Содержание

ВВЕДЕНИЕ

Актуальность работы

Современное состояние дел в естественных науках, в частности, в области химии и химического катализа, характеризуется высокой скоростью накопления данных, б**о**льшая часть из которых представлена в текстовой форме (статьи, патенты, тезисы конференций и т.д.). Размеры существующих в настоящее время текстовых коллекций (библиографических полнотекстовых БД) растут огромными темпами и на первый план выходят автоматизированные средства для работы с ними (такие, например, как Scopus®, SciFinder®, Reaxys®, WebOfScience® и т.д.). Это не удивительно, так как в современном мире скорость и качество проведения исследовательских работ, а также эффективность процесса создания новых технологических решений на базе существующих знаний, напрямую зависят от качества их информационного обеспечения, ориентированного, в основном, на расширенный поиск, извлечение и оперативный анализ нужной информации. Действительно, наличие массивов текстовых данных само по себе не представляет особой ценности без системы их интеллектуального автоматизированного анализа, наличие которого выходит на первый план при работе с очень большими и постоянно обновляющимися коллекциями текстов. При этом задачи, связанные с систематизацией (таксономией) текстовых коллекций представляет собой отдельную важную проблему.

Как правило, основой всех существующих методов анализа текстовых коллекций является предварительное индексирование имеющихся текстов в соответствии с наборами терминов (ключевых слов и фраз), известных как «терминологический индекс» и характеризующих содержание текста. Методы применения индекса для осуществления дальнейшего поиска и извлечения информации из текстовых коллекций основаны на использовании булевых комбинаций терминов индекса, векторных и вероятностных методов для решения задач таксономии текстов, использовании весов терминов и ряда других комбинированных подходов [1].

Можно утверждать, что структурной основой любой научной или технической текстовой информации является терминологическая база (ТБ) данной предметной области (ПО). При этом необходимо отметить, что язык науки меняется гораздо быстрее, чем происходит эволюция языка в целом. Особенно это заметно в таких областях, как химический катализ и молекулярная биология. Таким образом, процедура анализа терминологической базы текстовой коллекции должна осуществляться в

автоматизированном режиме, с использованием, в том числе, методов лингвистического анализа. Только автоматизированный системный подход может обеспечить работу с постоянно обновляющимся огромным массивом документов и имеющимся в нём массивом терминоподобных словосочетаний.

В принципе, терминологическая база конкретной ПО может быть представлена в виде информационно-поисковых тезаурусов (ИПТ), которые обычно создаются и регламентируются специалистами соответствующего профиля. Однако этот процесс является чрезвычайно трудоёмким, поэтому создаются и поддерживаются только ИПТ «верхнего» или «базового» уровня. Например, выработкой стандартов в использовании научной терминологии в химии занимается организация IUPAC (International Unit Pure and Applied Chemistry). Разработанные IUPAC списки терминов доступны в виде тезауруса на ресурсе «IUPAC Gold Book», представляющем список только основных «канонических» терминологических понятий и не претендующий на широкий охват всех предметных областей в области химии.

В связи с этим возникает потребность в создании программных систем, позволяющих, наряду с использованием известных терминологических понятий верхнего уровня (готовых ИПТ и т.д.), анализировать и извлекать из полнотекстовых источников максимально полную информацию об используемой терминологии (терминологический спектр публикации), соотнесённую с конкретными подразделами данной ПО. Такие системы могут быть использованы в аналитических и справочных целях, в качестве основы для специализированных поисковых систем, а также при выработке стандартов употребления научной терминологии. Таким образом, можно говорить о терминологическом спектре публикации как об универсальном научном представлении текста, состоящем из индексированного списка используемых в нём общенаучных понятий и терминов из известных тезаурусов, распознанных хим. соединений и аббревиатур, терминоподобных словосочетаний, с получением написания терминов на естественном языке, с естественным окружением и употреблением. С использованием таких специфичных терминологических словарей можно организовать как новые эффективные процедуры поиска, так и новые методы анализа текстовой информации.

Цель и основные результаты работы

Основной целью работы является разработка системы автоматизированного анализа научных текстов по химии и катализу, направленной на автоматизированное выявление терминоподобных словосочетаний и построение терминологического спектра коллекций текстовых документов, с последующим использованием результатов анализа в

системах принятия решений, специализированных поисковых сервисах, патентном анализе и т.д. Предлагаемый подход включает в себя разработку теоретических основ и практических алгоритмов для автоматизированного извлечения указанной информации путём детального анализа L-граммного спектра публикаций.

В качестве конкретной ПО на текущий момент выбрана химия и химический катализ, однако разработанные методики могут быть использованы и по отношению к другим ПО. Корпус текстов, рассматриваемый в работе, представляет собой материалы периодической, широкой по охвату научной конференции «Европейский Конгресс по Катализу – EuropaCat» за 2013, 2011, 2009, 2007, 2005 годы (всего порядка 6 тыс. документов). Упор на английском языке сделан, в первую очередь, ввиду публикации на нём подавляющего большинства научной периодической литературы.

Возможные применения системы терминологического анализа заключаются в частичной автоматизации решения следующих актуальных задач:

а) Построение и пополнение быстроменяющегося тезауруса ПО с частичным установлением связей между терминами.

б) Анализ и отслеживание динамики изменения ТБ ПО путём анализа научной периодической литературы, представленной на естественном языке и отражающей текущее состояние ПО (по материалам тематических конференций).

в) Поиск «горячих» направлений или тенденций развития ПО (перспективных материалов, катализаторов, методик, процессов и т.д.) путем анализа изменений ТБ во времени.

г) Организация направленного поиска по отдельным аспектам содержания научных текстов, осуществляющаяся с использованием индикаторных словарей (или словарей "подсказок").

д) Отслеживание смысловых связей между текстами путем сопоставления их терминологических спектров; таксономия текстов.

е) Отслеживание смысловых связей между документами с целью целевой подборки документов и формирования «ближайшей окрестности» заданного документа.

ж) Разработка технологий анализа научных текстов с целью выявления закономерностей и указанных экспертом элементарных смыслов. В том числе разработка технологий поиска и автоматизированной реконструкции ассоциативных сетей.

В рамках поставленной цели были достигнуты следующие основные результаты:

а) На основе L-граммного подхода разработана методика анализа и извлечения терминологического спектра из текстовых коллекций в области химии и химического катализа. Особенностью модели является применение расширенных правил фильтрации и отбора терминоподобных словосочетаний, использование морфологической и структурной информации.

б) На основе разработанного метода создана система терминологического анализа, выполняющая автоматизированный терминологический анализ текстовых коллекций, представленных в pdf-формате.

в) Разработан инструментарий отчетов, позволяющих пользователю-эксперту в предметной области анализировать полученные результаты, находить тренды и закономерности в динамике поведения терминов.

г) Разработан пользовательский веб-интерфейс, содержащий компоненты для просмотра терминологической разметки, исходных и преобразованных текстов, работы с модулем отчетов.

Структура и объем работы

Работа состоит из введения, литературного обзора, восьми основных глав, заключения и списка использованной литературы.

В *первой главе* – литературном обзоре – описываются существующие системы научного поиска, средства анализа химических текстов, инструменты для автоматического извлечения ключевых фраз.

Вторая глава содержит общее описание разработанной системы терминологического анализа.

Третья глава описывает механизм и реализацию построения логической модели pdf-документа.

Четвертая глава содержит общую схему модели терминологического анализа, включая структуру используемых правил и шаблонов.

Пятая глава детализирует общую модель терминологического анализа для случая химического катализа, в главе приводится список экземпляров правил и шаблонов с указанием примеров отбираемых L-грамм.

Шестая глава посвящена программной реализации системы. Приводятся диаграммы классов основных модулей системы, описывается структура компонентов.

Седьмая глава демонстрирует работу пользовательского интерфейса. Приводятся примеры отчетов, описываются основные характеристики L-грамм, вычисляемые в процессе построения отчетов.

В *восьмой главе* приводится оценка эффективности работы системы. Также в данной главе приведен анализ некоторых L-грамм, выделенных экспертом с помощью системы. Указаны основные тенденции развития данных L-грамм.

1 Литературный обзор

В данной главе представлены: обзор основных средств, используемых для научного текстового поиска; описание ряда стандартов для представления и хранения научных текстов; обзор средств анализа химических текстов, а также инструментов для решения задачи автоматического извлечения ключевых фраз.

1.1 Специализированные системы текстового поиска

Специализированные научные поисковые системы выполняют поиск по различным научным источникам, а также предоставляют продвинутый пользовательский интерфейс для создания сложных поисковых запросов и последующей обработки результатов поиска.

1.1.1 SciFinder

SciFinder [2] – информационный центр, обеспечивающий доступ к реферативным и справочным базам данных компании Chemical Abstracts Service. Доступна работа с тремя информационными ресурсами:

а) References – реферативные базы данных CAplus и MedLine (более 30 млн. статей и патентов).
б) Substances – структура, свойства, спектры более 50 млн. химических веществ.
в) Reactions – около 20 млн. одно– и многостадийных химических реакций.

Организована связь между этими тремя информационными массивами, например, от реферата статьи есть переход к описаниям веществ, упомянутых в статье, а от описания вещества — к рефератам тех статей, где говорится об этом веществе. Поисковый запрос может содержать от одного до трех понятий. Для отбора можно использовать фильтры, такие, как отсев документов по автору, названию, году публикации, типу документа и т.д. Результаты поиска представляются в четырех группах, приведенных в таблице 1.

Таблица 1. Группировка результатов поиска в поисковой системе SciFinder.

Значение	**Означает, что в документе поисковые термины были обнаружены...**
As entered	... точно в такой же форме, как в запросе.
Closely associated with one another	... в одном и том же предложении или в названии.
Present anywhere within a reference	... где-то в названии, реферате или указателе (вполне возможно, что на большом удалении друг от друга).
Containing the concept	... где-то в записи (либо сами термины, либо их синонимы или подобные термины).

После выполнения поиска пользователю предоставляется три инструмента для работы с его результатами: анализ (analyze), фильтрация (refine) и категоризация

(categorize). Инструмент «анализ» позволяет исследовать различные параметры множества найденных документов, такие, как распределения публикаций по базам данных, авторам, типам документов, языкам и т.д. Также он позволяет проанализировать наиболее часто встречающуюся терминологию (index terms).

Инструмент «категоризация» позволяет проводить фильтрацию результатов поиска в соответствии с желаемой категорией и используемой терминологией. Он работает следующим образом: первоначально результаты поиска группируются по категориям и подкатегориям, затем выделяется терминология, используемая в результатах поиска. При работе с данным инструментом пользователь сначала выбирает желаемую категорию и подкатегорию, а затем отмечает интересующие его термины. После проведения фильтрации в результатах поиска остаются записи, содержащие указанные термины.

Еще одним часто используемым инструментом является фильтрация (refine), позволяющий производить сужение результатов поиска до требуемых в соответствии с заданными критериями. Такими критериями могут быть автор, название компании, тип документа, год публикации, язык, база данных.

1.1.2 Scopus

Scopus [3] – библиографическая и реферативная база данных, а также инструмент для отслеживания цитируемости статей, опубликованных в научных изданиях. Индексирует 18 000 названий научных изданий по техническим, медицинским и гуманитарным наукам 5000 издателей. Имеется несколько видов поиска: простой поиск, поиск по автору, поиск по организации, расширенный поиск. В поисковом запросе можно использовать до двух поисковых терминов, соединяя их операторами AND, OR, NOT. Можно ввести ограничение по году издания или по времени добавления записи в базу Scopus. Также есть возможность ограничить поиск предметными областями.

В результатах поиска можно просмотреть общий обзор цитат, загрузить сразу несколько статей, посмотреть список ссылок или цитирующих статей. С каждой работой можно ознакомиться – просмотреть аннотацию и ссылки (Abstract + Ref), прочитать только аннотацию (Show Abstract), выйти на полный текст статьи на сайте издательства (View at Publisher). Статьи можно сортировать по любому столбцу. По умолчанию результаты поиска выведены по году издания.

Поиск по автору позволяет ознакомиться с индексом цитирования и списком работ определенного автора. Переходя в профиль автора, можно увидеть количество его статей в базе данных, количество источников, на которые ссылается автор, количество

цитирования, индекс Хирша, список соавторов, предметные области, историю публикаций.

1.1.3 Web Of Science

Поисковая система «Web Of science» [4] осуществляет поиск в более чем 12 000 журналов и 148 000 материалов конференций в области естественных, общественных, гуманитарных наук и искусства. Имеется возможность проводить поиск статей, оценивать их значимость и просматривать список работ, их цитирующих.

Функциональные возможности данной системы позволяют выполнять один из трех видов поиска: Search, Cited Reference Search или Advanced Search, также после проведения поиска возможно просмотреть историю поиска или перейти и списку отмеченных результатов (Marked list). В режиме поиска можно ввести до трех поисковых терминов, соединяя их операторами AND, OR, NOT, в случае необходимости можно ввести дополнительное поисковое поле. Поисковые термины можно вводить как самостоятельно, так и при помощи специального инструмента, позволяющего выбрать термин из списка значений. Можно ввести ограничение по времени добавления записи в базу данных Web of Science или по году издания статьи. Также возможно ограничить поиск одной из трех систем цитирования.

Страница результатов поиска делится на две части: первая часть предназначена для фильтрации выдачи, во второй части отображаются результаты поиска. Фильтрацию результатов можно осуществлять по таким критериям, как: предметная область, тип документа, имя автора, язык, страна и т.д. Результаты поиска можно распечатать, отправить по электронной почте или в список отмеченных результатов. Имеется два инструмента для работы с результатом поиска – Analyze Results и Create Citation Report. Analyze Results предназначен для упорядочения и группировки выдачи по различным параметрам, например, тип документа, номер гранта и др. Create Citation Report позволяет создавать отчет, в который включается диаграмма распределения статей и цитат по годам, среднее количество цитирований, а также индекс Хирша.

С помощью Web of Science также возможно проводить поиск работ, цитирующих выбранного автора, из определенного издания и года. Результат поиска отображается в виде таблицы, в первом столбце которой отображается имя автора, во второй – название журнала и т.д. В столбце Citing Articles отображается количество статей, цитирующих искомую работу.

1.2 Форматы представления текстовых данных

В настоящее время все большую актуальность приобретают задачи, связанных с автоматизированной обработкой текстов, в частности, обработкой полных текстов научных статей. Для эффективной работы необходимо, чтобы такие тексты были представлены в форме, удобной для машинной обработки – в них присутствовала метаинформация и вся необходимая разметка. Далее представлено два стандарта разметки текстовых данных, удовлетворяющие указанным требованиям.

1.2.1 TEI

TEI (Text Encoding Initiative) [5] – консорциум, разрабатывающий и поддерживающий стандарты хранения и отображения текстов в цифровой форме. Основу TEI составляют документы, представленные в виде xml-разметки. Существует базовая разметка, поверх которой могут существовать произвольное количество схем разметок для разных предметных областей, например, EpiDoc – расширение базовой разметки для эпиграфики.

Базовая разметка обладает достаточно разнообразными средствами, включающими:

а) базовые структурные и функциональные компоненты;
б) дипломатическую транскрипцию, изображения, аннотации;
в) ссылки, соответствия, выравнивание;
г) объекты, содержащие особые данные: дата, время, место, лицо, событие и т.д.;
д) метатекстовую аннотацию (исправления, удаления и т.п.);
е) все уровни лингвистического анализа;
ж) контекстные метаданные всех видов.

Применительно к обработке научных статей, документ в формате tei может содержать такие компоненты, как

а) название статьи;
б) авторы (в виде списка с указанием автора для переписки);
в) организации авторов;
г) соответствие между авторами и организациями (аффилиация);
д) заголовки различных уровней вложенности;
е) рисунки, таблицы, подписи;
ж) библиография.

Данный формат используется в некоторых средствах для извлечения логической структуры из pdf-документов, а также системах извлечения ключевых фраз.

1.2.2 JATS: Journal Article Tag Suite

JATS (Journal Article Tag Suite) [6] – xml-схема для представления журнальных статей. Задача JATS состоит в том, чтобы представить статью в виде, пригодной для обработки всеми заинтересованными сторонами – издателями, читателями, машинной обработкой и т.д. Набор тегов охватывает все области, связанные с созданием, изданием и публикации статьи. Структура JATS-документа предусматривает четыре части:

а) Вводная секция (front matter) – обязательная. Содержит метаданные о статье – название, журнал, в котором опубликована статья, дата, том, номер публикации, информация о правообладателе.

б) Основная часть (body) – опциональная. Содержит основное текстовое и графическое содержание статьи. Обычно состоит из параграфов и разделов, которые могут содержать рисунки, таблицы.

в) Информационная часть (back matter) – опциональная. Содержит глоссарий, приложения, библиографический список.

г) Документная часть (floating material) – опциональная. В некоторых случаях «плавающие» блоки, такие, как рисунки, таблицы, примечания могут находится в данной секции вне основного содержания статьи, тогда в тексте основной статьи будут ссылки в данную часть документа.

JATS используется в нескольких крупных репозиториях научных статей, например, PubMed Central.

1.3 Извлечение логической структуры из pdf-документов

В настоящее время множество разнообразных документов, в частности, научных статей и тезисов конференций, представлено в виде pdf-файлов. Сложность анализа таких файлов состоит в том, что pdf-документ не содержит в себе описания логической структуры документа. Под логической структурой документа понимается набор логических блоков, составляющих документ – название, список авторов, заголовки различных уровней, библиография и т.д. В основе формата pdf лежит набор текстовых блоков с указанием шрифта, абсолютного положения в документе, угла поворота. Задача по восстановлению логической структуры обычно выполняется в два этапа – на первом этапе производится извлечение отдельных слов и строк документа, а затем выполняется их классификация в соответствие с логической моделью документа. Далее представлено несколько средств анализа pdf-документов научных статей.

1.3.1 PDFx

PDFX [7] – средство для извлечения логической модели научной статьи из pdf-файла. Логическая модель документа, с которой работает PDFx, представлена в таблице 2.

Таблица 2. Логическая модель документа в PDFx.

Элементы начала документа	**Элементы тела документа**	**Элементы конца документа**
Название авторы описание (abstract) аффилиация	основной текст секция, подсекция заголовок секции, подсекции рисунок таблица подпись ссылка на рисунок/таблицу библиографическая ссылка	Элемент библиографии (библиографическая ссылка) URL email header/footer номер страницы

Процесс анализа pdf-файла состоит из двух этапов. На первом этапе выполняется построение геометрической модели документа. На следующем шаге геометрическая модель используется для выделения логических блоков, перечисленных выше.

Геометрическая модель строится с помощью Utopia Documents PDF Reader [8]. Определяются три основных элемента - страницы, слова и графическая информация (рисунки, линии). Каждый элемент является отдельным объектом с различными атрибутами – ограничивающий прямоугольник (bounding box), положением в документе, информацией о форматировании. Смежные слова с одинаковыми параметрами форматирования объединяются в один блок для последующего логического анализа. Также на этапе построения геометрической модели выполняется сбор статистики использования шрифтов на уровне страницы и документа в целом. Данная информация затем будет использоваться в логическом анализе документа.

Логический анализ состоит в просмотре полученной геометрической модели и классификации блоков логическими классами (автор, название и т.д.). Классификация выполняется последовательно различными классификаторами, при этом блоки просматриваются «в порядке чтения», т.е. так, как они бы просматривались бы читателем статьи. Первый шаг в логическом анализе – выделение основного текста статьи. Основной текст определяется как текст, имеющий самый распространенный шрифт в документе. Тегирование остальных блоков выполняется в соответствии со сложностью определения блоков в следующей последовательности:

1. рисунки;
2. DOI (Digital Object Identitifer) – уникальный идентификатор статьи;

3. список авторов публикации;
4. название публикации;
5. номера страниц;
6. заголовки верхнего уровня;
7. описание (abstract);
8. подписи к рисункам и таблицам;
9. заголовки более низких уровней;
10. примечания к автору (author foonotes);
11. оставшиеся регионы;
12. библиография и элементы библиографии;
13. остальные неклассифицированные регионы;
14. таблицы;
15. ссылки на таблицы, рисунки и библиографические ссылки;

Оценивание результатов выполнялось с использованием трех множеств текстов. Первое множество содержало 39 статей из области Computer Science. Второе множество содержало 1943 статьи из различных журналов подборки PMC Open Access Subset. Третье множество было построено из всех публикаций издательства Elsevier за 2008 год, предоставляемых для исследований. Из всего множества документов было отобрано около 50 000 статей. В таблице 3 представлены результаты для второго и третьего множества. Для оценки совпадения блоков использовался метод сравнения строк Ratcliff/Obershelp, представленный в [9], с 95% сходством между строками.

Таблица 3. Результаты работы PDFx.

Dataset	**Size**	**h3**	**table**	**h2**	**abstract**	**caption**	**author**	**h1**	**title**
Elsevier	50000	83.35	28.78	82.03	62.01	82.86	94.63	90.5	96.7
PMC sample	1943	6.05	13.27	27.19	32.41	54.53	61.65	77.45	85.42

PDFX является онлайн-системой и доступен для использования через веб-интерфейс по адресу http://pdfx.cs.man.ac.uk/.

1.3.2 SectLabel

SectLabel [10] – средство для анализа логической структуры pdf-документов. Данная библиотека относится к задаче выделения структуры документа как к задаче классификации: документ рассматривается как набор строк $L = \{l_1, l_2, \dots, l_n\}$. Каждой строке l_i необходимо присвоить класс (метку) ил множества классов $C = \{c_1, c_2, \dots, c_m\}$. Для классификации используется метод CRF (condition random fields).

Процесс анализа состоит из двух фаз – обучения и непосредственно анализа. В процессе обучения на вход классификатору подаются размеченные текстовые данные, и строится модель классификации. В процессе анализа строки текста классифицируются согласно построенной модели.

Логическая модель документа в SectLabel состоит из 23 классов: address, affiliation, author, bodyText, categories, construct, copyright, email, equation, figure, figureCaption, footnote, keywords, listItem, note, page, reference, sectionHeader, subsectionHeader, subsubsectionHeader, table, tableCaption, title. Класс note обозначает дополнительный текст сверху или снизу страницы, обычно содержащий детали проведения конференции.

В качестве атрибутов строк при классификации используются два типа атрибутов – относящихся непосредственно к строке (содержит ли строка числа, знаки пунктуации и т.д.), и атрибуты, относящиеся к документу (параметры форматирования строки, ее положение в документе). Такими атрибутами являются:

а) положение в документе (location) – содержит относительное положение строки в документе;
б) присутствие цифр (number) – отражает вхождение различных шаблонов, специфичных для иерархии заголовков, например, subsectionHeaders («1.1»), subsubSectionHeaders («1.1.1»);
в) признаки пунктуации (punctuation) – набор признаков для определения email адресов или веб ссылок;
г) признак длины (length) – содержит длину строки в токенах. Значения данного признака ограничены как 1 токен, 2 токена, 3 токена, 4 токена, 5 или более токенов. Признак полезен для определения строк, составляющих основное содержание документа.

Признаками форматирования являются:

а) положение (location) – характеризует положение строки на странице (координатное);
б) форматирование (format) – содержит такие признаки, как размер шрифта, жирность, курсив;
в) признаки объектов (object) – признаки, является ли строка частью таблицы или картинки, например, в виде подписей к осям графика.

1.4 Анализ химических текстов

1.4.1 OSCAR4

OSCAR4 (Open Source Chemistry Analysis Routines) [11] – open-source java-библиотека для выделения химических сущностей. Библиотека является средством для решения задачи NER (Named Entity Recognition – распознавание именованных сущностей) применительно к предметной области химии. OSCAR4 позволяет выделять следующие типы химических сущностей:

а) CM – химическое вещество, молекула;
б) RN – реакция;
в) ONT – онтологический термин.

Схема работы библиотеки выглядит следующим образом: на вход подается строка, являющаяся предложением исходного текста. Данная строка разбивается на список токенов, который затем подается в блок распознавания именованных сущностей. Выходом данного блока является список сущностей, где каждая сущность состоит из одного или нескольких подряд идущих токенов.

Всего в OSCAR4 реализовано три NER модуля: RegexRecogniser, PatternRecogniser и MEMMRecogniser. RegexRecogniser использует регулярные выражения для нахождения номеров веществ в различных базах данных, например, «NSC-2648». PatternRecogniser использует технику n-граммного анализа: для текущего слова рассматриваются 1-4 буквенные подпоследовательности в слове, и ищутся вхождения таких последовательностей в множество химических и нехимических слов. MEMMRecogniser использует Maximum Entropy Markov Model для выделения химических сущностей.

Можно привести следующий пример работы OSCAR4: пусть исходным предложением является «Saponite is intercalated in an effective way by Cr and Al-Cr polycation, providing solid with basal spacing of ca». Выделенные химические сущности показаны в таблице 4.

Таблица 4. Пример работы OSCAR4.

Токен	Тип сущности
Saponite	CM
Cr	CM
Al-Cr	CM
polycation	ONT

1.4.2 ChemicalTagger

ChemicalTagger [12] является open-source библиотекой для анализа неструктурированной химической литературы. Конечным результатом анализа является набор предложений данного текста, где каждое предложение является деревом. Узлы дерева представляют собой токены, или группы токенов, объединенных логически в соответствии с моделью предложения.

Процесс обработки текста можно разделить на пять этапов – нормализация текста, токенизация, тегирование, разбиение на фразы, и идентификация действий.

Нормализация текста подготавливает исходный текст для дальнейшей обработки. В частности, на данном этапе фильтруются непечатные символы, добавляется необходимое количество пробелов между специальными символами (скобки), числа с плавающей запятой приводятся к унифицированному представлению.

Следующий этап обработки – токенизация текста, т.е. разбиение его на отдельные слова (токены). В качестве токенайзера используется модифицированный white-space tokenizer, поскольку токенизация по непробельным символам может вызывать проблемы в химических текстах.

Далее выполняется разметка полученных токенов. ChemicalTagger использует три различных таггера, которые выполняются один после другого. Сначала используется OSCAR, который ставит в соответствие токенам их химическую роль – вещество, процесс, реакция, онтологический термин и т.д. Затем выполняется Regex-tagger. Он помечает токены, специфичные для химических текстов, которые не были помечены OSCAR на предыдущем шаге. Для поиска токенов Regex-tagger использует регулярные выражения, хранящиеся в файле вместе с соответствующими тегами. В набор правил данного таггера входят регулярные выражения для идентификации размерностей, агрегатных состояний веществ, действий, проводимых в процессе химического эксперимента. Последний используемый таггер – POS tagger, который помечает каждый токен меткой – частью речи данного токена. В качестве реализации выбран POS tagger, входящий в состав библиотеки Stanford openNLP.

После того, как процесс разметки текста завершен, выполняется построение фраз. Цель данного этапа – представить каждое предложение текста в виде дерева, описывающего структуру предложения. Поскольку химия является достаточно формальной предметной областью, то в качестве компонента построения дерева предложения используется ANTLR (Another Tool for Language Recognition). ANTLR является LL(*) парсером, и используется по большей части для генерации ACT

формальных языков, например, языков программирования, применение этого средства для анализа структуры предложений естественного языка является новым подходом.

Последним этапом обработки является приписывание ролей различным частям полученного синтаксического дерева предложения. В текущей версии реализовано два вида такого приписывания – приписывание ролей к фразам и приписывание ролей к молекулам. Приписывание ролей к фразам работает на основе выделенных экспертами слов, характеризующих действие (всего 21 слов). Например, фраза «DMAP (2.48 g, 11.8 mmol) was dissolved in THF (50 mL)» будет помечена как «Dissolve-Phrase», поскольку во фразе присутствует глагол «dissolved». Приписывание ролей к молекулам работает аналогичным образом.

1.5 Извлечение ключевых фраз

1.5.1 Общее описание задачи и подходов к решению

Задача извлечения ключевых фраз из текста является родственной решаемым задачам, использует схожие методы и алгоритмы. Суть задачи состоит в том, чтобы извлечь из полного текста документа набор ключевых фраз – слов и словосочетаний, наиболее полно характеризующих содержание документа. Существует два основных подхода к решению данной задачи – подход, использующий машинное обучение, и подход, работающий с правилами, изначально заложенные в систему.

Общая схема извлечения ключевых фраз состоит из следующих этапов:

1. предварительная обработка текста;
2. отбор кандидатов в ключевые фразы;
3. вычисление признаков для каждого кандидата;
4. отбор ключевых фраз из числа кандидатов.

На этапе предварительной обработки устраняется «шум» - рисунки, таблицы, номера страниц. Отбор кандидатов обычно производится L-граммным методом: используется скользящее окно переменной длины 1..L, которое каждый раз смещается на один токен вправо, каждая фраза, попавшая в скользящее окно, обрабатывается независимо. Для фильтрации L-грамм часто используются стоп-словари и фильтрация по морфологическим признакам – удаление предлогов, междометий и т.д.

На следующем этапе для каждого кандидата в ключевые фразы рассчитывается ряд признаков, позволяющих принять решение, является ли данный кандидат ключевой фразой, или нет. Набор кандидатов ранжируется по значениям признаков, и отбираются первые лучшие кандидаты, либо кандидаты, преодолевший минимальный порог значений признаков.

1.5.2 Оценка эффективности

Для оценки эффективности методов извлечения ключевых фраз обычно используются две характеристики – точность и полнота. Данные характеристики сравнивают ключевые слова, найденные автоматически, с ключевыми словами, выделенными читателями-экспертами. Для объединения точности и полноты в одну характеристику используется F-мера.

Точность позволяет оценить, насколько ключевые фразы, найденные автоматически, совпадают с ключевыми фразами, найденными экспертами. Точность определяется как отношение количества *экспертных* ключевых фраз, найденных автоматически, к общему количеству *найденных автоматически* ключевых фраз.

$$Precision = \frac{|T_{exp} \cap T_a|}{|T_a|}$$

где T_{exp}-множество ключевых фраз, найденных экспертами,

T_a- множество ключевых фраз, найденных автоматически.

Полнота указывает, насколько ключевые слова, найденные автоматически, покрывают ключевые слова, найденными экспертами. Полнота определяется как отношение количества *экспертных* ключевых фраз, найденных автоматически, к общему количеству *экспертных* ключевых фраз.

$$Recall = \frac{|T_{exp} \cap T_a|}{|T_{exp}|}$$

F-мера является взвешенным гармоническом средним точности P и полноты R и определяется как

$F = \frac{1}{\alpha\frac{1}{P}+(1-\alpha)\frac{1}{R}}$, $\alpha \in [0,1]$

При $\alpha = 0.5$ F-мера придает одинаковый вес точности и полноте (сбалансированная, или F_1 мера), и записывается как

$$F_1 = \frac{2PR}{P+R}$$

1.5.3 KEA

KEA (keyword extraction algorithm) [13] – один из первых алгоритмов, использующих машинное обучение (наивный Байесовский классификатор) для извлечения ключевых фраз. В данном алгоритме используется два признака для классификации – TF-IDF и признак первого вхождения (first occurrence). Эти признаки

используются во множестве других алгоритмах, и называются «стандартными признаками».

TF-IDF показывает специфичность данной фразы по отношению к остальным фразам документа и вычисляется как произведение TF (Term Frequency) на IDF (Inversed Document Frequency), и описывается следующим соотношением:

$$TFIDF(t, D) = \frac{freq(t, D)}{size(D)} * \left|\log_2 \frac{df(t)}{N}\right|$$

где $freq(t, D)$ – число вхождений фразы t в документ D,

$size(D)$– число слов в D,

$df(t)$ – число документов рассматриваемого текстового корпуса, содержащих t,

N – количество документов в корпусе.

Второй признак, используемый в KEA – признак первого вхождения (first occurrence). Он вычисляется как позиция первого вхождения первого слова фразы, деленная на количество слов в документе. Значения признака лежат в диапазоне [0..1], и показывают, насколько далеко находится ключевая фраза относительно начала документа.

1.5.4 Sztergak

Sztergak [14] решает задачу воспроизведения ключевых фраз, отобранными экспертами, как задачу обучения с учителем. Используется стандартный подход к извлечению ключевых фраз, описанный выше. На заключительном этапе отбираются top-15 фраз.

На этапе предварительной обработки устраняются строки, содержащие бесполезную информацию, например, содержание ячеек таблиц, номера страниц и т.д. Это происходит путем сравнения длины строки со средней длиной в тексте, а также использованием регулярных выражений. Затем выполняется разметка текста – выделение токенов и предложений, частеречная разметка.

Следующий этап – отбор кандидатов в ключевые фразы – выполняется L-граммным методом со скользящим окном длиной 1..4. Отбираются кандидаты, не содержащие стоп-слова на концах L-граммы, и также состоящие из слов с POS-метками из множества {существительное, прилагательное, глагол}. Также используется стемминг для работы с кандидатами в унифицированной форме.

1.5.4.1 Набор признаков

Признаки, использующихся для отбора ключевых фраз, могут быть сгруппированы в четыре группы:

1. Признаки уровня фразы
2. Признаки уровня документа
3. Признаки уровня корпуса
4. Признаки, основанные на внешних знаниях

Также используются «стандартные признаки», введенные в алгоритме KEA и рассмотренные выше.

1.5.4.1.1 Признаки уровня фразы

Признаками уровня фразы являются такие признаки, которые можно рассчитать, опираясь только на текущую фразу, без использования внешних контекстов. Такими признаками являются:

а) длина фразы – число слов в текущей фразе;
б) POS – признак, содержащий части речи слов, входящих во фразу;
в) признак суффикса (suffix feature) – булев признак, показывающий, содержит ли исходная фраза окончание из множества Michigan Sufficiency Exams Suffix List.

1.5.4.1.2 Признаки уровня документа

Для данной группы признаков контекстом является документ целиком. Эта группа содержит:

а) Признак аббревиатуры (acronymity feature) – булев признак, принимающий значение истина, если текущая фраза является расшифровкой аббревиатуры, которая содержится в этом же документе.
б) Признак поточечной взаимной информации (PMI, pointwise mutual information) – обобщает меру поточечной взаимной информации для фраз с произвольным количеством слов. Значение признака вычисляется как

$$pmi(t_1, \dots, t_n) = \frac{\log \frac{p(t_1, \dots, t_n)}{p(t_1) * p(t_2) * \dots * p(t_n)}}{\log p(t_1, t_2, \dots, t_n)^{n-1}}$$

где $p(t_i)$ – вероятность появления i-го токена фразы в документе

в) Синтаксический признак – вычисляется как средняя минимальная нормированная глубина листьев синтаксического дерева, которые содержат слова ключевой фразы.

1.5.4.1.3 Признаки уровня корпуса

а) Sf-idf признак предназначен для работы с секциями документов. По аналогии с TF-IDF, признак характеризует специфичность ключевой фразы относительно секций документов – значение будет высоким, если фраза содержится в большом числе

секций текущего документа, и в малом числе секций документов корпуса. Признак вычисляется как

$$sf - isf(k, d) = sf(k, d) * isf(k)$$

Где k – текущая фраза, d – документ.

б) Информативность – булев признак, принимающий значение истина, если текущая фраза является одной из множества всех ключевой фразой, определенных экспертами для тестового корпуса документов.

1.5.4.1.4 Внешние признаки

Внешние признаки позволяют получить знания о кандидатах в ключевые фразы, основываясь на внешних источниках.

а) Признак Wikipedia – булев признак, принимающий значение истина, если существует статья в Wikipedia с таким же названием, как текущий кандидат в ключевые фразы.

1.5.4.2 Результаты

В качестве обучающего и тестового корпуса были использованы соответственно 144 и 100 публикации из библиотеки ассоциации по компьютерной лингвистике (ACL). Оценки были получены путем сравнения top-15 выделенных ключевых фраз с ключевыми фразами, назначенными читателями библиотеки. Результаты оценки представлены в таблице 5.

Таблица 5. Результаты работы Sztegak.

Комбинация признаков	**F-мера**
Стандартные признаки (SF)	14.57
SF + признак длины фразы	20.93
SF + грамматический признак (POS)	19.60
SF + суффикс признак	16.35
SF + акроним признак	16.87
SF + поточечной взаимной информации (PMI)	15.68
SF + синтаксический признак	14.20
SF + sf-idf	14.79
SF + информативность	15.17
Все признаки	23.82
Все признаки исключая информативность	22.11

Из таблицы видно, что существенное улучшение получено добавлением к стандартным признакам признака длины фразы. Это показывает, что хотя стандартные признаки выдают неплохие результаты, но они не учитывают, что ключевые фразы, назначенные читателями, часто состоят из нескольких слов. Также следует отметить, что морфологические признаки, такие, как признак части речи (POS) или признак суффиксов

были среди наиболее эффективных, что свидетельствует о некоторой общей структуре ключевых фраз.

1.5.5 Wingnus

Отличительной особенностью системы Wingnus [15] является то, что она использует логическую структуру документа в процессе определения кандидатов в ключевые фразы. Данный подход не использует стандартные признаки, и может комбинироваться с другими системами для большей эффективности. Под логической структурой документа понимается иерархия компонентов – название, авторы, организации, заголовки 1-3 уровней и т.д.

1.5.5.1 Обработка текстов

Для восстановления логической структуры документа Wingnus использует оригинальные pdf-документы. Система работает также и с информаций в текстовом виде, но точность распознавания логической структуры при этом будет низкой. Для получения оригинальных статей используется поисковая система Google scholar, для нахождения статей используется сходство заголовков – если значение меры сходства больше 0.7, то текущая статья считается найденной. Из 140 статей тренировочного и 100 статей текстового корпуса было найдено 116 и 76 статей соответственно.

Для получения логической структуры документа используется разработанное ПО SectLabel. SectLabel классифицирует каждую строку документа логическом классом - название, авторы, заголовки различных уровней и т.д. Используется информация о форматировании документа – размер шрифта, положение строку в документе. Если pdf-файл статьи не удалось найти, SectLabel может обрабатывать простой текст с потерей точности и полноты.

1.5.5.2 Извлечение ключевых фраз

Для извлечения ключевых фраз предлагается использование регулярных выражений. Эксперименты показали, что наибольшее значение полноты обеспечивается при обработке таких разделов документа, как название, заголовки различных уровней, введении, заключении и первой строки каждого параграфа тела документа.

Для проведения классификации используется Наивный Байесовский классификатор, встроенный в программный комплекс Weka. Используются следующие признаки:

а) F1 – значение TF-IDF меры;

б) F2 - частота ключевых фраз;

в) F3 - частота подстрок ключевых фраз;

г) F4, F5 - первое и последнее вхождение (смещение слова);

д) F6 - количество слов в фразе;

е) F7 - typeface attribute - двоичный признак указывающий на выделение одного из слов фразы жирным шрифтом или курсивом (доступен если найдена статья в PDF формате);

ж) F8 - inTitle - Двоичный признак указывающий на содержание фразы в названии;

з) F9 - titleOverlap – сколько раз фраза встречается в других документах;

и) F10-F14 - двоичные признаки указывающие на содержании фразы в заголовке, аннотации, вступлении, обзоре существующих решений или заключении;

к) F15-F19 - частота содержания фразы в заголовке, аннотации, вступлении, обзоре существующих решений или заключении.

Для оценки эффективности признаки F1 и F4 были объединены в базовый набор признаков. Расчет F-меры проводился для базового набора и совместно с каждым из остальных признаков. Результаты представлены в таблице 6.

Таблица 6: Результат работы Wingnus.

Признак	**F-мера, %**	**Признак**	**F-мера, %**
Base (Базовый)	23.42	Base + F11	23.42
Base + F2	21.11	Base + F12	23.42
Base + F3	24.57	Base + F13	23.75
Base + F5	24.08	Base + F14	22.28
Base + F6	25.06	Base + F15	22.11
Base + F7	23.42	Base + F16	23.59
Base + F8	22.77	Base + F17	22.60
Base + F9	22.28	Base + F18	23.26
Base + F10	23.42	Base + F19	21.95

1.5.6 KP-Miner

KP-miner [16] – система извлечения ключевых фраз из текстов на английском и арабском языках. В отличие от остальных систем, KP-miner основана не на модели машинного обучения, а на подходе, в котором к кандидатам в ключевые фразы применяются различные правила для фильтрации для их отбора. Процесс извлечения ключевых фраз проходит в три этапа:

а) выделение кандидатов в ключевые фразы;

б) подсчет весов кандидатов;

в) окончательное уточнение ключевых фраз.

1.5.6.1 Выделение кандидатов в ключевые фразы

На начальном этапе сканируется последовательность слов до момента, пока не встретится знак пунктуации или стоп-слова. Полученная фраза разбивается на L-граммы длиной от единицы до длины фразы минус единица. Полученные L-граммы сохраняются в двух вариантах – исходном и нормализованном, для нормализации используется слабый стемминг – только первый шаг Porter stemmer.

Для фильтрации полученных L-грамм используются два правила. Первое правило состоит в том, что ключевая фраза должна быть представлена в документе не менее n раз, в английской версии системы по умолчанию n=3. Если документ короткий, то n понижается в зависимости от длины документа.

Второе правило ограничивает позицию первого вхождения кандидата в ключевые фразы. Было замечено, что в длинных документах фразы, встречающиеся первый раз достаточно далеко от начала документа, редко являются ключевыми. В правиле определена константа отсечки, определяющая число слов, после которого фраза не рассматривается, если она появилась в первый раз позже, чем данная константа. В настоящее время значение константы равно 400 слов.

В отличие от других систем, K-miner не устанавливает ограничений на длину ключевых фраз, но было замечено, что ключевые фразы редко имеют длину больше трех слов.

1.5.6.2 Подсчет весов кандидатов

Для подсчета весов кандидатов и последующего их ранжирования используется скорректированная мера TF-IDF. Использование TF-IDF напрямую для ранжирования кандидатов в ключевые фразы показывает неудовлетворительные результаты, поскольку частота вхождения длинных фраз обычно значительно ниже, чем частота вхождения коротких. Поэтому к стандартной TD-IDF мере добавлены повышающие коэффициенты для компенсации длинных фраз:

$$W = Tf * IDF * B * P$$

Где W – итоговый ранг фразы

B – коэффициент компенсации длинных враз

P – атрибут, определяющий положение фразы в документе. Если позиционная информация не учитывается, то P=1

1.5.6.3 Окончательное уточнение ключевых фраз

KP-miner позволяет указать необходимое количество ключевых фраз, которое нужно получить (по умолчанию, пять). При отборе кандидатов учитывается только

наиболее длинная фраза. Для выполнения этого требования при окончательном отборе top ключевых фраз проверяется, являются ли данные фразы частью более длинных фраз. Если являются, то их частота вхождения уменьшается на число вхождений более длинной фразы, затем веса фраз пересчитываются, и кандидаты отсортировываются по новым значениям.

2 Общее описание системы терминологического анализа

Основным назначением разрабатываемой системы терминологического анализа является выявление терминологического спектра публикаций – автоматизированное выделение из текстов терминоподобных словосочетаний, общенаучных терминов, названий химических соединений и реакций, аббревиатур и т.д., с последующим проведением различных видов анализа полученных спектров в масштабах коллекции документов.

Под терминоподобными словосочетаниями в настоящей работе понимаются одно или несколько подряд идущих слов (токенов) в тексте, претендующих на роль термина – значимого понятия предметной области или конкретного научного исследования. Терминоподобные словосочетания, выделяемые из текста, по своему написанию часто не похожи на общепринятые термины (публикуемые в канонизированном виде в тезаурусах), поскольку их написание отражает употребление понятия в различном контексте, с естественным языковым окружением. Однако именно такие терминоподобные словосочетания отражают реальное содержание публикации, а с ним и суть реальных явлений и процессов, встречающихся в научных исследованиях, что делает анализ таких словосочетаний чрезвычайно полезным. С использованием таких специфичных терминологических словарей можно организовывать как новые эффективные процедуры поиска, так и новые методы анализа текстовой информации.

Для проведения анализа динамики изменения терминологической базы необходимо анализировать однородные текстовые подборки за некоторый временной промежуток. Для этого в процессе терминологического анализа, одновременно с выделением терминоподобного словосочетания сохраняется контекст его употребления – место в тексте, в котором встретилось словосочетание, а также дата публикации данного текста. Хранение контекста употребления позволяет вычислять различные характеристики словосочетаний для заданного временного диапазона, выявлять восходящие и нисходящие тренды, искать особенности во временной динамике терминов.

Разработанная система терминологического анализа состоит из следующих основных модулей (рисунок 1):

а) ***Модуль преобразования текстовых материалов***. Предназначен для обработки полнотекстовых источников (текстовых коллекций) с их преобразованием из pdf формата во внутреннее представление (пригодное для дальнейшей обработки) и сохранением в БД. Для каждого документа подборки происходит выделение логических блоков и нормализация текста.

б) ***Модуль терминологического анализа***. Выполняет анализ текстов с одновременным выделением терминоподобных словосочетаний, общенаучных терминов, названий химических соединений и реакций, аббревиатур и т.д. В процессе терминологического анализа вся получаемая информация о терминах заносится в базу данных, в том числе морфологические и другие признаки. Для выделения терминоподобных словосочетаний используется модифицированный метод L-граммного анализа, согласно которому из текста выделяются всевозможные словосочетания длины L, которые затем тестируются на «терминоподобность».

в) ***Модуль отчетов***. Предназначен для анализа полученных терминологических спектров коллекций документов, с генерацией различного вида отчётов для экспертов в предметной области. Например, отчет может представлять собой таблицу, состоящую из терминоподобного словосочетания и его основных характеристик: частота встречаемости в текстах по годам, среднеквадратичное отклонение частоты встречаемости и т.д.

г) ***База данных***. Предназначена для хранения размеченных обработанных текстов, терминов, служебной информации т.д.

д) ***Веб-интерфейс пользователя***. Позволяет эксперту в предметной области работать с полученными результатами – просматривать тексты, терминологическую разметку, выполнять и анализировать различные отчеты.

Элементы общей архитектуры системы терминологического анализа представлены на рисунке 1.

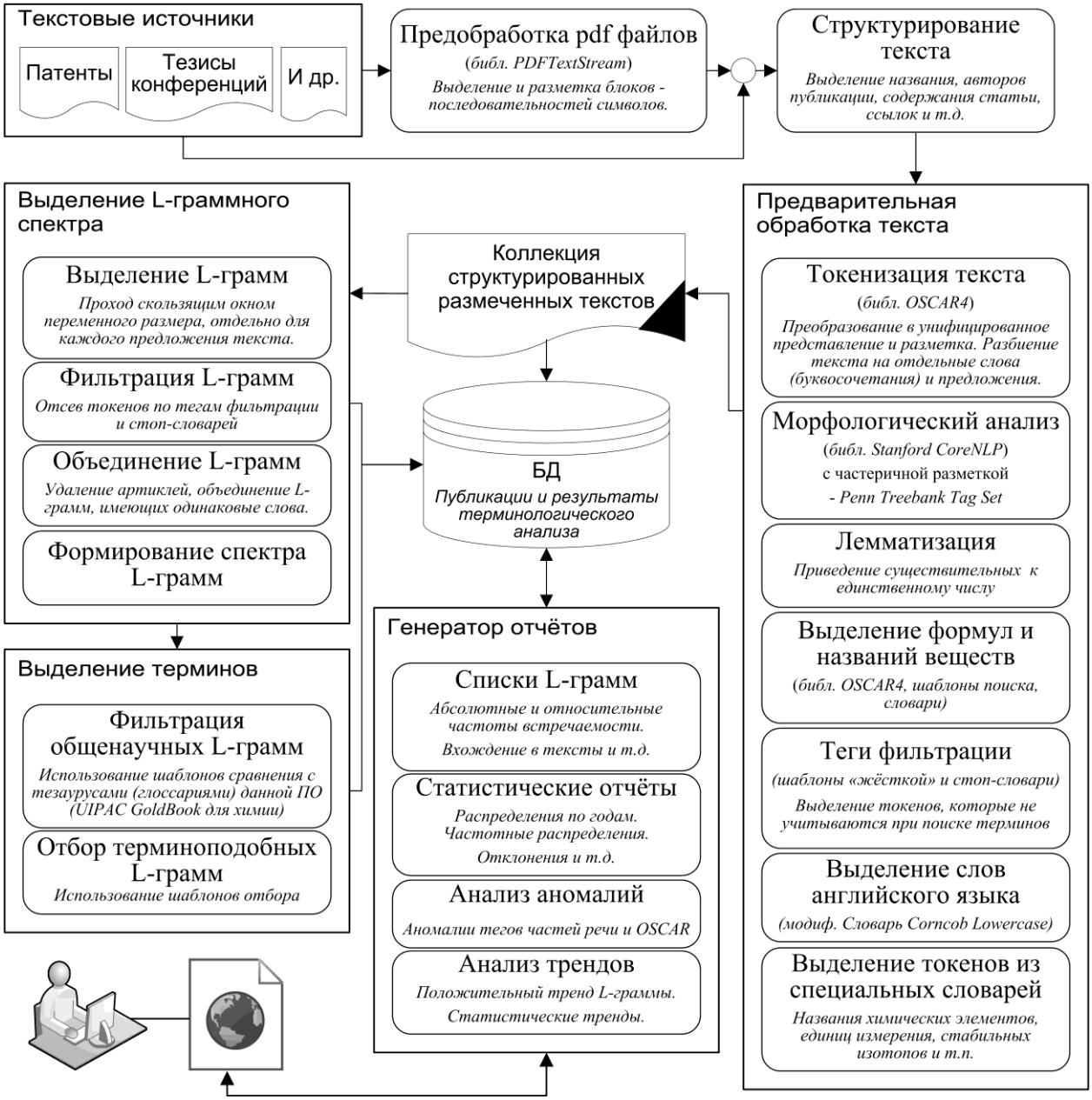

Рисунок 1. Общая схема работы системы терминологического анализа.

Преобразование текстовых PDF материалов. В настоящее время большинство публикаций представлено в формате pdf. Для эффективной обработки информации, содержащейся в таких файлах, необходимо обеспечить их корректный перевод в текстовое представление, в том числе обеспечить выделение структурированных данных: названия, авторов публикации, содержания статьи, ссылок и т.д.

В рамках работы апробирован метод выделения текста из pdf файлов публикаций, использующий при выделении структуры информацию о размере, положении, форматировании текста в блоках документа. Результатом обработки pdf-файла публикации является полностью структурированный документ, который затем сохраняется в БД для последующего анализа.

Предварительная обработка текста заключается в преобразовании текста в унифицированное представление и его разметке. При этом происходит разбиение текста на отдельные слова и предложения (процесс токенизации), с последующим проведением морфологического анализа текста (в том числе выделении формул и названий химических соединений), фильтрации ненужных “шумовых” слов и буквосочетаний и т.д.

Морфологический анализ (применяется библиотека Stanford CoreNLP) сопоставляет каждому слову набор тегов частеречной разметки (Penn Treebank Tag Set). Выделение формул и названий веществ, необходимое для дополнительной фильтрации терминоподобных словосочетаний, осуществляется с помощью библиотеки OSCAR 4 (Open-Source Chemistry Analysis Routines), а также путём использования разработанных шаблонов отбора.

На этапе предварительной обработки выставляются теги «жесткой» фильтрации, отмечающие слова-токены, которые не должны учитываться при поиске терминов (используются специально разработанные шаблоны и стоп-словари). На данном этапе учитываются такие параметры токена, как вхождение цифр, символов размерностей, символов химических элементов, скобок и т.д.

Отмечаются общеупотребительные слова английского языка (например, *«ability»*, *«about»* и др.) за исключением общенаучных слов, представляющих интерес для данной ПО (*«abrasion»*, *«absorption»* и др.). Для этой цели используется словарь английских слов, созданный на основе словаря Corncob Lowercase (~57 тыс. слов в н.м.).

Выявление терминоподобных словосочетаний основано на использовании модели L-граммного анализа текста. Под L-граммой далее понимается последовательность из $L \geq 1$ последовательно идущих слов (токенов) текста, с возможным пропуском отдельных слов. Выделение L-грамм осуществляется в отдельности для каждого предложения, внутри которого осуществляется проход скользящим окном переменного размера, с применением к такому окну процедур отбора L-грамм.

В рамках работы разработаны алгоритмы отбора терминоподобных L-грамм, а сам метод L-граммного анализа текста адаптирован к области химии, особенностью которой является наличие большого числа формул, названий веществ, специальных символов и т.д. Процедуры отбора терминоподобных L-грамм с учётом особенностей данной ПО осуществляются путём применения различных правил (шаблонов) фильтрации. Шаблоны основаны на использовании: тегов морфологического анализа, вхождения слов в словарь общеупотребительного английского языка, в глоссарии общехимических терминов и аббревиатур (UIPAC GoldBook) и т.д. Например, используются специально созданные

словари названий химических элементов, единиц измерения, стабильных изотопов и т.п. Для L-грамм с $L \geq 2$ применяются шаблоны для первого, последнего и срединного слова.

Окончательное формирование спектра выделенных терминоподобных L-грамм и создание базы данных публикаций являются последними этапами процедуры автоматизированного терминологического анализа. Для разработанной модели терминологического анализа выполнена программная реализация на платформе Java. Работа пользователя–эксперта с БД в настоящий момент возможна через веб-интерфейс.

3 Преобразование материалов конференций

В настоящее время большинство научных публикаций распространяются в виде pdf-файлов. Для эффективной обработки текстовой информации, содержащейся в таких файлах, необходимо обеспечить их корректный перевод в текстовое представление – выделение названия, авторов публикации, содержания статьи, библиографических ссылок и т.д. Однако, как было отмечено в литературном обзоре, выделение таких логических блоков из pdf-документа является весьма нетривиальной задачей, поскольку pdf файл не содержит никакой информации о логической структуре документа.

В рамках работы был разработан метод выделения текста из файлов публикаций, использующий при выделении структуры информацию о размере, положении, форматировании текста в блоках документа. Вся последовательность действий по обработке pdf-файла публикации состоит из следующих этапов:

1. Выделение блоков. Блоком считается последовательность символов исходного документа, имеющих одинаковое форматирование текста (одинаковые параметры bold, italic, underline). Выделение последовательности символов выполняется при помощи java-библиотеки PdfTextStream [17]. Каждый символ является отдельным объектом с такими атрибутами, как номер строки, размер и стиль шрифта, координатное положение в документе. Задача PdfTextStream состоит в том, чтобы распознавать границы строк, выделять колонки при многоколоночной верстке документа, определять правильную последовательность символов.
2. Нормализация. Из полученного на предыдущем этапе списка блоков удаляются пустые (состоящие только из пробельных символов) блоки.
3. Объединение блоков. На данном этапе объединяются блоки, находящиеся на одной строке исходного документа. Результатом данного этапа является список суперблоков, где каждый суперблок представляет собой один или более объединенных блоков.
4. Разметка суперблоков. Каждому суперблоку присваивается тег (метка), соответствующий элементу логической структуры документа – название, авторы и т.д. Разметка выполняется с помощью набора таггеров. Каждым таггером последовательно принимается решение о присвоении текущему суперблоку тега. При принятии решения таггер может руководствоваться такими параметрами суперблока, как номер первой и последней строки, средневзвешенные параметры форматирования текста, информацией о положении блока на странице и другими.
5. Фильтрация суперблоков. На данном этапе устраняются некоторые ненужные суперблоки, в частности, неклассифицированные блоки до секции с названием

публикации. Обычно такие блоки содержат различную служебную информацию, например, дата и место проведения конференции.

6. Нормализация текста. Перед сохранением итогового документа в БД выполняется предварительная обработка текста. Обработка состоит из замены всех вариантов написания тире, дефисов и кавычек на один фиксированный вариант, а также из удаления пробелов перед скобками в написании кристаллографических индексов. Данный этап выполняется с помощью набора регулярных выражений для поиска и замены.

Для корректной работы таггерам необходима настройка на формат документа, который обрабатывается в данный момент. Обычно публикации из одного источника, например, тезисы конференции одного года, имеют одинаковую структуру документа, поэтому достаточно настроить таггеры одни раз для всех документов из фиксированного источника. Настройка таггеров производится через конфигурационный файл. Применительно к анализу тезисов конференции, для каждого года проведения конференции в конфигурационном файле имеется секция, описывающая параметры всех таггеров для данного года.

Всего в системе имеются следующие таггеры:

а) TitleClassifier – определяет блоки, относящиеся к названию публикации. Использует параметры MaxEndLine (максимально возможный номер последней строки) и MaximumNotBoldSymbols (максимально возможное количество символов, не являющихся жирными). Обычно все символы в заголовке являются жирными, но иногда могут встретится нежирные символы, например, в случае надстрочных и подстрочных индексов в формулах веществ. Для работы с такими случаями используется параметр MaximumNotBoldSymbols.

б) AuthorClassifier – определяет блоки, относящиеся к авторам публикации. Использует параметры MinStartLine, MaxStartLine, MinimumItalicLength. Параметр MinimumItalicLength - определяет минимально возможное количество курсивных символов в блоке. Если блок имеет меньшее количество курсивных символов, чем значение данного параметра, то текущий суперблок не будет рассматриваться как блок с авторами публикации.

в) OrganisationClassifier – определяет организации авторов публикации. Используемые параметры: AverageItalic (отношение количество символов, имеющих курсивное начертание, к количеству остальных символов), MinimumBoldLength (минимально возможное количество символов, имеющих жирное начертание), MaxEndLine.

г) AuthorEmailClassifier – определяет e-mail авторов публикации. Для определения блоков данного класса используется регулярное выражение для проверки email-адресов. Предварительно из строки вырезается символ «*» (с него часто начинается новая строка с email-адресами авторов), строка разделяется по символу «;» (обычно указываются несколько email подряд, разделенных данным символом).

д) ReferencecClassifier – определяет секцию с библиографическими ссылками. Для определения начала секции использует возможные названия заголовков секции. Параметр ReferencesNames определяет возможные названия заголовков, например, «references, recent papers, selected publications».

Таким образом, результатом обработки pdf-файла публикации является структурированный документ с такими полями, как название, авторы, полный текст публикации и т.д. Полученный документ затем сохраняется в БД для последующего анализа.

4 Модель терминологического анализа

4.1 Основные этапы анализа

Для модуля терминологического анализа была разработана модель, учитывающая особенности тестов химической направленности. В основе модели лежит метод L-граммного анализа [18]. Общая схема метода показана на рисунке 2. Данный метод предусматривает несколько этапов обработки текста:

1. Предобработка – текст разбивается на отдельные предложения и слова (токены), вычисляются характеристики предложений и токенов.
2. Обработка – по тексту осуществляется проход скользящим окном длины L. К каждой позиции скользящего окна (L-грамме) применяются шаблоны для определения, является ли текущая L-грамма терминоподобной, или нет. Если текущая L-грамма является терминоподобной, то она сохраняется в локальное хранилище. Длина скользящего окна L меняется в диапазоне $[1 \,..\, L_{max}]$.
3. Постобработка – на данном этапе используются различные фильтры для отсева отобранных L-грамм. Примером фильтра может служить фильтрация по количеству вхождений L-граммы в тексты анализируемой подборки.

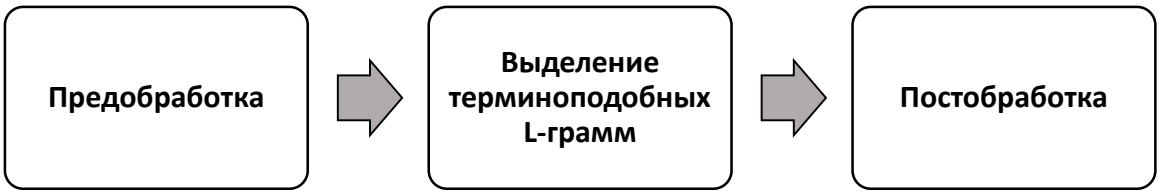

Рисунок 2. Общая схема модели терминологического анализа.

Для работы с L-граммами (фильтрации, сортировки, построение отчетов) используется две основные характеристики, на основании которых вычисляются все остальные:

а) f_{abs} – абсолютная частота вхождения L-граммы во все тексты анализируемой подборки. Показывает, сколько раз L-грамма суммарно встретилась во всех текстах подборки.

б) f_{text} – текстовая частота вхождения L-граммы. Показывает, в скольких текстах подборки встречается данная L-грамма.

4.2 Предварительная обработка текста

Предварительная обработка текста включает в себя следующие этапы: токенизация, разметка токенов, модификация токенов.

Рисунок 3. Этапы предварительной обработки текста.

Этап токенизации заключается в разбиении текста на отдельные слова и предложения. Изначально для выполнения этапа использовался токенайзер из библиотеки Stanford CoreNLP [19], но было обнаружено, что данный токенайзер имеет проблемы с обработкой химических текстов, в частности, разбивает формулы и названия веществ на отдельные токены. Поэтому в настоящее время для выполнения токенизации используется модифицированный токенайзер библиотеки OSCAR4. Данный токенайзер спроектирован с учетом обработки текстов химической и биологической направленности, учета сложных химических формул.

Сборка токенов в предложения выполняется после этапа токенизации. Данный этап выполняется при помощи WordToSentenceAnnotator библиотеки Stanford CoreNLP.

На этапе разметки каждому токену ставится в соответствие набор тегов с их значениями – характеристики токенов. Такими характеристиками являются:

а) Позиция токена в тексте. Задается в виде пары [offsetBegin, offsetEnd], где offsetBegin – количество символов от начала текста до первого символа данного токена, offsetEnd – количество символов до последнего символа токена. Значение данного тега вычисляется на этапе токенизации.

б) Морфологическая разметка. Данная разметка сопоставляет токену его морфологический тег – часть речи. Разметка выполняется с помощью Stafond CoreNLP, в качестве набора тегов используется Penn Treebank tag set [20].

в) Лемма токена. Данный тег содержит нормальную форму слова.

г) Тег Oscar. Характеризует химическую составляющую токена, показывает, является ли данный токен частью химического вещества, реакцией или онтологическим термином.

д) Тег жесткой фильтрации. Показывает, нужно ли рассматривать L-грамму, содержащую данный токен, как терминоподобную. Механизм жесткой фильтрации подробно рассматривается далее.

Этап модификации токенов состоит из двух частей:

а) Замена токенов – существительных во множественном числе на их лемму (существительное в единственном числе именительного падежа). Эксперименты показали, что замена только такого класса токенов является допустимой.

б) Объединение токенов, описывающих состав катализатора, в один токен. Примерами объединенных токенов являются «1%Pd-1%Cu», «2%Pd-0.7%Re».

4.3 Процедура «жесткой» фильтрации

Процедура жесткой фильтрации предназначена для более полного отсева L-грамм. Если токен помечается тегом жесткой фильтрации, то L-грамма, содержащая такой токен, не будет рассматриваться как терминоподобная.

В основе жесткой фильтрации лежат список предикатов правил и исключений. Общий алгоритм процедуры жесткой фильтрации показан на рисунке 4.

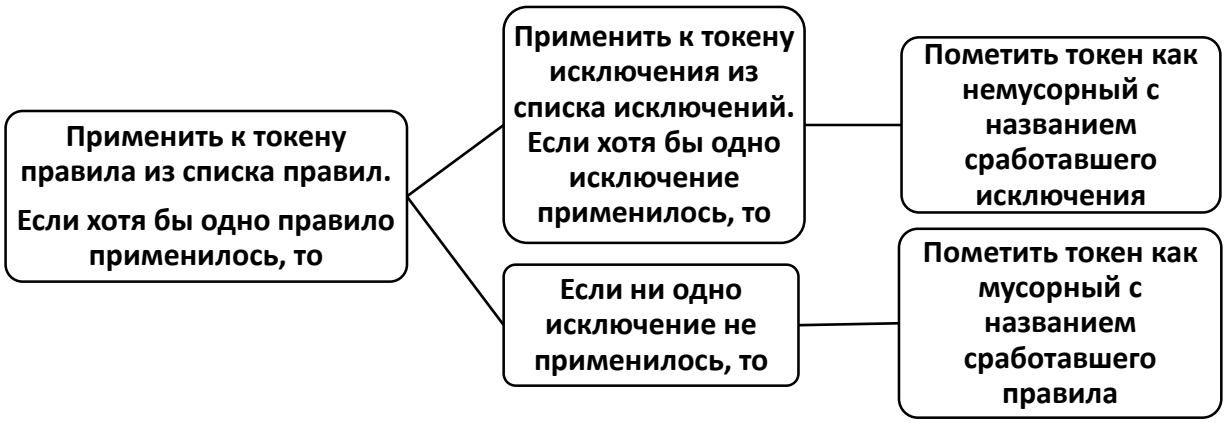

Рисунок 4. Схема применения правил и исключений жесткой фильтрации.

Всего в системе имеются следующие правила и исключения:

а) Правило специальных символов. Истинно, если токен содержит хотя бы один символ из списка специальных символов.

б) Правило стоп-слов. Истинно, если токен содержится в словаре стоп-слов.

в) Правило регулярных выражений. Истинно, если токен удовлетворяет хотя бы одному регулярному выражению из списка регулярных выражений.

г) Правило размерностей. Истинно, если на конце токена имеется подстрока из словаря размерностей.

д) Исключение OSCAR. Истинно, если токен имеет тег Oscar из множества допустимых тегов, а также удовлетворяет указанному правилу.

е) Исключение расширенных регулярных выражений. Истинно, если слово удовлетворяет хотя бы одному расширенному регулярному выражению из списка. Под расширенным регулярным выражением понимается регулярное выражение, в котором дополнительно можно использовать следующие заполнители:

- EL – обозначение любого химического элемента;

- IS – обозначение стабильного изотопа.

4.4 Выделение терминоподобных L-грамм

После того, как выполнена разметка токенов, происходит выделение терминоподобных L-грамм. Данный процесс происходит при помощи различных шаблонов. Задача шаблона состоит в том, чтобы определить, что текущая L-грамма является терминоподобной, либо зафиксировать правило, согласно которому нельзя рассматривать данную L-грамму в качестве терминоподобной.

Шаблон состоит из набора правил, которые последовательно применяются к L-грамме. Некоторые правила являются общими для всех шаблонов, некоторые специфичны только для одного конкретного шаблона. Описание списка шаблонов, списка правил и соответствие между шаблонами и правилами определяется в конфигурационном файле.

Схема работы шаблона выглядит следующим образом:

1. Применить к L-грамме правила для данного шаблона
2. Если хотя бы одно правило из списка правил успешно применилось, то
    a. добавить L-грамму в список отфильтрованных L-грамм с указанием сработавшего правила
3. Иначе
    a. добавить L-грамму в список терминоподобных L-грамм с указанием названия текущего шаблона

Всего в системе имеются следующие общие правила:

1. L-грамма содержит токен с тегом жесткой фильтрации.
2. L-грамма содержит как минимум два одинаковых токена.
3. L-грамма содержит только короткие токены (токены длины меньше трех символов). Данное правило нужно для исключения различного шума, присутствующего в обработанном pdf-документе, например, подписей к осям графиков.
4. L-грамма содержит на конце размерность из списка размерностей. Правило работает различным способом в зависимости от длины L-граммы: для L-граммы, состоящей из одного токена проверяется, не является ли L-грамма частью размерности. Для L-граммы, состоящей из двух и более токенов, проверяется, содержится ли на конце L-граммы размерность. При этом для поиска размерности рассматривается не последний токен L-граммы, а исходная строка целиком, поскольку некоторые размерности могут состоять из нескольких токенов, например, размерность «g/h» состоит из трех токенов [“g”, “/”, “h”].

5. L-грамма содержит два подряд идущих токена с морфологическими тегами из списка морфологических тегов – исключений. Данное правило применяется к L-граммам длины L>=2, и используется для дополнительной фильтрации длинных L-грамм.

Всего в системе используется пять шаблонов, перечисленных далее:

а) UnigrammPattern – шаблон для отбора однограмм (L-грамм с L = 1). Используются общие правила, а также следующие правила:
   - L-грамма является словарным словом из словаря однограмм. Данное правило позволяет исключить известные нетерминоподобные L-граммы, такие, как общеупотребительные слова.
   - L-грамма не удовлетворяет морфологическим шаблонам однограмм. Морфологическим шаблоном в данном правиле является список допустимых частей речи.
   - L-грамма удовлетворяет набору регулярных выражений для разных категорий. Имеется шесть общих категорий регулярных выражений, а так же категории для ионов и катионов.

б) BigrammPattern – шаблон для отбора биграмм (L=2). Используются общие правила, а также:
   - L-грамма не удовлетворяет морфологическим шаблонам биграмм. Морфологическим шаблоном является набор допустимых морфологических тегов для первого токена L-граммы, и набор допустимых тегов для второго токена L-граммы. L-грамма удовлетворяет шаблону, когда первый токен имеет любой морфологический тег из списка допустимых тегов первого токена, и второй токен *независимо от первого* имеет морфологический тег из списка допустимых тегов второго токена. Использование данной структуры шаблона позволяет задать большое число сочетаний тегов для токенов L-граммы, используя коротие списки независимо для каждого токена. Для исключения L-грамм с зависимыми морфологическими тегапи используется общее правило №5, а также следующее правило.
   - L-грамма содержит два токена с морфологическими тегами из списка морфологических тегов-исключений биграмм. Данное правило является частным случаем общего правила №5 применительно к биграммам.

в) ManyGrammPattern – шаблон для отбора L-грамм с L > 2. Используются общие правила, а также:

- L-грамма не удовлетворяет морфологическим шаблонам. Морфологический шаблон состоит из списка допустимых морфологических тегов первого, последнего, а также срединных токенов. Каждый токен проверяется независимо друг от друга по аналогии с морфологическим шаблоном биграмм. Для исключения L-грамм с зависимыми морфологическими тегами используется общее правило №5.
- Морфологические теги первого и второго токена L-граммы не удовлетворяют шаблону.

г) ChemicalUnogramm – шаблон для отбора одногramм, имеющих непосредственный химический смысл, например, название вещества или реакции. Используется общие правила №1, 4, а так же:

- L-грамма имеет любой тег OSCAR, и удовлетворяет морфологическому шаблону. Данное правило нужно для отбора тех L-грамм, которые отфильтровались в шаблоне UnigrammPattern по морфологическому признаку, но являются значимыми терминоподобными L-граммами.

4.5 Постобработка

Заключительным этапом терминологического анализа является постобработка отобранных L-грамм. Имеется два этапа постобработки – тегирование и фильтрация L-грамм.

Тегирование L-грамм на этапе постобработки необходимо для присвоения L-граммам меток, которые могут использоваться в дальнейшем для построения отчетов. В отличие от тегирования на этапе предобработки, на постобработке тегируется L-грамма целиком, а не отдельные токены L-граммы. В настоящее время имеется один таггер – таггер общенаучных химических L-грамм. Задача данного таггера состоит в том, чтобы найти L-граммы, похожие на общехимические термины, и пометить их соответствующим тегом. В качестве критерия сходства используется следующее правило: если L-1 любых слов L-граммы совпадает с L-1 любыми словами некоторого общехимического термина, то данная L-грамма считается общехимической. В качестве словаря общехимических терминов используется IUPAC Goldbook [21].

Фильтрация на этапе постобработки необходима для дополнительного отсева L-грамм, который невозможно провести на других этапах, например, фильтрация по частотным критериям. В системе используется фильтрация по критерию вложенности. Данный критерий основан на поиске вложенных L-грамм, т.е. L-грамм, имеющих подстроки, которые также являются L-граммами. Суть критерия состоит в том, чтобы

отфильтровывать более короткие L-граммы, если в тексте присутствуют длинные L-граммы с такой же абсолютной частотой встречаемости. Схема работы критерия следующая:

1. Для currentLength от 3 до L_{max} (currentLength – текущая длина L-грамм)
    1.1. Для всех L-грамм длины currentLength
        1.1.1. Создать строку из [1 .. currentLength – 1] токенов L-граммы
        1.1.2. Если данная строка является L-граммой в текущем тексте и абсолютная частота данной L-граммы совпадает с абсолютной частотой текущей L-граммы, то удалить короткую L-грамму из списка L-грамм

Также на этапе сохранения списка L-грамм используется неявная фильтрация, основанная на поиске L-грамм в нормальной форме. Если L-грамма в нормальной форме уже существует на момент сохранения в БД, то новая L-грамма не будет создана, вместо этого будет обновлена имеющаяся L-грамма. Под нормальной формой L-граммы понимается L-грамма, в которой исключены союзы («a», «the»), и оставшиеся токены расположены в лексикографическом порядке. Например, для L-граммы «apparently grouped into cluster» нормальной формой является «apparently cluster grouped into», а для «vanadium ion on the catalyst surface» нормальной формой является «catalyst ion on surface vanadium».

5 Описание элементов модели

В данной главе описывается применение разработанной модели в области химического катализа, приводится список используемых словарей, экземпляров правил и исключений.

5.1 Словари

В процессе терминологического анализа используется ряд словарей для фильтрации L-граммного спектра на различных уровнях. Данный набор словарей представлен в таблице 7.

Таблица 7. Словари, использующиеся в процессе терминологического анализа.

Название словаря	Цель использования	Описание	Ссылка	Примеры
Common_science_terms_dictionary	Выделение общенаучных терминов (химических, физико-математических и др.)	Насчитывает около 7500 общенаучных терминов в области химии, физики и математики.	http://goldbook.iupac.org/ [21]	rotational diffusion coefficient, naphthenes, solvation energy, osmotic pressure, reaction dynamics, dynamic field mass spectrometer
Common_words_dictionary	Фильтрация общеупотреби-тельных слов английского языка	Насчитывает около 58 тысяч терминов, широко используемых в английском языке. Создан на основе Corncob Lowercase, из которого были исключены слова, представляющие интерес для катализа (abrasion, absorption, absorptive, aerosol, adhesion и др.).	http://ru.scribd.com/doc/147594864/Corncob-Lowercase [22]	abbreviate, academic, accelerate, accompany, ablation, abomination

Stop_words	Фильтрация L-грамм, не являющихся терминами	Насчитывает около 2060 слов. Составлен из слов и сокращений, которые не должны входить в состав терминоподобных L-грамм		e.g., de, ca., fig., al., co-exist, et, etc., i.e., ltd
Stable_isotops	Фильтрация L-грамм, в состав которых входят цифры	Насчитывает около 250 изотопов. Составлен на основе The Berkeley Laboratory Isotopes Project's	http://ie.lbl.gov/education/isotopes.htm [23]	1H, 2H, 3He, 4He, 6Li, 7Li
Chem_elements	Фильтрация L-грамм, в состав которых входят цифры	Насчитывает 126 химических элементов. Составлен на основе длиннопериодной формы Периодической таблицы Д.И. Менделеева, утверждённой ИЮПАК в качестве основной.	http://ru.wikipedia.org/	H, He, Li, Be, B, C, N, O, F
dimensions	Фильтрация L-грамм, в состав которых входят размерности	Насчитывает около 100 записей. Составлен на основе ИЮПАК GoldBook, в который добавлены сокращения, например, mm, mm2, mm3, mol, g-1, gcat и др.	http://goldbook.iupac.org/	(a.u.), (ev), a.u, °C, ppm, kV, mol, g-1, ml-1, gcat, gcat h

5.2 Правила и исключения жесткой фильтрации

Всего в системе имеются следующие экземпляры правил и исключений:

5.2.1 Правила

1. Правило специальных символов

1.1. Используются следующие символы: «.-,/:()[]+=@®»

2. Правило стоп-слов:

2.1. Используется словарь StopWords. Данный словарь составлен из слов и сокращений, которые не должны входить в состав терминоподобных L-грамм (с L>=1), например, abbreviation, according, additively, afterward, result, discussion, figure и др. Насчитывает около 2060 слов.

3. Правило регулярных выражений:

3.1. 4DigitRule. Отбирает токены, содержащие 4 и более цифры в любом месте. Примеры отбираемых токенов – Ru(0001), Co(0001), Co(0001)-carbide, α-Fe_2O_3(0001), S-2009, 8500C, FQM-3994, AvaSpec-3648.

3.2. 3DigitRule. Отбирает токены, содержащие 3 цифры в любом месте. Примеры отбираемых токенов - CuO(111), Al_2O_3/NiAl(110), BaO_x/Pt(111), $H_4Co_2Mo_{10}O_{38}^{6-}$, ^{129}Xe. Регулярное выражение – «\d{3}»

3.3. 2DigitRule. Отбирает токены, имеющие 1 или две цифры *в начале*. Примеры отбираемых токенов - 5-bromo-3-(N,N-diethylamino-ethoxy)-2-methylindole, 1-hexene, $^{16}O_2$, 13WT%Ni-13WT%Co. Регулярное выражение – «^\d{1,2}».

3.4. dimensions. Отбирает токены, обозначающие размерности. Примеры отбираемых токенов – 100oC, 298K, 2M, 5h.

5.2.2 Исключения

1. RegexpWithOscarRule

1.1. Отбирает токены, имеющих тег CM и удовлетворяющих следующим регулярным выражениям: «\-[A-Za-z]{2}», «\{«, «\[*[A-Za-z]», «\(*[A-Za-z]».

Примеры отбираемых токенов - β-cyclodextrin, γ-Al_2O_3, 2MgO-2Al_2O_3-5SiO_2, [MoO_2Cl{HC(BIM)$_3$}]X, Pt{111}, (PNP*)Ir(I), (PNP*)Ir(III)(H)(CH_2COCH_3).

Примеры отфильтрованных токенов - 128°-y-rotated, π-backdonation, T.C_3H_8(0.9), toluene(%), Mo12~NiMo(0.35), xylene(%), xNO_3-(%).

2. Правила расширенных регулярных выражений:

2.1. Cryst_Index_4digit. Отбираются токены, обозначающие грани катализаторов с четырьмя кристаллографическими индексами.

Примеры отбираемых токенов - Ru(0001), Co(0001), Co(0001)-carbide, α-Fe_2O_3(0001). Примеры отфильтрованных токенов - S-2009, 8500C, FQM-3994, AvaSpec-3648.

2.2. Cryst_Index_3digit. Отбираются токены, обозначающие грани катализаторов с тремя индексами Миллера. Примеры отбираемых токенов - CuO(111),

Al_2O_3/NiAl(110), BaO_x/Pt(111), BaO_2/Pt(111), Ni-Pt-Pt(111), Pt{111}. Примеры отфильтрованных токенов - HZ-180, B-154, R423, P25-2-700.

2.3. Substances_3digit. Отбираются токены, обозначающие химические вещества, в названии которых в любом месте присутствуют три цифры подряд. Примеры отбираемых токенов - $^{15}N_2{}^{18}O$, $H_2{}^{35}S$, $H_2{}^{18}O$-SSITKA, $H_2{}^{16}O/H_2{}^{18}O$, $H_4Co_2Mo_{10}O_{38}{}^6$.

2.4. Isotopes. Отбираются токены, обозначающие изотопы. Используется словарь изотопов, составленный на основе The Berkeley Laboratory Isotopes Project's, который насчитывает около 250 изотопов. Примеры отбираемых токенов - $^{16}O_2$, ^{129}Xe, ^{1}H-NMR, ^{1}HNMR $^{12}C^{16}O$, $^{12}C^{16}O$-$^{13}C^{16}O$, $^{12}CO_2$, $^{15}NO_2$, $^{18}O_2$-$^{16}O_2$.

2.5. Substances_2digit_initial. Отбираются токены, обозначающие различные вещества, названия которых начинаются на одну или две цифры. Примеры отбираемых токенов - 5-bromo-3-(N,N-diethylamino-ethoxy)-2-methylindole, 1-hexyne, 2,2,4-trimethypentane, 2.2-dimethyl-[1,3]dioxin-4-yl)-methanol..

2.6. catalysts. Отбираются токены, обозначающие различные каталитические системы, в названиях которых присутствует символ «.». Примеры отбираемых токенов - 1.5Au/C, 1.0CuCoK/ZrO_2, spherical $Ce_{0.9}Pr_{0.1}O_2$ particle, $Cu_{0.2}Co_{0.8}Fe_2O_4$, $Mg_3Zn_{3-x}Fe_{0.5}Al_{0.5}$, $LaFe_{0.7}Ni_{0.3}O_{3-\delta}$, $Ce_{0.8}Gd_{0.2}O_{2-\delta}$, $Mn_{0.8}Zr_{0.2}$.

2.7. comp. Отбираются токены, обозначающие состав катализаторов. Примеры отбираемых токенов - 1%Pd-1%Cu, 2%Pd-0.7%Re, 30%Cu-SiO_2-3%HZSM5-120, 0.4%Pt-0.1%Pd(0.4Pt-PdZWS-S), 0.3WT%Pt-0.2WT%Pd(0.3Pt-PdZWS-S).

2.8. Cryst_hydrates. Отбираются токены, обозначающие разнообразные кристаллогидраты. Примеры отбираемых токенов - $(NH_4)_6Mo_7O_{24}.4H_2O$, $FeCl_3.6H_2O$, $Na_2Si_{20.7}Al_{0.3}O_{29}.9H_2O$, $Ni_{0.364}Cu_{0,371}Fe_{0,264}(OH)_2(CO_3)_{0,132}.0.5H_2O$, $Ni_{0.39}Cu_{0.35}Cr_{0.26}(OH)_2(CO_3)_{0.13}.0.54H_2O$, $CrCl_3$-$6H_2O$, $Mg(Cl_2)X6H_2O$, $Ni_{0.37}Co_{0.38}Cr_{0.25}(OH)_2(CO_3)_{0.12}.0.82H_2O$, $MnCl_3.4H_2O$.

2.9. dimension. Используемые регулярные выражения: “(1|2|3)D” Отбираются токены, обозначающие 1-, 2- и 3-размерные методы или структуры. Примеры отбираемых токенов - 2D-SAXS, 1D-3D copper-oxide structure.

2.10. names. Отбираются токены, обозначающие имена собственные (имена и фамилии ученых), в состав которых входят символы, отфильтровываемые по правилу специальных символов. Примеры отбираемых токенов – Brønsted acid site, Brönsted acidity, Mössbauer spectroscopy, Béchamp process.

5.3 Шаблоны отбора L-грамм

В системе имеются пять шаблонов, содержащих следующие правила:

1. UnogrammPattern. Шаблон для отбора униграмм (L-грамм с L = 1). Содержит следующие правила (кроме общих правил):
   1.1. unigram.dictionary. Используется словарь, созданный на основе Corncob Lowercase, из которого были исключены слова, представляющие интерес для катализа (abrasion, absorption, absorptive, aerosol, adhesion и др.). Этот словарь насчитывает около 58 тысяч терминов, широко используемых в английском языке. Слова из этого словаря (например, copy, corpora, correction, correspondence и др.) исключаются из списка униграмм.
   1.2. unigram.pos. Срабатывает, если униграмма имеет один из следующих допустимых тегов - VBG,NN,NNPS,NNS. Отбираются токены, являющиеся существительными или герундием. Примеры отбираемых униграмм – hydrocalcite, acetylacetone, cracking, ageing, anchoring. Примеры отфильтрованных униграмм – suddenly, skeletal, behind, later.
   1.3. unigramm.additionalRules – список правил для расширеной фильтрации униграмм. В списке содержатся несколько правил, позволяющих отфильтровать униграммы, обозначающие различные ионы, обозначения, подписи к рисункам, например,

   F, AA, XT, D(C), GK(R), Ag0, S(C6), P-25, (A)(B), P1,P2, Pt,Pd, FEIII, Pd(I), Ba^{2+}, Ce(3+).
2. Bigrammpattern – шаблон для отбора биграмм (L-грамм с L = 2). Содержит следующие правила:
   2.1. BigrammPos
      2.1.1. Теги первого токена: JJ,JJR,FW,VBG,VBD,VBN,NN,NNP,NNPS,NNS. Первое слово в биграмме может быть существительным, прилагательным, герундием или причастием прошедшего времени, иностранным словом. Примеры отбираемых биграмм: первое слово является существительным - Andronov bifurcation, cobalt nitrate, Na_2CO_3 impregnation, nickel catalyst; первое слово является причастием - supported MgO, anchored lysine, stirred glass; герундием – homocoupling reaction, subliming $FeCl_3$; прилагательным - carbonaceous particle, temperature-programmed adsorption, Fischer-Tropsch catalyst; имеет тег FW - in-situ EXAF, UV-VIS spectroscopy, Raman spectroscopy.
      2.1.2. Теги второго токена: FW,VBG,NN,NNP,NNPS,NNS. Второе слово в биграмме может быть существительным, герундием или иностранным словом.
   2.2. bigrammMorphTagException

2.2.1. Список исключений: «VBG,VBG», «VBG,FW», «NNP,FW». Отбираются ошибочные нетерминоподобные биграммы, вырванные из контекста, имеющие неподходящие для биграмм сочетания тегов. Примеры отбираемых биграмм - involving reforming, reforming minimizing, imploung in-situ, using in, Shimada et, Baddeley et.

3. ManyGramPattern – шаблон для отбора L-грамм с L > 2. Содержит следующие правила:
   3.1. ManygrammPos
      3.1.1. Теги первого токена: NN,NNP,VBG,VBD,VBN,JJ,JJR,RB,RBS,FW. Первое слово в L-граммах с L>2 может быть существительным, герундием, причастием, прилагательным, наречием и иностранным словом. Примеры отбираемых L-грамм (L>2) : первое слово является существительным - X-ray fluorescence spectrometer, Brønsted basic site, lauric acid conversion, Pd(110) surface oscillation; герундием - doping CsPW with platinum, modifying HZSM5 zeolite with metal; причастием - supported manganese-cerium oxide, supported proton-conductive membrane, catalyzed N_2O decomposition, programmed CO oxidation; прилагательным - crystalline phase transition, characteristic diffusion pathlength; наречием - heterogeneously catalyzed liquid phase oxidation, catalytically active cell; FW - infra red emission spectroscopy, in-situ UV/VIS reflectance.
      3.1.2. Теги срединного токена: NN,NNP,VBG,VBD,VBN,JJ,JJR,RB,RBS,FW,IN,DT. Срединное слово в длинных L-граммах может помимо тегов для первого слова быть предлогом и артиклем. Примеры отбираемых L-грамм: catalyzed oxidation of NO, complete photoreduction of Pd(II), oxidative coupling of methane.
      3.1.3. Теги последнего токена: VBG,NN,NNP,NNPS,NNS. Последнее слово в длинных L-граммах может быть существительным или герундием.
   3.2. FirstSecondWord
      3.2.1. Список тегов-исключений: «VBG -> NN, IN», «VBN -> NN, JJ». Отбираются L-граммы, начинающиеся с герундия или причастия прошедшего времени, обозначающие явления, системы или процессы, отфильтровываются нетерминоподобные L-граммы, вырванные из контекста. Примеры отбираемых L-грамм – начинающиеся с герундия, за которым идет существительное или предлог: promoting effect, propagating thermosynthesis, reforming of the biomass, drying inside the microscope column, drying of the catalyst bed; начинающиеся с причастия, за которым следует существительное

или прилагательное: activated carbon, oxidized platinum site, fixed bed, ordered micro-mesoporous Al_2O_3, supported ionic liquid catalyst.

Примеры удаляемых L-грамм – used during steam reforming, catalyzed by metalloporphyrin, investigated by XRD, using atomic absorption, yielding mainly saturated hydrocarbon.

4. ChemicalUnogrammPattern – Шаблон для отбора одногramм, помеченных тегами OSCAR.
   4.1. oscarAndPos. Срабатывает, если токен имеет любой тег OSCAR, а также следующие морфологические теги: FW,NNP. Отбираются одногramмы, являющиеся именами собственными, распознанные OSCAR как имеющие химический смысл. Примеры отбираемых одногramм – XANES, TEOM, HMS, GSA. Примеры удаляемых одногramм – Exner, Languedoc, Gomes.

6 Программная реализация

Программная реализация системы состоит из серверной и клиентской части. Серверная часть выполняет преобразование материалов конференций, терминологический анализ, генерацию отчетов. Клиентская часть предоставляет пользователю интерфейс для доступа к результатам терминологического анализа, просмотру и выполнению отчетов.

6.1 Серверная часть

Программная реализация серверной части выполнена на платформе Java SE 6 [24]. Для управления и контроля над исходным кодом системы используется vcs git, для удаленной работы с git репозиторием используется решение на базе gitlab.

В качестве системы сборки проекта и управления зависимостями используется maven [25]. Данная система сборки позволяет разделить проект на несколько модулей, декларативно управлять конфигурацией модулей, описывать зависимости между модулями, а также зависимости модулей от сторонних библиотек. На рисунке 5 представлена диаграмма модулей maven для данного проекта.

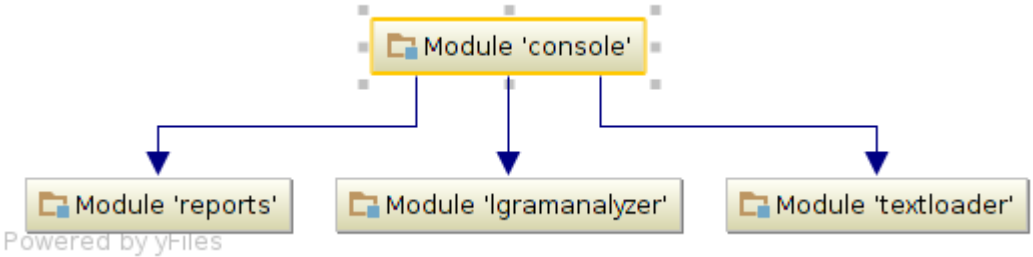

Рисунок 5. Диаграмма maven-модулей проекта.

Всего в системе имеется четыре модуля. Модуль textloader ответственен за разбор и загрузку в БД материалов конференций в pdf-формате. lgramanalyzer предназначен для выполнения терминологического анализа текстов. Ответственность модуля Reports состоит в генерации отчетов на основе полученного L-граммного спектра. Console предназначен для проведения экспериментов – запуск, выполнение и сохранение результатов эксперимента.

Для управления конфигурацией приложения используется принцип Inversion of control (IOC) [26]. Суть данного подхода состоит в том, что зависимости между классами задаются не напрямую в коде приложения, а описываются во внешнем источнике. Управляя описанием зависимостей в этом источнике, можно менять структуру и состав приложения, не изменяя его исходный код. В качестве реализации данного подхода используется компонент Spring DI фреймворка Spring [27].

Рассмотрим более детально структуру основного модуля системы – lgramanalyzer. Задача, которую решает модуль, состоит в терминологическом анализе текстов. На рисунке 6 показана class-диаграмма классов, ответственных за терминологический анализ.

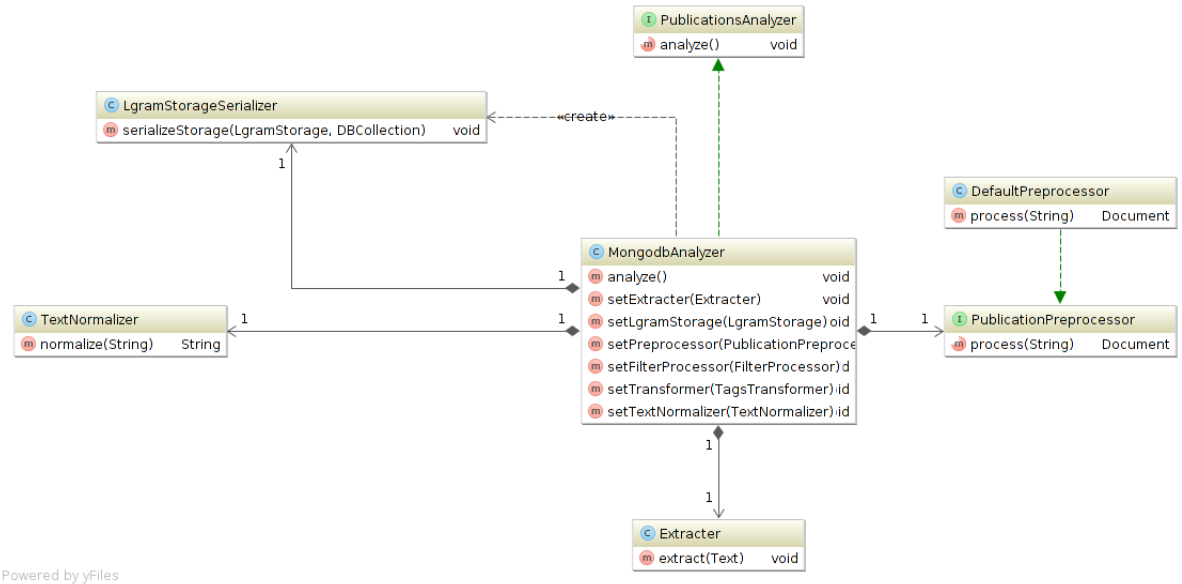

Рисунок 6. Основные классы модуля Lgramanalyzer.

Основной интерфейс модуля – PublicationAnalyzer, содержащий метод analyze, выполняющий анализ текстов. Основную реализацию представляет класс MongoDbAnalyzer, работающей с БД mongodb. Данный класс загружает тексты публикаций, выполняет терминологический анализ, и сохраняет полученный L-граммный спектр в БД. Для выполнения анализа используется зависимые классы: TextNormalizer для предварительной нормализации текста, PublicationPreprocessor для выполнения разметки текста, PublicationAnalyzer для L-граммного анализа и LgramStorageSerizlizer для сохранения результатов L-граммного анализа.

В процессе разметки текстов используется несколько сторонних библиотек (Stanford CoreNLP, OSCAR4), обладающих различными интерфейсами, и по-разному представляющих размеченные данные. Для решения проблемы объединения таких библиотек разметки была разработана архитектура, позволяющая абстрагироваться от конкретной библиотеки разметки. В основе решения лежит представления всех структур данных, относящихся к тексту (документ, предложение, токен) в виде Map с ограниченным набором ключей (тегов). Диаграмма классов данных структур представлена на рисунке 7.

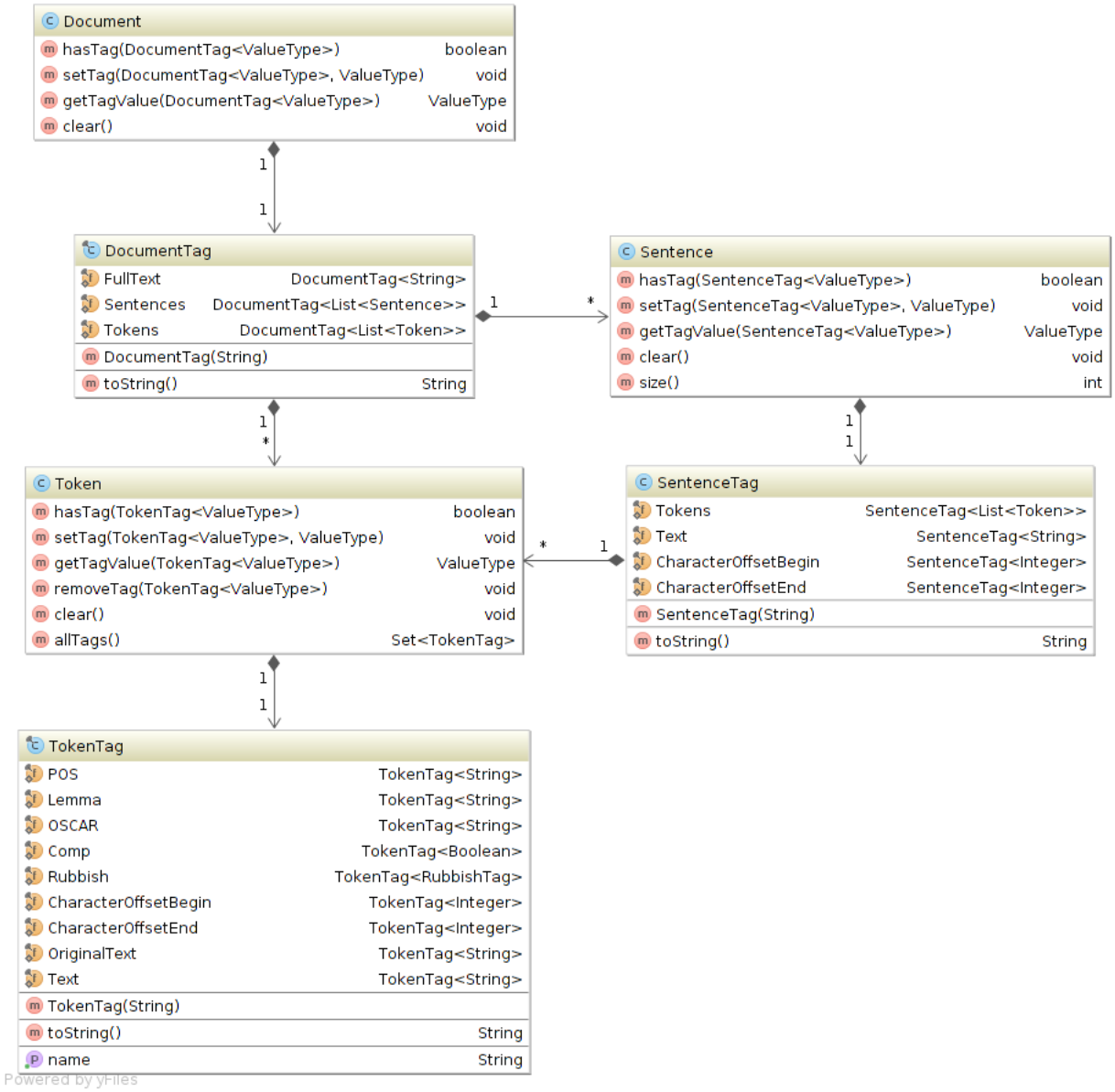

Рисунок 7. Диаграмма классов документа.

Всего имеется три основных структуры – Document, Sentence, Token. Каждая структура представляет из себя Map, в котором ключами являются объекты соответствующих классов (DocumentTag для Document, SentenctTag для Sentence, TokenTag для Token). Данные классы ключей являются обобщенными классами с параметрами, определяющими значение, которое может быть передано по ключу. Таким образом обеспечивается строгий контроль типов для возможных ключей и значений.

Для использования данных структур совместно с различными сторонними библиотеками разметки используется шаблон проектирования «адаптер». Задача адаптера состоит в преобразовании имеющихся данных в формат, пригодный для обработки внешней библиотеки, и последующем обратном преобразовании результата обработки в имеющуюся структуру данных. Классы основных адаптеров представлены на рисунке 8.

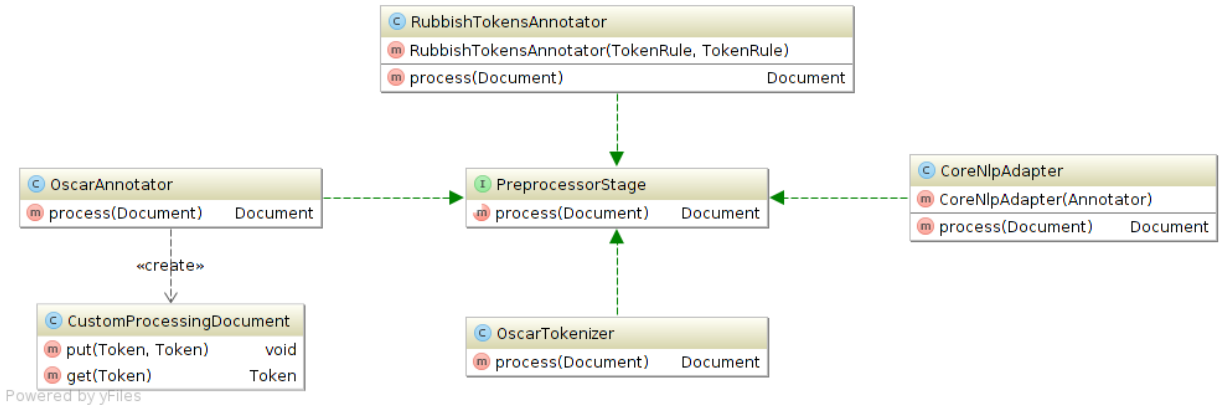

Рисунок 8. Класс-диаграмма адаптеров.

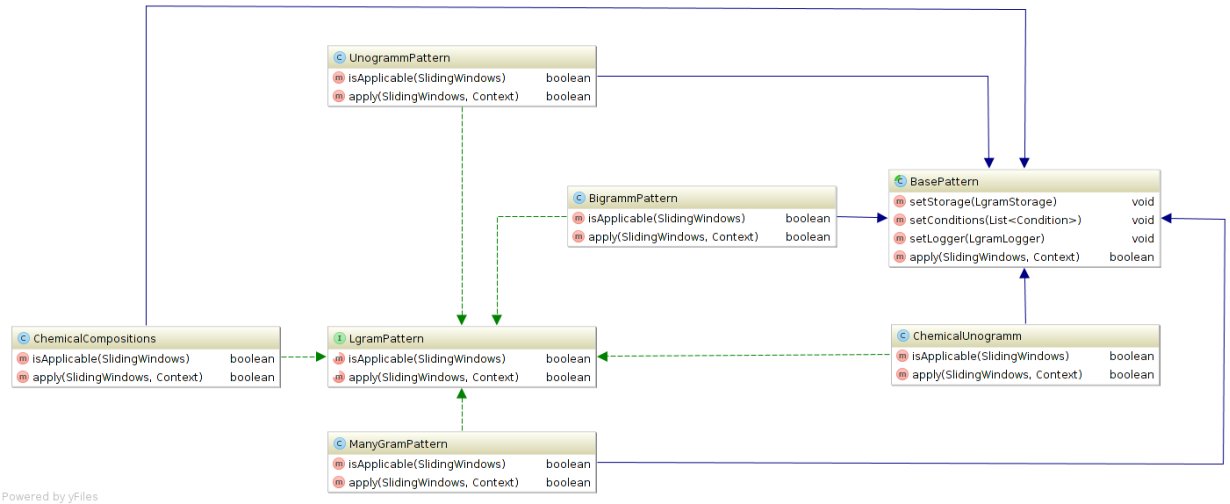

Рисунок 9. Диаграмма основных классов L-граммного анализа.

L-граммный анализ осуществляется с помощью интерфейса LgrammPattern и его реализации BasePattern. Реализация данного класса состоит в применении к исходной L-грамме списка правил. Если хотя бы одно правило сработало, то L-грамма помечается как отфильтрованная, в противном случае делается пометка о терминоподобности L-граммы. Описание объектов интерфейса LgramPattern и правил, составляющих конкретный шаблон, задается в spring-конфигурации. На рисунке 10 представлены основные правила отбора терминоподобных L-грамм.

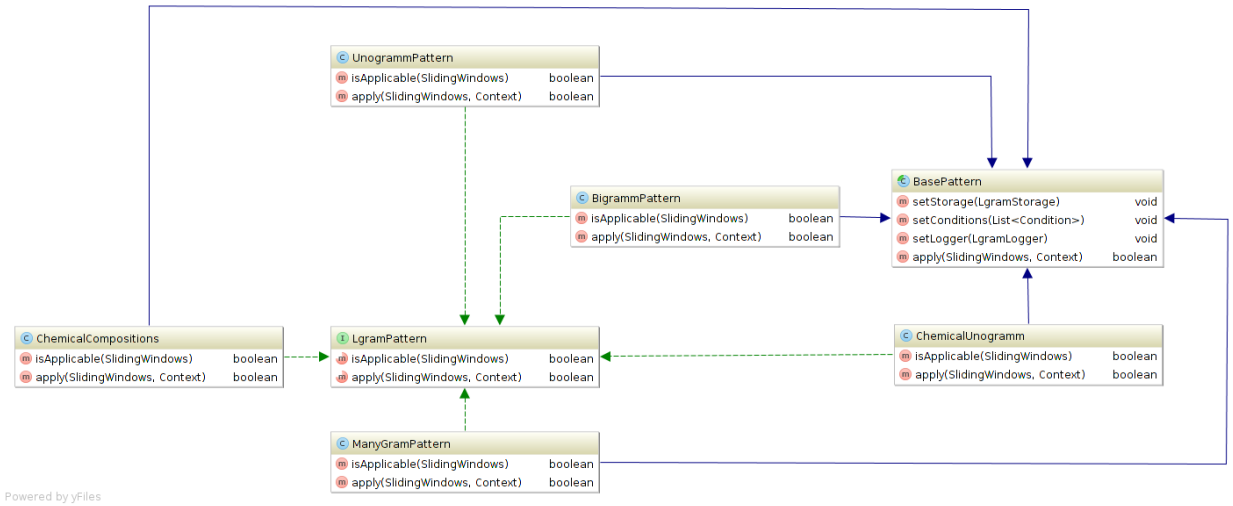

Рисунок 10. Диаграмма основных классов правил L-граммных шаблонов.

6.2 Клиентская часть

Клиентская часть приложения реализована на языке PHP. В качестве основы приложения используется микрофреймфорк silex [28], содержащий в себе основные средства для поддержки веб-приложений – модули для работы с базой данных, шаблонами, локализацией, обработки http-запросов и т.д. В качестве средства управления javascript-зависимостями используется bower. Для удобного описания css-стилей используется формат sass, который преобразуется в обычный css при помощи compass.

Для удобства создания новых и поддержки существующих отчетов был разработан подход, позволяющий декларативно описывать отчеты. По такому описанию затем автоматически генерируются все необходимые компоненты, такие, как формы поиска и фильтрации, таблица с данными отчета. Декларация отчета состоит из трех классов, описывающих характеристики фильтруемых полей (FilterFields), полей сортировки (SortFields) и полей, которые отображаются в таблице с данными отчета (ReportFields). Каждое поле содержит следующие характеристики:

а) Translation – ключ для перевода названия поля

б) FieldPath – путь к полю в БД. Если поле является вложенным, то компоненты пути разделяются точкой, например, “context.publicationYear”.

в) Template – используется для указания шаблона отображения значения поля в таблице с данными отчета. Использование шаблонов необходимо в случае, когда значение поля представляет собой сложную конструкцию, например, массив или объект, и необходимо вывести данный объект в виде скалярного значения.

г) DefaultValue – задает значение по умолчанию, которое будет отображено в таблице с данными отчета в случае, когда значение отсутствует в текущей записи.

д) Filter – описывает характеристики поля, относящиеся к фильтрации. Основная характеристика – тип поля, которая определяет выражение для фильтрации данных по данному полю. Доступные значение – integer, double, string, regexp, mongoid.

6.3 Описание базы данных

Для хранения данных, генерируемых в процессе работы системы используется нереляционная документноориентированная СУБД mongodb [29]. Выбор документноориентированной субд обусловлен структурой хранимых данных – описание L-граммы содержит объекты уровня вложенности до пяти. При этом для обработки L-граммы необходимо загрузить все данные о L-грамме целиком, что в случае использования реляционной БД потребовало бы пять join выражений.

Всего в базе данных имеется три коллекции – texts, tokens, lgramms. В коллекции Texts находятся анализируемые тексты. Документы данной коллекции имеют следующий набор полей:

а) _id – идентификатор записи;
б) year – год выхода публикации;
в) filePath – путь к исходному файлу публикации;
г) collectionId – идентификатор коллекции, в которую входит текст. Для конференций значением данного поля обычно является название конференции.
д) authors – массив авторов публикации;
е) organizations – массив организаций, указанных в публикации;
ж) references – библиографические ссылки публикации;
з) title – название публикации;
и) content – основной текст публикации.

Документ коллекции tokens описывает текст, разбитый на токены, а также структуру токенов. Элементы коллекции имеют следующие поля:

1. _id – идентификатор записи
2. textId – идентификатор текста
3. sentences – массив предложений текста. Каждое предложение является массивом токенов.

   3.1. token

   3.1.1. originalText – оригинальная строка токена (в точности так, как написание токена в исходном документе)

   3.1.2. text – нормализованная строка токена. Применяемые нормализации приведены в главе с описанием модели терминологического анализа.

3.1.3. characterOffsetBegin – количество символов от начала текста до первого символа токена.

3.1.4. characterOffsetEnd – количество символов от начала текста до последнего символа токена

3.1.5. pos – часть речи токена

3.1.6. lemma – лемматизированный токен

3.1.7. rubbish – поле, показывающее, был ли токен помечен в ходе жесткой фильтрации.

Элемент коллекции lgramms описывают конкретную L-грамму в совокупности со всеми ее вхождениями во все анализируемые тексты. Структура документа следующая:

1. _id – идентификатор L-граммы. Идентификатором является строка, состоящая из токенов L-граммы, приведенных к верхнему регистру, например, «PLASMA MODIFICATION OF CATALYST».
2. normalizedName – содержит нормализованную L-грамму. Нормализованной является L-грамма, в которой удалены токены-союзы, и оставшиеся токены находятся в лексикографическом порядке. Например, для L-граммы «vanadium ion on the catalyst surface» нормальной формой является «catalyst ion on surface vanadium».
3. occurences – массив, элементы которого описывают одно вхождение L-граммы в текст.
   3.1. context – объект, описывающий контекст вхождения L-граммы в текст.
      3.1.1. publicationId – идентификатор текста, в который входит L-грамма.
      3.1.2. publicationYear – год выхода публикации, в которую входит L-грамма. Данное поле необходимо для построения различный отчетов и вычисления временных характеристик L-грамм без обращения к сторонним коллекциям.
   3.2. lgram – объект, описывающий характеристики L-граммы применительно к данному конкретному вхождению. Характеристики L-граммы могут меняться в зависимости от конкретного вхождения, например, морфологические теги могут быть различными из-за ошибок морфологического анализатора. Поэтому в каждом вхождении L-граммы хранится полная информация о этой L-грамме.00
      3.2.1. tokens – список токенов L-граммы. Поля каждого токена такие же, как в коллекции tokens.
      3.2.2. tags - список тегов L-граммы для данного вхождения. Теги позволяют сохранить дополнительную информацию о вхождении L-граммы, например, название шаблона, который пометил L-грамму как терминоподобную. Теги хранятся как обычные javascript объекты и могут иметь любую структуру.

7 Пользовательский интерфейс

Пользовательский интерфейс предназначен для взаимодействия пользователя – эксперта в предметной области с системой. При помощи пользовательского интерфейса пользователю доступны следующие операции:

а) просмотр терминологической разметки;

б) просмотр исходных текстов;

в) просмотр результата отчетов.

Просмотр терминологической разметки необходим для того, чтобы оценить корректность работы терминологического анализа, просмотреть терминологическую структуру текста, скорректировать словари и шаблоны алгоритма. Окно с терминологической разметкой показано на рисунке 11.

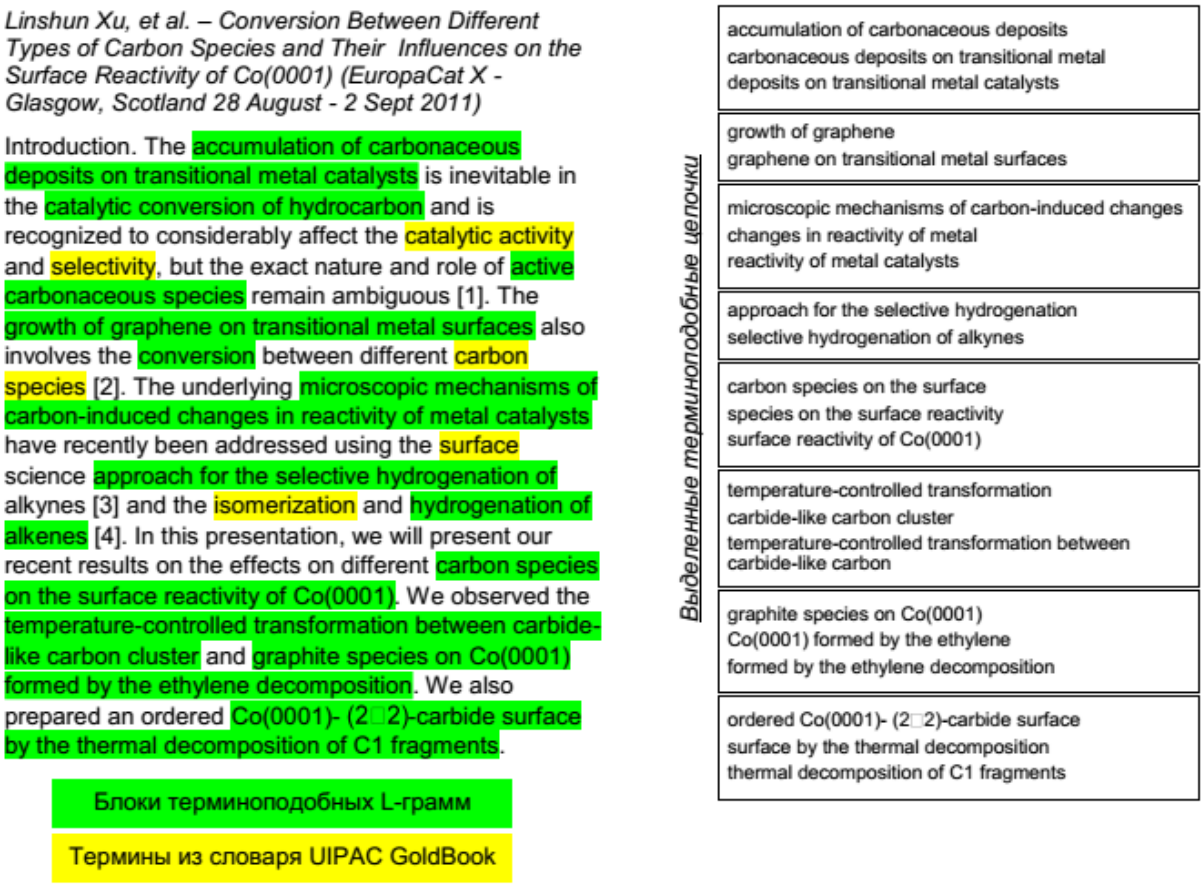

Рисунок 11. Пример терминологической разметки.

Функция просмотра текстов позволяет оценить корректность преобразования pdf-файлов публикаций, найти нужную фразу в тексте. Форма поиска содержит поля для поиска текстов по году публикации, названию файла полного текста, содержанию. Также пользователь может указать произвольную javascript-функцию для фильтрации текстов.

Основной задачей пользовательского интерфейса является предоставление пользователю возможности работы с отчетами. Отчетом является прямоугольная таблица, в которой первый столбец содержит L-грамму, а остальные столбцы представляют различные характеристики данной L-граммы, такие, как частота встречаемости, среднеквадратичное отклонение и т.д. Имеется функция экспорта результатов отчета в excel для последующего детального анализа.

У каждого отчета имеется форма поиска и фильтрации, с помощью которых пользователь выполняет анализ отчета. На рисунке 12 представлена форма поиска и фильтрации для отчета по трендам.

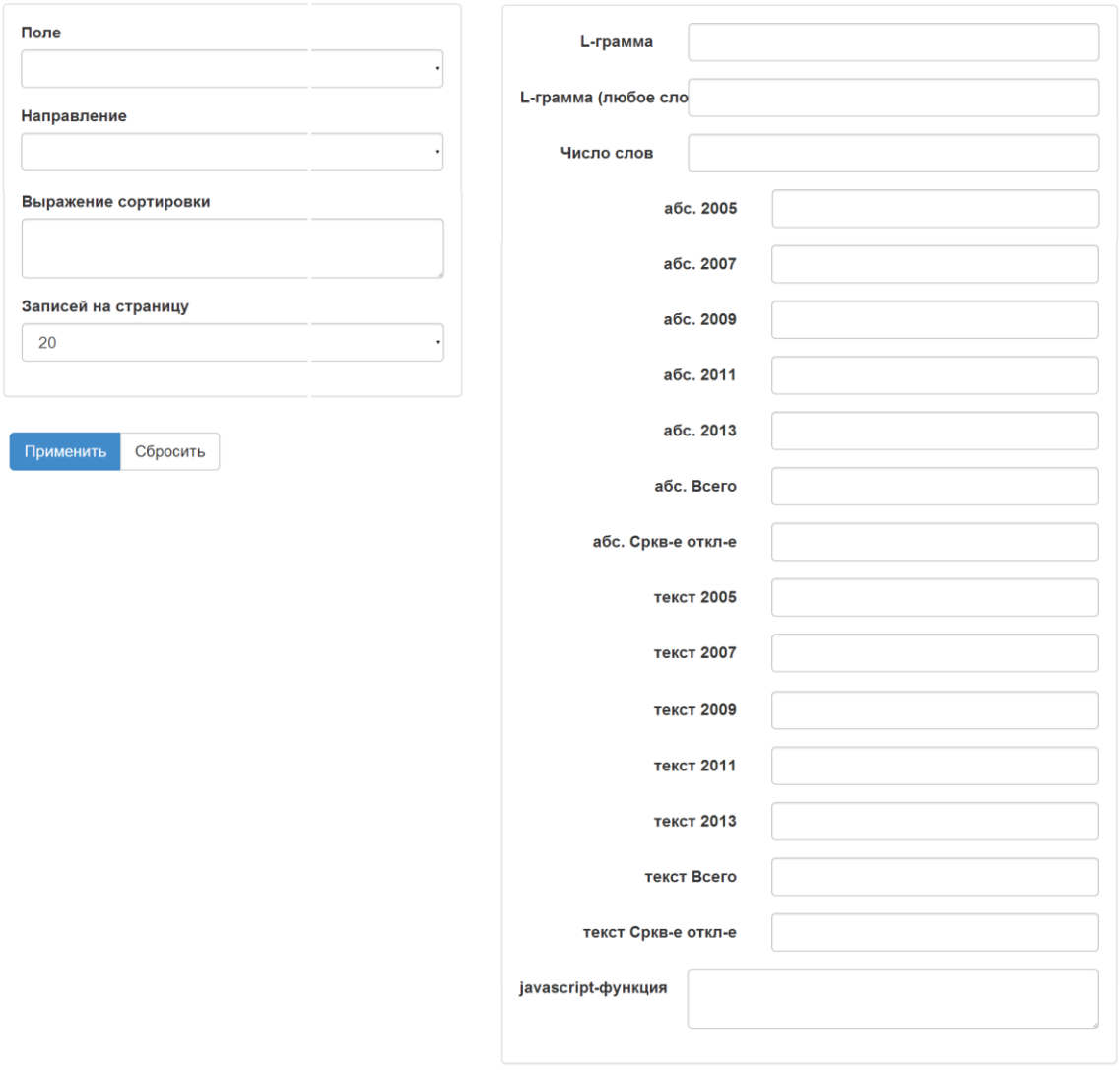

Рисунок 12. Форма поиска и фильтрации отчета.

Форма состоит из двух частей: правая часть предназначена для фильтрации результатов отчета, левая часть – для сортировки отчета. Для того, чтобы произвести сортировку, необходимо выбрать поле и направление сортировки. Поле сортировки определяет, ко какому полу из списка полей отчета будет произведена сортировка,

направление задает порядок сортировки – по возрастанию значений поля или по их убыванию.

Фильтрация позволяет отобрать необходимые L-граммы отчета при помощи поисковых полей. Каждое поисковое поле позволяет задать необходимое значение для отбора нужных L-грамм. При поиске по числовым полям допустимо задание диапазона значений, например, выражение «1-5, 7-» отфильтрует L-граммы со значением числового поля от 1 до 5, а также больше 7.

Наиболее гибким механизмом фильтрации является произвольная javascript-функция, которую можно задать в специальном поле формы фильтрации. Данная функция может обращаться к полям объекта текущей записи для определения того, удовлетворяет ли запись условиям отбора, или нет. С помощью механизма javascript-функций можно выполнять отбор L-грамм по критерию любой сложности.

Для отбора L-грамм полезно в качестве характеристики L-граммы кроме распределений абсолютных и текстовых частот встречаемости по годам иметь также величины, показывающие различные характеристики данных распределений. В качестве такой величины используется величина среднеквадратичного отклонения, которая вычисляется стандартным образом:

$$\sigma = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2},$$

где σ – среднеквадратичное отклонение

x_i-значение частоты встречаемости (абсолютной или текстовой) за год i

$\bar{x}$ –среднее значение частоты встречаемости.

Всего в пользовательском интерфейсе представлены следующие отчеты:

а) L-граммы и их общие характеристики. Каждая запись отчета содержит L-грамму, суммарные абсолютную и текстовую частоты встречаемости за все годы.

б) Частотное распределение. Запись отчета содержит распределение абсолютной и текстовой частот встречаемости по годам, а также среднеквадратичное отклонение данных частот.

в) L-граммы с тегом OSCAR. Модификация отчета о частотном распределении, в котором имеются только L-граммы, в которых хотя бы один токен помечен тегом OSCAR.

г) Аномалии POS. Модификация отчета о частотном распределении, в котором присутствуют только L-граммы с тегами части речи, различными в зависимости от конкретного вхождения L-граммы в текст. Например, L-грамма «Ru loading level»

имеет два варианта морфологических тегов: «NN,VBG,NN» и «NN,NN,NN» или «polyramnogalacturonan-copper complex» - «NN,NN» и «JJ,NN». Данный отчет фактически позволяет определить ошибки морфологического анализатора.

д) Аномалии OSCAR. Аналогичен предыдущему отчету, но вместо различных морфологических тегов присутствуют только L-граммы с различными тегами OSCAR. Например, L-грамма «Cr_2O_3 catalyst» имеет два варианта тегов OSCAR – [CM,ONT] и [CM,].

е) Отчеты о положительных трендах. Положительным трендом считается L-грамма, текстовая частота встречаемости которой монотонно не убывает в течении некоторого периода времени. В настоящее время используется два отчета данного типа – с одним строгим переходом и с двумя. Под строгим переходом понимается переход между двумя соседними значениями частоты встречаемости, значение частоты которого во второй точке строго больше, чем в первой точке. Например, переход 17 -> 20 является строгим, переходы 17 -> 17 и 17 -> 15 не являются строгими.

ж) Отчет об общенаучных терминах содержит L-граммы, помеченные общенаучным фильтром на последнем этапе терминологического анализа.

8 Некоторые практические результаты использования системы терминологического анализа

Основным критерием успешности разрабатываемого подхода, разумеется, является проверка его работоспособности на реальном практическом материале. В настоящей работе разработанные методики были использованы для предварительного анализа периодической, возможно, самой широкой по охвату научной конференции по химическому катализу – «Европейский Конгресс по Катализу – EuropaCat» за 2013, 2011, 2009, 2007, 2005 годы (5 конгрессов за 10 лет, всего порядка 6 тыс. документов).

Проведены оценки точности и полноты анализа, производимого терминологической системой, а также получены предварительные результаты анализа динамики изменения терминологической базы со временем. Даная работа проводилась при участии сотрудников Института катализа СО РАН – к.х.н. Ильиной Л.Ю. и к.х.н. Кузьмина А.О.

8.1 Оценки точности и полноты

Для оценки эффективности использованного метода выделения терминоподобных L-грамм необходимо количественно ответить на вопросы о полноте и точности извлечения системой терминоподобных словосочетаний (методика расчёта данных понятий приведена в литературном обзоре).

Для ответа на эти вопросы, специалистами – экспертами был осуществлен поиск терминоподобных словосочетаний, а также общенаучных терминов и названий химических соединений в ручном режиме. Использовались результаты независимого экспертного анализа пяти тезисов, выбранных случайным образом. Для включения словосочетания в список терминоподобных требовалось его выделение обоими экспертами (то же самое касалось и выделения общенаучных терминов и т.д.) Данные, полученные в ходе данной «ручной» экспертной терминологической разметки сравнивались с данными, получаемыми автоматически. Рассчитывалась точность и полнота анализа, для расчёта объединённой характеристики использовалась F-мера. Результаты представлены в таблице 8.

Точность позволяет оценить, насколько множество терминоподобных словосочетаний, найденных автоматически, совпадает с экспертным множеством. Полнота указывает, насколько полно автоматически полученное множество охватывает экспертное множество терминов. F-мера является взвешенным гармоническом средним точности и полноты.

Для анализа использовались следующие тексты:

- № 1 – Design, synthesis and catalysis of recoverable catalysts assembled in emulsion and its application in deep desulphurization of fuel oil, C. Li et al. (2005);
- № 2 – Understanding Reaction Pathways on Model Catalyst Surfaces, F. Gao et al. (2007);
- № 3 – Solid Acid Catalysts Based on H3PW12O40 Heteropoly Acid: Acid and Catalytic Properties at a Gas-Solid Interface, A.M. Alsalme et al. (2011);
- № 4 – Advantages of using TOF-SIMS method in surface studies of heterogeneous catalysts, M. I. Szynkowska et a. (2005);
- № 5 – ECS-Materials: Synthesis and Characterization of a New Class of Crystalline Organic-Inorganic Hybrid Alumino-Silicates, G. Bellussi et al. (2007).

Таблица 8. Точность и полнота нахождения терминоподобных словосочетаний.

Текст	**Кол-во экспертных терминов**	**Кол-во автомат. терминов**	**Кол-во пересечений**	**Точность, %**	**Полнота, %**	**F-мера, %**
№ 1	164	221	135	61	82	70
№ 2	155	174	96	55	62	58
№ 3	151	150	96	64	63	64
№ 4	68	119	40	34	59	42
№ 5	125	215	106	50	84	62
среднее				53	70	59

Из данной таблицы видно, что уже первая версия системы терминологического анализа обеспечивает точность и полноту поиска терминов на достаточно высоком уровне. В качестве примера приведём ещё раз данные из главы «Литературный обзор» для достигаемого в других системах значения F-меры: Sztergak - 22÷23%, Wingnus -22÷24 %.

8.2 Анализ трендов

Среди множества аналитических задач, которые можно исследовать с использованием терминологического спектра текстовой коллекции, наиболее очевидной является анализ динамики изменения употребления термина со временем.

Был проведен предварительный анализ динамики изменения терминоподобных L-грамм во времени для тезисов конференций «EuropaCat». Для этого в спектре полученных L-грамм выявлялись случаи возрастания текстовой и абсолютных частот со временем. Часть полученных результатов (в последовательном порядке выдачи) приведена в таблице 9. В таблице объединены значения для синонимов термина (альтернативные написания приведены в скобках).

Таблица 9. Часть обнаруженных L-грамм с положительной временной динамикой текстовой (и абсолютной) частоты.

№	L-грамма	2005 г.	2007 г.	2009 г.	2011 г.	2013 г.
1	nanoparticles	78 (203)	146 (451)	158 (494)	170 (543)	350 (1525)
2	nanorod (nanorods, nano-rod)	0 (0)	0 (0)	3 (26)	8 (30)	21 (89)
3	platform molecule	0 (0)	0 (0)	1 (2)	5(5)	14(17)
4	lignocellulosic biomass	1 (1)	1 (1)	3 (3)	7 (7)	19 (23)
5	lignocellulose	0 (0)	1 (2)	3 (3)	3 (4)	17 (29)
6	hydrolysis of cellulose	0 (0)	0 (0)	3 (3)	3 (9)	10 (25)
7	renewable feedstock	2 (3)	3 (5)	7 (8)	11 (12)	15 (16)
8	biogas (bioga)	1 (2)	1 (1)	6 (12)	7 (17)	19 (65)
9	biooil (bio-oils, bio-oil, bio oil)	6 (29)	11 (33)	6 (17)	14 (38)	23 (90)
10	lignin	0 (0)	8 (20)	5 (16)	9 (48)	27 (177)
11	humin	0	0	0	0	4 (60)
12	HMF (hydroxymethylfurfural, hydroxymethyl furfural)	1 (9)	1 (2)	11 (79)	15 (84)	21 (138)
13	glycerol oxidation (oxidation of glycerol)	1 (2)	2 (2)	1 (1)	2 (4)	13 (50)
14	oxidative coupling of methane (methane oxidative coupling)	2 (2)	4 (5)	4 (5)	8 (9)	12 (30)
15	OCMOL	0	0	0	2 (3)	4 (10)
16	dry reforming	6 (13)	13 (32)	16 (27)	23 (33)	31 (68)
17	DRM (methane dry reforming, CH_4 dry reforming, dry reforming of CH_4, dry reforming of methane)	4 (7)	14 (28)	11 (25)	23 (75)	27 (85)

Как и следовало ожидать, огромное число публикаций посвящено исследованию и использованию наночастиц в катализе, и их число неуклонно возрастает. Текстовая частота термина *nanoparticles* с 78 (2005 год) возросла в 4.5 раза и составила 350 текстов в 2013 году.

Интересно, что термин *nanorod* (наностержень и катализаторы с использованием таких материалов) до 2009 года не встречался в нашей подборке тезисов (и даже не был известен нашей группе на момент проведения терминологического анализа), после чего стал все активнее использоваться для обозначения материалов, все характерные размеры которых составляют от 1 до 100 нм, при этом отношение длины к ширине составляет от 3

до 5. Например, в 2009 году термин *nanorod* упоминается всего в двух текстах анализируемой конференции, тогда как в 2013 году его используют уже в 17 работах.

Это же касается и недавно появившегося термина *platform molecule* (текстовая частота за 2009 для этого термина равна 1, тогда как текстовая частота за 2013 составляет 14). Термин *platform molecule* в подборке тезисов появился в 2009 году и все более широко используется для обозначения получаемых из биомассы базовых молекул с множественными функциональными группами, которые потенциально могут быть превращены в различные, имеющие широкое применение (например, компоненты дизельного топлива) молекулы. Это, например, глицерин (*glycerol*), 5-гидроксиметилфурфурол (*HMF*), уксусная, лауриновая, янтарная кислоты и т.д.

Действительно, термины из диапазона 3-13 непосредственно связаны с процессами переработки биомассы, что является известным трендом последних лет, так как последние десятилетия отмечены бурным развитием технологий производства альтернативных видов топлива. Микробиологическая конверсия возобновляемых ресурсов биосферы с целью получения полезных продуктов, в том числе биотоплива, в настоящее время является актуальной технологической проблемой.

В настоящее время наиболее перспективным источником биомассы является массовое сырье второго поколения (непищевые остатки культивируемых растений, травы и отходы древесины), в частности *лигноцеллюлоза* (*lignocellulose*). Так называют несъедобный растительный материал, состоящий в основном из целлюлозы (*cellulose*) (40-50%), гемицеллюлозы (25-35%) и лигнина (*lignin*) (15-30%). Именно это сырьё считают будущим биотопливной промышленности.

Возрастающий интерес к химии фурановых соединений также обусловлен широкой перспективой использования возобновляемого растительного сырья для синтеза химических продуктов, в том числе таких полимерных материалов, как полиэфиры, полиамиды и полиуретаны. Фурфурол и 5-гидроксиметилфурфурол являются наиболее подходящими мономерами для синтеза указанных полимеров. И в то время как промышленное производство фурфурола составляет около 200 тыс. т в год, технология получения 5-гидроксиметилфурфурола (*HMF*) стала отрабатываться на пилотном уровне, что видно из тренда данного термина. Глицерин (*glycerol*) также является промежуточным продуктом производства биодизельного топлива, поэтому всё большее количество работ посвящено исследованию процесса окисление глицерина (*glycerol oxidation*).

Интересно также, что в последнее время растет интерес к разработке способов утилизации гумина (*humin*) – побочного продукта биологической конверсии

лигноцеллюлозы (*lignocellulose*). Так, в работах [30, 31] рассматриваются вопросы растворения и каталитической газификации гуминов как перспективных способов его утилизации.

Очень перспективным типом биотоплива является биогаз (*biogas*). Он вырабатывается в ходе водородного или метанового брожения биомассы под воздействием специальных бактерий и представляет собой смесь метана и углекислого газа. В 2012 г. в России начат проект по строительству 50 биогазовых электростанций в 27 регионах. По оценкам экспертов, потенциальное производство в России биогаза может составить до 72 млрд кубометров в год [32]. В настоящее время, однако, большая часть этого топлива выбрасывается и не используется для получения энергии, потому что высокое содержание CO_2 в биогазе уменьшает температуру нагрева и стабильность пламени газовой смеси, что приводит к увеличению выбросов при сжигании.

Из динамики изменения терминологической базы со временем отчётливо видно, что процесс сухого риформинга (*methane dry reforming*, *dry reforming*), который позволяет преобразовывать недорогой биогаз в синтез-газ (сырье для производства жидкого топлива или ценных химических продуктов), вызывает в мире всё больший интерес.

Таким образом, число исследований, связанных с биотопливом и переработкой биомассы, неуклонно растет, что нашло свое отражение и в нашем анализе зависимости частоты встречаемости терминоподобных словосочетаний от времени.

Была изучена информация о том, в каких текстах встречаются трендовые термины и найдено пересечение текстов их содержащих. В качестве примера на рисунке 13 приведены диаграммы Венна для некоторых пересечений (количество общих текстов) из трех терминов. Так, например, *HMF* и *platform molecule* встречаются вместе в 4-х документах, что составляет 20 % от общего числа текстов, в которых встречается термин *platform molecule* и 8,2 % от общего числа текстов с термином *HMF*. Пересечение *HMF* и *renewable feedstock* равняется 4, что составляет 8,3 % от всех текстов с *HMF* и 10,5 % от всех текстов, содержащих *renewable feedstock*. Общая картина всех пересечений терминов из таблицы 9 изображена на графе (рисунок 14), из которого видно большое количество пересечений текстов, содержащих термины из области переработки биомассы.

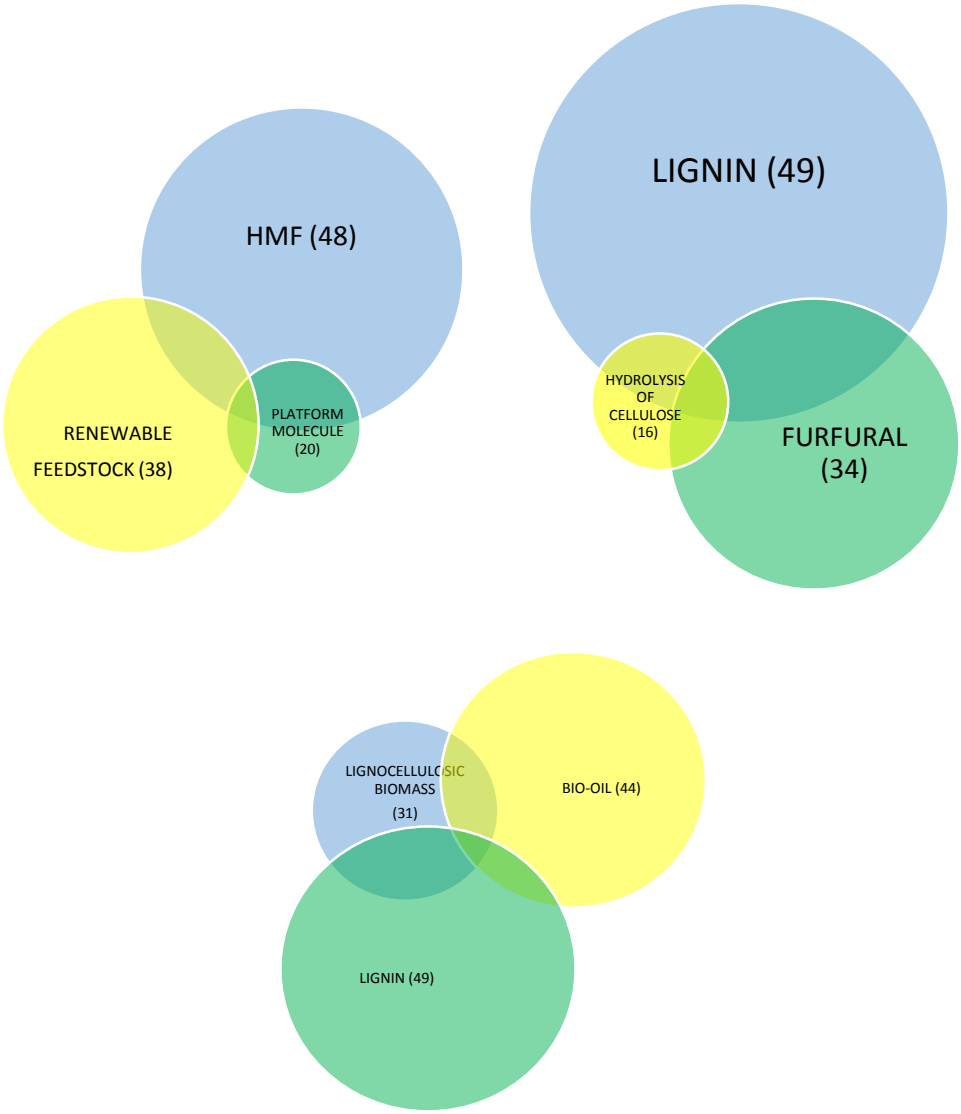

Рисунок 13. Диаграммы Венна некоторых пересечений текстов, содержащих термины из таблицы 9.

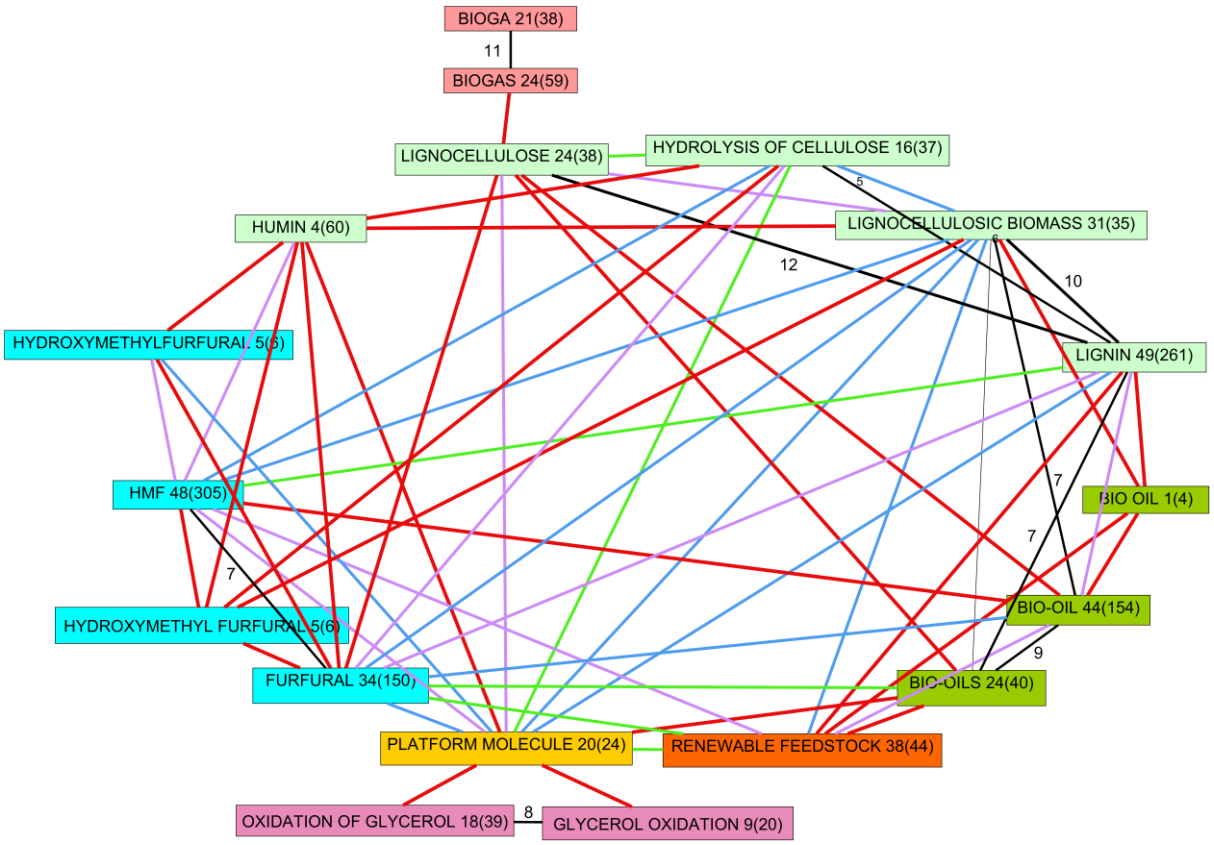

Рисунок 14. Граф пересечений текстов, содержащих трендовые термины. У термина указана текстовая (абсолютная) частоты встречаемости. Цвет линии обозначает количество общих текстов: красный – 1; зелёный– 2; синий – 3; сиреневый – 4; чёрный – более 4.

Одним из удивительных наблюдений является тот факт, что в последние годы появляется все больше работ, посвященных процессу окислительной конденсации метана *oxidative coupling of methane* и т.д. Текстовая частота термина *oxidative coupling of methane* с 1 в 2005 году возросла до 12 в 2013 году. Это показалось удивительным, поэтому мы попытались проанализировать причины этого явления.

Известно, что об осуществлении реакции окислительной конденсации метана (ОКМ) впервые сообщил Митчелл в 1980 году. Реакция ОКМ оказалась настолько привлекательной, что после первой публикации произошел настоящий бум публикаций по этой теме. Достаточно сказать, что в 80-е годы и в первой половине 90-х годов в мире проводилось до десяти ежегодных международных конференций, посвященных процессам переработки природного газа, и реакции ОКМ в частности. В короткий срок появилось огромное число публикаций и патентов. Были протестированы сотни катализаторов, однако процесс не удалось довести до практической реализации, так как в процессе реакции образуются продукты (этан, этилен, пропилен и т.д.), более реакционноспособные, чем исходный метан.

В конце концов, неспособность обнаружить селективный катализатор привело к постепенной потере интереса к ОКМ. Начиная с середины 1990-х годов, научно-исследовательская деятельность в этой области практически прекратилась. Однако оказалось, что в последнее время интерес к окислительной конденсации метана вновь начинает расти, что связано с возможностью использования наноматериалов (*nanoparticles)* в качестве катализаторов данного процесса. Из анализа текстовых пересечений терминов удалось установить, что в течение пяти лет (с сентября 2009 по август 2013 года) осуществлялся проект *OCMOL* (7-я Европейская рамочная программа научно-технологического развития), в рамках которого решались основные технологические проблемы в области окислительной конденсации метана, олигомеризации этилена, мембранного и адсорбционного разделения, сухого риформинга метана, синтеза кислородных соединений и их конверсии в жидкие вещества.

Следующий этап анализа с применением разработанной терминологической системы связан с извлечением информации об использованных катализаторах, методах их приготовления и т.д., который уже выходит за рамки изложения данной работы.

Заключение

В рамках работы получены следующие результаты:

а) На основе L-граммного метода разработана методика анализа и извлечения терминологического спектра из текстовых коллекций в области химии и химического катализа. Особенностью модели является использование расширенных правил фильтрации терминоподобных словосочетаний, гибкой системы правил отбора и исключения терминов, а также морфологической и структурной информации.

б) На основе разработанного метода создана система терминологического анализа, выполняющая автоматизированный терминологический анализ текстовых коллекций, представленных в pdf-формате.

в) Разработано средство для выделения логической структуры из pdf-файлов публикаций (тезисов конференций, журнальных статей). Данное средство позволяет выделять такие структурные части публикации, как название, список авторов, организаций, основной текст публикации, библиографию.

г) Разработан инструментарий отчетов, позволяющих пользователю-эксперту в предметной области анализировать полученные результаты, находить тренды и закономерности в динамике поведения терминов.

д) Разработан пользовательский веб-интерфейс, содержащий компоненты для просмотра терминологической разметки, исходных и преобразованных текстов, работы с модулем отчетов.

Публикации

1. Альперин Б.Л. Программная система терминологического анализа научной периодической литературы в области химии // Материалы 52-й Международной научной студенческой конференции МНСК-2014: Информационные технологии / Новосиб. гос. ун-т. Новосибирск, 2014. 1 с.
2. Альперин Б.Л, Гусев В.Д., Ильина Л.Ю., Кузьмин А.О., Саломатина Н.В., Пармон В.Н. Разработка открытого терминологического веб-ресурса и создание начальных версий тезаурусов по катализу // Российский Химический Журнал (Журнал Российского химического общества им. Д.И.Менделеева). – 2013. – T.LVII. – №6. – 13 с.
3. Альперин Б.Л. Создание терминологического ресурса по катализу и его применение в системе текстового поиска // 50-я юбилейная Международная научная студенческая конференция «Студент и научно-технический прогресс». Новосибирск, 13-19 апреля 2012г. – Новосибирск, издательство Новосибирского государственного университета, 2012. – С.185.
4. Альперин Б.Л., Кузьмин А.О., Саломатина Н.В., Гусев В.Д. Веб - приложение для работы с тезаурусами и рубрикаторами // Свидетельство о государственной регистрации №50201001265.

Литература

1. SALTON G. Developments in Automatic Text Retrieval // SCIENCE - 1991 - VOL. 253 - P. 974-980.
2. SciFinder - The choice for chemistry research. URL: http://www.cas.org/products/scifinder (дата обращения: 27.05.2014).
3. Реферативная база данных Scopus. URL: http://health.elsevier.ru/electronic/product_scopus/ (дата обращения: 27.05.2014)
4. Russian microsite - IP&Science - Thomson Reuters. URL: http://wokinfo.com/russian/ (дата обращения: 27.05.2014).
5. Edward Vanhoutte. An Introduction to the TEI and the TEI Consortium // Literary and Linguistic Computing – 2004 – VOL 19(1). doi: 10.1093/llc/19.1.9
6. ANSI/NISO Z39.96-2012 standard, JATS: Journal Article Tag Suite. URL: http://jats.niso.org/
7. Constantin A., Pettifer S., Voronkov A. PDFX: Fully-automated PDF-to-XML Conversion of Scientific Literature // DocEng '13 Proceedings of the 2013 ACM symposium on Document engineering - 2013. - P. 177-180.
8. Teresa K Attwood, Douglas B Kell, Philip McDermott, James Marsh, Steve Pettifer, and David Thorne. Utopia documents: linking scholarly literature with research data // Bioinformatics – 2010 - VOL 26(18) – pp i568–i574.
9. JohnW. Ratcliff and David Metzener. Pattern matching: The gestalt approach // Dr. Dobb's Journal – 1988 – P.46
10. Kan M.-Y., Luong M.-T., Nguyen T. D. Logical Structure Recovery in Scholarly Articles with Rich Document Features // Int. J. Digit. Library Syst. - 2010. - VOL. 1(4) - P. 1-23.
11. Jessop D. M., Adams S. E., Willighagen E. L., Hawizy L., Murray-Rust P. OSCAR4: a flexible architecture for chemical text-mining //J Cheminform. - 2011 -VOL. 3(1) - P. 41.
12. Hawizy L., Jessop D. M., Adams N., Murray-Rust P. ChemicalTagger: A tool for semantic text-mining in chemistry // J Cheminform. - 2011. - VOL. 3(1) - P. 17.
13. Witten I. H., Paynter G. W., Frank E., Gutwin C., Nevill-Manning C. G. KEA: practical automatic keyphrase extraction // Proceedings of the fourth ACM conference on Digital libraries. - Berkeley, California, USA: ACM, 1999. - P. 254-255.
14. G´abor Berend R. F. SZTERGAK: Feature Engineering for Keyphrase Extraction // Proceedings of the 5th International Workshop on Semantic Evaluation, ACL, Uppsala, Sweden, 15-16 July 2010. - 2010. - P. 186-189.
15. Thuy Dung Nguyen M.-T. L. WINGNUS: Keyphrase Extraction Utilizing Document Logical Structure // Proceedings of the 5th International Workshop on Semantic Evaluation, ACL, Uppsala, Sweden, 15-16 July 2010. - 2010. - P. 166-169.
16. El-Beltagy S. R., Rafea A. KP-Miner: A keyphrase extraction system for English and Arabic documents // Information Systems. - 2009. - VOL. 34 (1) - P. 132-144.
17. PDF Text Extraction for Java &.NET – Snowtide. URL: http://snowtide.com/ (дата обращения 27.05.2014)
18. Gusev V. D., Salomatina N. V., Kuzmin A. O., Parmon V. N. An express analysis of the term vocabulary of a subject area: the dynamics of change over time // Autom. Doc. Math. Linguist. - 2012. - VOL. 46(1) - P. 1-7.
19. Stanford CoreNLP - A Suite of Core NLP Tools URL: http://nlp.stanford.edu/software/corenlp.shtml (дата обращения 27.05.2014)
20. Taylor A., Marcus M., Santorini B. The Penn Treebank: An Overview // Treebanks / Abeillé A.Springer Netherlands, 2003. – 2003 - P. 5-22.
21. IUPAC Compendium of Chemical Terminology (Gold Book) URL: http://goldbook.iupac.org/about.html (дата обращения 27.05.2014)

22. The Corncob list of more than 58 000 English words. URL:www.mieliestronk.com/wordlist.html (дата обращения 27.05.2014)
23. The Berkeley Laboratory Isotopes Project's URL: http://ie.lbl.gov/education/isotopes.htm (дата обращения 27.05.2014)
24. J.Bloch. Effective java (2 edition) // Prentice Hall – 2006 – pp. 384. ISBN: 978-0321356680
25. Apache Maven – software project management and comprehension tool. – URL: http://maven.apache.org/ (дата обращения 27.05.2014)
26. Гамма, Э. Приемы объектно–ориентированного проектирования // Э. Гамма, Р. Хелм, Р. Джонсон, Д. Влиссидес – СПб: Питер, 2001 – 368 с
27. Rod Johnson, Juergen Hoeller, Alef Arendsen, Thomas Risberg, Colin Sampalean Professional Java Development with the Spring Framework // Wiley Publishing, Inc. - 2005 – 676 pp. ISBN-13: 978-0-7645-7483-2
28. Silex - The PHP micro-framework based on Symfony2 Components URL: http://silex.sensiolabs.org (дата обращения 27.05.2014)
29. Kristina Chodorow MongoDB: The Definitive Guide, 2nd Edition // O'Reilly Media – 2013 - 423 pp, ISBN 978-1-44934-468-9
30. Ilona van Zandvoort, Yuehu Wang, C.B. Rasrendra, Pieter C.A. Bruijnincx, H.J.Heeres, Bert M. Weckhuysen Towards catalytic valorization of humin by-products formed during biomass processing // 11th European Congress on Catalysis – EuropaCat-X - 2013 - 2 pp.
31. Thi Minh Chau Hoang, Leon Lefferts and K. Seshan Catalytic gasification – a potential route for valorisation of humin based by-products formed during biomass processing // 11th European Congress on Catalysis – EuropaCat-X - 2013 -2 pp.
32. McKenzie Primerano Kohn, Catalytic Reforming of Biogas for Syngas Production, Columbia University – 2012 -155pp.