того, это позволяет сравнивать качество предложенных эвристик на промежуточных этапах, не дожидаясь построения полного дерева.

3. Предложены 3 группы эвристик, улучшающие качество синтаксического анализа. Каждая группа эвристик является отдельным «блоком».

- Первая группа эвристик имеет лингвистическую природу и выполняется на этапе поиска каждого чанка.
- Вторая группа эвристик является набором фильтров, накладываемых на все обнаруженные чанки в сегменте после окончания работы первой группы эвристик. Она также имеет лингвистическую природу.
- Третья группа эвристик также является набором фильтров и имеет математическую природу. Эти эвристики выполняются после эвристик из второй группы.

4. Описана математическая постановка задачи частичного синтаксического анализа в логико-математической нотации.

5. Предложена расширенная нотация математической постановки задачи частичного синтаксического анализа с использованием инструментария модификаторов грамматических

категорий, позволяющая описывать в единой манере «неудобные» с вычислительной точки зрения явления естественного языка, такие как: сослагательное наклонение, отрицание, предлоги, союзы и/или, составные глаголы.

ПРОГРАММНЫЙ КОМПЛЕКС ПОДДЕРЖКИ ПРИНЯТИЯ РЕШЕНИЯ «FINDING CHUNK»

В главе 3 приводится обзор функций программного комплекса «Find-Chunk», разработанного в рамках диссертационной работы для решения широкого круга задач, связанных с областью патентного поиска с использованием частичного синтаксического анализа.

Описываются методика и результаты экспериментального исследования модели частичного синтаксического анализа, метода разработки эвристик, позволяющих увеличить точность проводимого анализа.

Приводится описание методики и результатов тестирования модуля патентного поиска.

Программный комплекс «Find-Chunk» предназначен для проведения частичного синтаксического анализа текстов на русском языке и поддержки принятия решения в области патентной поиска.

После загрузки текста результат его анализа выводится в удобном для пользователя виде (рисунок 2).

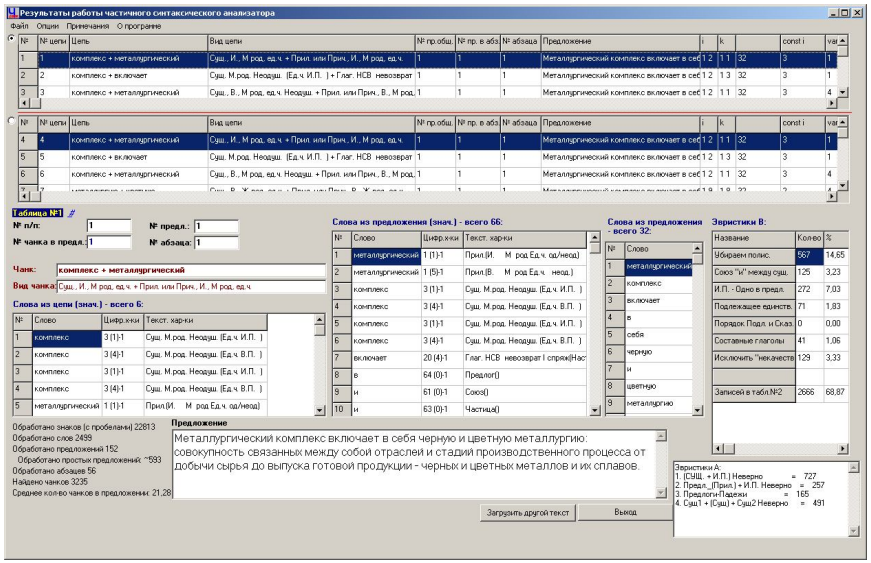

Рисунок 2 – Окно встроенного синтаксического анализатора

Также в программный комплекс встроены несколько дополнительных функций:

- Модуль для проведения статистических исследований в исследуемом тексте. Позволяет исследовать наиболее встречаемые наборы омонимичных чанков, принимая во внимание все предложения (сегменты) из анализируемого текста, а также дает возможность определять распределение чанков в тексте в зависимости от расстояния между словами в чанках, количества слов в сегменте и других характеристик.
- Модуль для проведения морфологического анализа отдельных слов из текста. Позволяет проводить морфологический разбор каждого слова из предложения.
- Модуль принятия решения, позволяющий проводить патентный поиск на русском языке. Данный модуль позволяет проводить поиск патентов на основе анализа патентных формул каждого из патентов.

Процедура патентного поиска содержит следующие шаги:

1. Пользователь задает патентную формулу для поиска патентов-аналогов.
2. После начала поиска производится анализ введенной пользователем патентной формулы для выявления в ней всех возможных чанков и слов для поиска.
3. Пользователь выбирает (отмечает) из всех возможных чанков и слов те, которые, по его мнению, наилучшим образом описывают введенную им патентную формулу.
4. Далее производится обнаружение чанков и слов в каждом патенте, среди которых производится поиск патентов аналогов.
5. После этого согласно формуле (3) вычисляется индекс релевантности для каждого патента с использованием обычного метода совпадений, а также меры TD-IDF. Далее производится сортировка патентов. Патенты, имеющие индекс релевантности менее 0,05, считаются нерелевантными поисковому запросу пользователя.
6. Патенты, имеющие индекс релевантности более или равный 0,05, в упорядоченном виде представляются пользователю для ознакомления.
7. После просмотра найденных патентов-аналогов пользователь принимает решение о необходимости уточнения условий поиска (изменение заданной для поиска патентной формулы – возврат к шагу №1 или изменение набора отобранных чанков – возврат к шагу №3) или об удовлетворении своего запроса и прекращении дальнейшего уточнения условий поиска.

Тестирование модуля частичного синтаксического анализа и модуля для проведения статистических исследований; методика формирования эвристик

Исходными данными для тестирования программного комплекса явился научно-популярный текст про металлургический комплекс России. В тексте присутствует 2499 слов, которые состоят из 20128 знаков без пробелов или 22813 с пробелами. Текст состоит из 65 абзацев и 207 предложений.

При проведении части тестирования, в которой предполагалось сравнивать результаты работы программного комплекса с мнением эксперта, использовалась часть этого текста (1170 слов, которые состоят из 7412 знаков без пробелов или 8701 с пробелами; этот текст состоит из 44 абзацев и 105 предложений).

Тестирование модуля частичного синтаксического анализа программного комплекса производится методом сравнения результатов работы алгоритма (количество обнаруженных чанков) с истинным количеством чанков в обработанном тексте, выделенным экспертом. Оценка качества производится с использованием меры F_1.

На рисунке 3 представлена сводная диаграмма значений P_r, R_e, F_1 для случаев с применением всех разработанных эвристик и для случая применения только одной эвристики, ограничивающей область анализа (значения обозначены как $\_P_r$, $\_R_e$, $\_F_1$).

Была проведена аналитическая работа по выявлению вклада каждой из эвристик в окончательный результат работы частичного синтаксического анализатора.

Распределение вклада по группам эвристик следующее:

- Влияние эвристик группы «A» составило 63,85%;
- Влияние эвристик группы «B» составило 36,15%.

Качество работы частичного синтаксического анализатора удовлетворяет требованиям многих прикладных задач, требующих проведение анализа большого количества текста при ограничении временных ресурсов, в том числе патентного поиска.

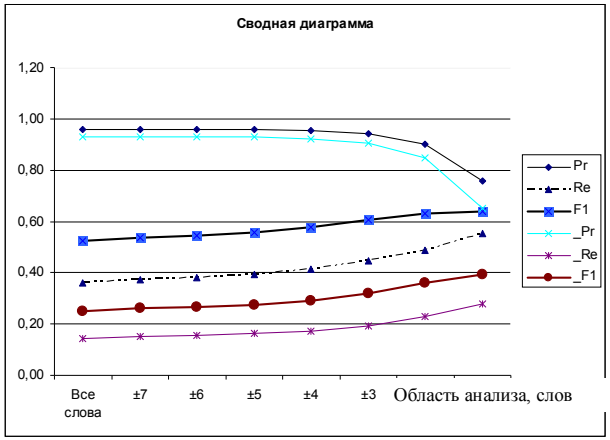

Рисунок 3 – Сводная диаграмма: P_r, R_e, F_1 - величины, полученные при применении всех эвристик; $\_P_r$, $\_R_e$, $\_F_1$ - величины, полученные при применении только одной эвристики, ограничивающей область анализа

Тестирование модуля поддержки принятия решения в области патентного поиска

В качестве исходных данных для проведения тестирования модуля патентного поиска было использовано 320 текстов – патентных формул[1] из различных областей науки – каждый из которых содержит от 100 до 800 слов.

Для тестирования модуля патентного поиска в поле запроса вводится некий текст – патентная формула или часть патентной формулы некоторого патента, отсутствующего среди тех, по которым производится патентный поиск. Это позволяет избежать витальных запросов[2]. В каждом из 25 экспериментов в поиске участвовали 100 патентов, случайно выбранных из 320 таким образом, чтобы среди них было не менее 5 патентов, «похожих» на запрос (так называемых, «патентов-аналогов»).

[1] Для краткости в дальнейшем по тексту будет использоваться слово патент.
[2] Витальный запрос – такой запрос, на который возможно получить единственный верный ответ.

При выборе формулы для расчета релевантности патентов в модуле патентного поиска производился сравнительный анализ шести методов поиска.

1. Расчет простого частотного индекса по словам.
2. Расчет простого частотного индекса по чанкам.
3. Расчет комбинированных частотных индексов по словам и чанкам.
4. Расчет частотного индекса по методике TF-IDF по словам.
5. Расчет частотного индекса по методике TF-IDF по чанкам.
6. Расчет комбинированных частотных индексов по методике TF-IDF по словам и чанкам.

Результаты сравнительного анализа представлены на рисунке 4. На их основании можно сделать вывод, что наилучшее качество поиска достигаются при использовании комбинированного метода анализа на основе применения частотных индексов по методике TF-IDF, полученных при поиске по словам и чанкам.

<u>По результатам третьей главы сделаны следующие выводы:</u>

1. Итогом проведенной работы стало создание программного комплекса «Find-chunk», в состав которого входят следующие модули:

- Модуль для проведения статистических исследований в исследуемом тексте. Позволяет исследовать и группировать наиболее встречаемые наборы омонимичных чанков, принимая во внимание все предложения из анализируемого текста, а также дает возможность определять распределение чанков в тексте в зависимости от расстояния между словами в чанках, количества слов в сегменте и других характеристик.
- Модуль для проведения морфологического анализа отдельных слов из текста. Позволяет проводить морфологический разбор каждого слова из предложения.
- Модуль принятия решения, позволяющий проводить патентный поиск на русском языке.

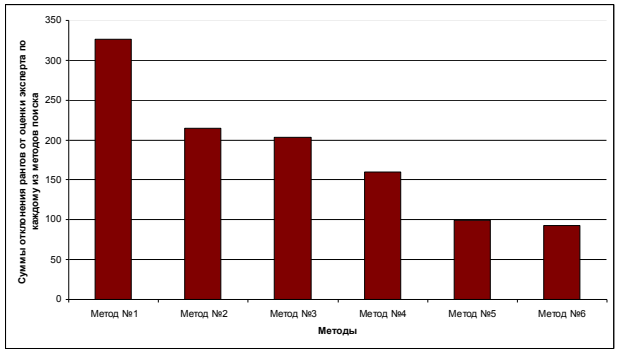

Рисунок 4 - Суммы отклонения рангов от оценки эксперта по 6 методам поиска

2. Программный комплекс «Find-chunk» был апробирован в Казанском (Приволжском) федеральном университете.

3. Для тестирования алгоритма парсинга были разработаны специальные приемы, которые позволили оценить точность его работы.

4. В результате тестирования программного комплекса свою состоятельность доказал блочный подход, применяющийся при синтаксическом анализе для увеличения точности его работы. При этом выяснилось, что этот подход также позволяет проводить настройку параметров анализа на этапе выполнения алгоритма, что оказывается очень удобным при анализе текстов разного рода

5. Мера F_1 при работе частичного синтаксического анализатора с использованием эвристик выросла с 0,25 до 0,6, при этом P_r вырос с 0,9 до 0,91, а R_e с 0,14 до 0,45.

6. Точность работы частичного синтаксического анализатора соответствует требуемой точности для работы прикладных задач, например, задачи патентного поиска.

7. При разработке модуля поддержки принятия решения в области патентного поиска было проведено сравнение его работы с работой обычного поискового алгоритма, основанного на поиске по ключевым словам. Сравнение показало, что использование алгоритма, основанного на гибридном поиске с использованием меры TD-IDF по чанкам и словам, имеет преимущество перед обычным поиском по словам.

ЗАКЛЮЧЕНИЕ

В ходе данной работы была предложена модель принятия решения в области патентного поиска, позволяющая с помощью все более глубокого уточнения условий поиска пользователем на каждой итерации получать максимально релевантный список патентов-аналогов.

Также было проведено исследование работы частичного синтаксического анализатора для русского языка, исследованы статистические параметры для чанков в русском языке.

Анализ статистических параметров дал возможность выявить большое количество закономерностей, описание которых позволило сформулировать набор эвристик, позволивших значительно увеличить точность работы частичного синтаксического анализатора. Так результирующая мера F_1, оценивающая качество работы частичного синтаксического анализатора с использованием эвристик выросла с 0,25 до 0,6, при этом показатель точности классификации P_r вырос с 0,93 до 0,94, а показатель отказа классификации R_e вырос с 0,14 до 0,45.

В ходе исследования была предложена новая методика для проведения синтаксического анализа, опирающаяся на «блочный» подход. Согласно данной методике возможно отдельное функционирование, разработка и настройка каждого «блока» модели. Информация с результатами работы одного «блока» поступает на вход следующего «блока». Каждый «блок» состоит из набора эвристик или правил. Плюсом данного подхода является то, что каждый «блок» является независимой частью алгоритма, а уточнение параметров функционирования «блока» может производиться пользователем на этапе выполнения алгоритма. Таким образом, увеличилась точность работы синтаксического анализатора.

Созданные в ходе данной работы модели и программный комплекс «Find-chunk» дали возможность убедиться в эффективности частичного синтаксического анализа в виде отдельной задачи, а также как прикладной задачи в составе, например, поисковой системы, разработанной с целью нахождения патентов-аналогов.

Проделанная работа привела к следующим результатам и выводам:

1. Выполнена формальная постановка задачи принятия решения итерационного поиска патентов-аналогов на основе анализа чанков.

2. Исследованы и описаны закономерности согласования слов в русском языке, которые позволили сформулировать часть эвристик, вошедшие в алгоритм частичного синтаксического анализа.

3. Исследованы и описаны наборы омонимичных чанков, наиболее встречающихся в научных текстах на русском языке, которые дали возможность сформулировать часть эвристик, существенно повысивших точность работы частичного синтаксического анализатора.

4. Разработан алгоритм частичного синтаксического анализа с использованием условий поиска чанков, а также эвристик, позволяющих значительно повысить точность работы частичного синтаксического анализатора.

5. Предложена методика и общий алгоритм для проведения частичного синтаксического анализа, основывающаяся на блочном подходе, применение которого может привести к максимальной точности работы частичного синтаксического анализатора.

6. Создан алгоритм для системы поддержки принятия решения в области патентного поиска.

7. На основе моделей и алгоритмов, предложенных в данной работе, создан и апробирован опытный программный комплекс системы поддержки принятия решения «Find-chunk».

Теоретические вопросы диссертации освещаются в 10-и научных публикациях, в том числе две публикации в изданиях, рекомендованных ВАК:

1. Буштедт В. А., Поляков В. Н. Частичный синтаксический анализатор для корпоративной поисковой системы. // Труды Казанской школы по компьютерной и когнитивной лингвистике (TEL-2006), Казань, Отечество, 2007, с. 4-16.

2. Vladislav Bushtedt, Vladimir Polyakov. Finding chunks with restriction of distance to dependent word. Text Processing and Cognitive Technologies. Paper Collection. N 13. (Edited by V. Solovyev, R. Potapova, V. Polyakov). Kazan: KSU, 2007, p. 37-46.

3. Vladislav Bushtedt, Vladimir Polyakov. Partial parsing with use of heuristics directed on the search of false chunks. Text Processing and Cognitive Technologies. Paper Collection. N 15. (Edited by V. Solovyev, M. Bergelson, V. Polyakov). Kazan: KSU, 2008, p. 204-228.

4. Буштедт В. А. Частичный синтаксический анализатор с применением эвристик, повышающих точность его работы. // 64-е дни науки студентов МИСиС: международные, межвузовские и институтские научные конференции. М.: МИСиС, 2009, с. 365-367.

5. Буштедт В. А., Поляков В. Н. Использование частичного синтаксического анализа текстов для патентного поиска в области нанотехнологии. Труды российско-японско-казахстанской научной конференции «Перспективные технологи, оборудование и аналитические системы для материаловедения и наноматериалов», Волгоград, 2009, с. 1026-1034.

6. Буштедт В. А., Поляков В. Н. Эвристики для улучшения работы частичного синтаксического анализатора. Ученые записки Казанского Государственного Университета, 2009, т. 151, книга 3, с. 214-228.

7. Буштедт В. А., Поляков В. Н. Блочный алгоритм для синтаксического анализатора // TEL'09. – Казань: Фэн. 2010. с. 46-64.

8. Буштедт В. А. Модель синтаксического анализа в задачах обработки патентной информации // 65-е дни науки студентов МИСиС: международные, межвузовские и институтские научные конференции. М.: НИТУ «МИСиС», 2010, с. 529-530.

9. Буштедт В. А. Тестирование модуля патентного поиска с использованием модели синтаксического анализа в задачах обработки патентной информации // 66-е дни науки студентов МИСиС: международные, межвузовские и институтские научные конференции. М.: НИТУ «МИСиС», 2011, с. 416.

10. Буштедт В. А., Поляков В. Н. Блочный алгоритм для синтаксического анализатора с использованием расширенной нотации // Естественные и технические науки № 2. М.: «Спутник+», 2011. с. 410-413.

Соискатель В.А. Буштедт