МИНИСТЕРСТВО ОБРАЗОВАНИЯ И НАУКИ РОССИЙСКОЙ ФЕДЕРАЦИИ

Новосибирский государственный университет (НГУ)

Факультет информационных технологий

Кафедра компьютерных систем

Магистерская диссертация

Ыдырыс Жулдызай Сериккызы

“Разработка и реализация алгоритма извлечения из текста географических названий, отражающих содержание документа”

Научный руководитель
д.т.н., доцент, с.н.с. ИВТ
СО РАН
Барахнин Владимир
Борисович

Новосибирск, 2013г

СОДЕРЖАНИЕ

ВВЕДЕНИЕ

В настоящее время с учетом возрастающей потребности общества в информационном обеспечении, в том числе связанным и с географическим аспектом информации, всё большую актуальность приобретают разработки, направленные на интеграцию «негеографических» информационных систем с информационными системами, изначально ориентированных на обработку географической информации. Добавление географического аспекта к информации, хранящейся в таких системах, как, например, электронные библиотеки, позволяет существенно повысить функциональность навигационных, поисковых и визуализационных сервисов этих систем, в частности, находить информацию, которая относится к конкретному географическому региону.

Следует отметить, что существующие программные комплексы для организации электронных библиотек не обладают необходимой функциональностью по хранению и обработке географических данных. Наделение же их требуемой функциональностью осложняется отсутствием единых стандартов на поиск и представление данных, связанных с географическим аспектом, которые сопрягались бы с существующими геоинформационными системами, т.е. с системами, для которых географический аспект информации является основным [3]. Отсюда вытекает актуальность и перспективность создания технологии, обеспечивающей обработку географической информации в «негеографических» информационных системах общего назначения.

1. ПОСТАНОВКА ЗАДАЧИ

Были поставлены следующие задачи:

1.Создание тезауруса географических названий

2.Разработка и реализация алгоритма извлечения из текста географических названий, отражающих содержание документа.

Должны были решаться вопросы:

1. Извлечения из текста документа названий, входящих в соответствующий тезаурус
2. Пополнения лексического словаря словоформами географических названий
3. Выявления омонимов и определение названий, отражающих содержание документа.

Актуальность работы

Для задачи извлечения из текста географических наименований необходимо иметь соответстующий тезаурус, а также все их словоформы. Поиск и склонение столь большой выборки слов вручную – не эффективно и не дает гарантий безошибочного ввода данных. А значит, возникает необходимость создать алгоритм, позволяющую извлекать из текста географические названия и разработать удобный и достаточно простой для использования интерфейс.

Целью диссертационной работы является реализовать алгоритм извлечения ключевых слов из текстов произвольной тематики, адаптировав его для работы с географическими названиями.

2. СОЗДАНИЕ ТЕЗАУРУСА ГЕОГРАФИЧЕСКИХ НАЗВАНИЙ

2.1 Основные понятия тезауруса

Географический аспект информации может быть зафиксирован на уровне метаданных, описывающих содержание документа. При этом географические метаданные объекта могут быть заданы двумя способами:

- с помощью количественного геометрического описания географического объекта на основе координат;
- с помощью ссылки на элемент некоторого тезауруса, включающего географические названия соответствующих объектов.

Первый вариант более предпочтителен, но он не очень удобен по причине необходимости внесения существенных изменений в уже существующие информационные системы, в отличие от второго варианта, который может быть реализован на базе существующих парадигм информационных систем при условии их небольшой модернизации. Поэтому далее речь пойдет только о втором варианте.

Существует множество тезаурусов географических наименований, однако сложность их использования заключается в том, что географический аспект объектов, хранящиеся в электронных библиотеках, зачастую относится не к текущему, а к прошедшему моменту, в то время как большинство тезаурусов содержит информацию, относящуюся только к текущему моменту.

Важно подчеркнуть, что могут меняться не только географические названия, к чему все уже привыкли, но и границы геометрических объектов, которые соответствуют объектам географическим. При этом любые изменения географических названий и геометрических объектов, ассоциированных с ними, как правило, привязываются к какому-нибудь нормативному документу, будь то постановление того или иного органа власти или соответствующая историческая хроника.

Таким образом, для использования в информационных системах (в электронных библиотеках) географического аспекта в его любом виде необходим справочный аппарат (тезаурус), который бы включал в себя не только географический аспект информации, но и ее временной (исторический) аспект.

Существует большое количество тезаурусов, описывающих понятийные и терминологические системы многих предметных областей. Высококачественные тезаурусы в большинстве своем создаются вручную. Процесс поддержания тезаурусов в актуальном состоянии довольно трудоемок, особенно в быстро развивающихся областях.

Таким образом, ручное построение тезаурусов становится «узким местом» для практической реализации проектов, использующих тезаурусы для решения своих задач, требуются методы автоматизации их наполнения и поддержки. Одной из проблем, возникающей при автоматическом наполнении тезаурусов, является большое количество «шума», который надо эффективно отсеивать. В связи с этим, наряду с автоматическими методами используют последующую ручную обработку полученного материала для получения данных большей точности (такие методы называются автоматизированными).

Тезаурус предоставляет набор нормализованной лексики, состоящий из ключевых понятий с множеством заданных семантических отношений между ними.

Тезаурус (от греческого - сокровище) – многозначный термин. Наиболее адекватный русский перевод слова «тезаурус» – это «мир знаний и интересов». Например, «мир знаний и интересов художника – тезаурус художника», «мир знаний и интересов бизнесмена – тезаурус бизнесмена» или иначе говоря, тезаурус предметной области «Бизнес». и т.п. [5]. Тезаурус - полный систематизированный набор данных по какой-либо предметной области, позволяющий человеку или компьютеру в ней ориентироваться.

Построения тезауруса – непростая задача. Несмотря на то, что существуют определенные стандарты построения тезаурусов, не всегда возможно напрямую воспроизвести существующие методики. Причинами этого являются, во-первых, отсутствие источников лексической информации (например, размеченных корпусов текстов) во-вторых, специфика подъязыка предметной области.

Первую проблему можно решить, создав подобный лексикографический ресурс самостоятельно.

Решение второй проблемы не является тривиальным, проблема усугубляется еще и тем, что необходимо реализовать инструмент, который был бы универсальным, то есть, подходил бы для автоматизированного построения тезаурусов любых предметных областей. Для этого необходимо спроектировать структуру тезауруса таким образом, чтобы можно было легко настаивать тезаурус под любую область.

Важнейшей задачей, возникающей в процессе добавления к описанию документа географических метаданных, является извлечение из его текста географических названий, входящих в тезаурус и отражающих содержание документа [10]. Ввиду того, что электронные библиотеки нередко содержат десятки тысяч (а иногда и миллионы) документов, решение указанной задачи невозможно без ее максимальной автоматизации.

2.2 Структура тезауруса

При наличии соответствующих данных тезаурус включает следующие основные сведения:

• нормализованное наименование географического объекта;

• принятые на текущий момент наименования географических объектов (иные варианты

названий, сокращенные названия);

• географические координаты;

•род географического объекта;

• административный статус населенного пункта;

• административно-территориальная привязка географического объекта (наименование субъ-

екта РФ, наименование административного района);

• соответствующий индекс (территориальное типовое деление);

• источники данных и другие примечания.

Структура записи обеспечивает учет смысловых связей между географическими названиями. На рис.1 приведена структура тезауруса.

Рисунок 1 – Структура тезауруса.

При содержательном наполнении тезауруса связанных с географическими объектами (населенные пункты, реки и т. д.), расположенными на территории

Российской Федерации, целесообразно использовать тезаурус географических названий Российской государственной библиотеки [8, 9]. К его достоинствам следует отнести полноту и наличие ссылок на нормативные документы, определяющие наименование объекта.

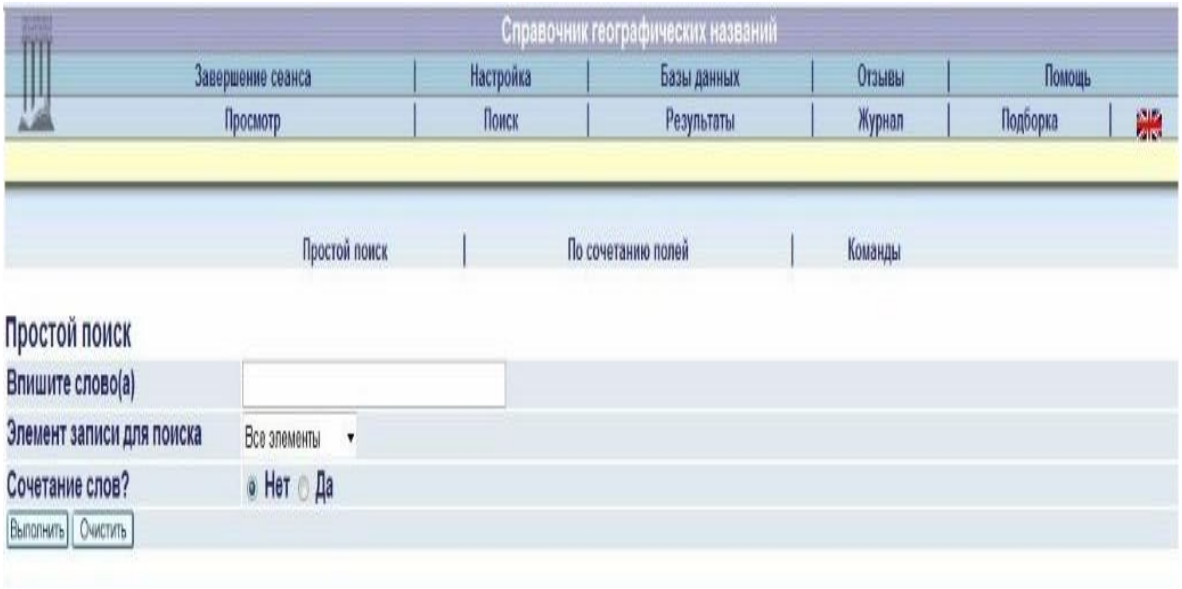


Рисунок 2 – Справочник географических названий Российской государственной библиотеки.

В справочнике географических названий РГБ подготовлены данные более чем о 140 тыс. географических объектов России (64 региона РФ) и загружены нормативные записи примерно для 80000 географических объектов. Для входа в справочник географических названий нужно выбрать в верхнем меню функцию «Базы данных», затем в списке найти раздел «Электронные справочники» и выбрать «Географические названия». (рис.2).

Поиск ведется по произвольным словам, сочетаниям слов и их частей с учетом иерархических и ассоциативных связей между лексическими единицами тезауруса.

2.3 Реализация тезауруса с использованием словаря географических названий Российской государственной библиотеки.

Из словаря Российской государственной библиотеки мы взяли достаточное количество географических наименовании (рис.4). Пользуясь этими данными мы создаем собственный тезаурус.

Графический интерфейс тезауруса (рис.5,6) должен иметь удобный вид, возможность добавления новых названий или пополнения их словоформами. Названия должны быть отсортированы в алфавитном порядке.

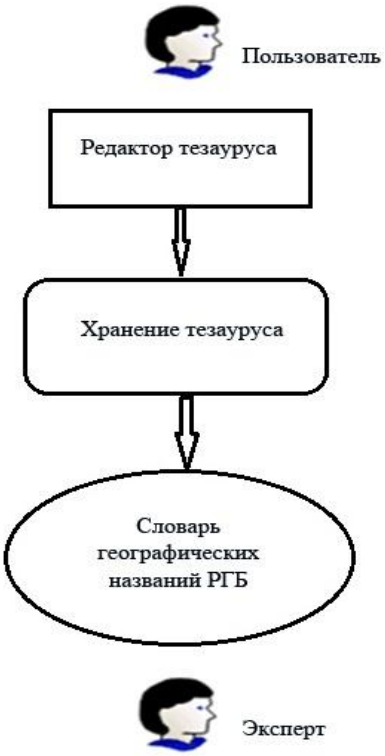


Рисунок 3 –РГБ–как выбор словаря для тезауруса.

Агапа	Бабушкина Имени	Ванавара	Кирс
Агаповка	Бабушкинский курорт	Ванино	Кирсанов
Агаповское месторождение	Бабынино	Ванкарем	Кисегач
Агвали	Бавлы	Варангер-Фьорд	Киселевск
Агидель	Багаевский	Варгаши	Кисловодск
Агинская степь	Баган	Варзи-Ятчи	Китойские Гольцы
Агинский Бурятский автономный округ	Багдарин	Варзуга	Кичменгский Городок
Агинское	Багратионовск	Варна	Кия
Агинское	Баджальский хребет	Варнавино	Киясово
Аграханский полуостров	Бадяриха	Варьеганское (Варьеганское) месторождение	Клетня
Агрия	Баево	Васильевский	Клетский
Агрыз	Баженовское месторождение	Васильсурск	Климово
Агульские Белки	Базардюзю	Васюган	Климовск
Адайхох	Базарные Матаки	Васюганская равнина	Клин
Адамовка	Базарный Карабулак	Вах	Клинско-Дмитровская гряда
Адзьва	Байдарацкая губа	Вача	Клинцы
Адлер	Байкал	Вачи	Кличкинский хребет
Адыгейск	Байкал	Вашка	Клухори
Адыге-Хабль	Байкало-Амурская магистраль	Ведено	Клухорский перевал
Адыгея	Байкалово	Вейделевка	Ключевская сопка
Адыча	Байкало-Ленский заповедник	Велиж	Ключи
Азас	Байкальск	Великая	Ключи
Азау	Байкальский заповедник	Великая	Ключи
Азнакаево	Байкальский хребет	Великие Луки	Клявлино
Азов	Байкит	Великий Устюг	Клязьма
Азово	Баймак	Великовечное	Клязьминское водохранилище
Азово-Кубанский артезианский бассейн	Бай-Тайга	Великое	Княгинино
Азовское море	Бай-Хаак	Вель	Кобан
Ай	Бакал	Велье	Кобра
Айваседапур	Бакалы	Вельмо	Ковдозерское (Княжегубское) водохранилище

Рисунок 4-Географические названия полученные из словаря Российской государственной библиотеки.

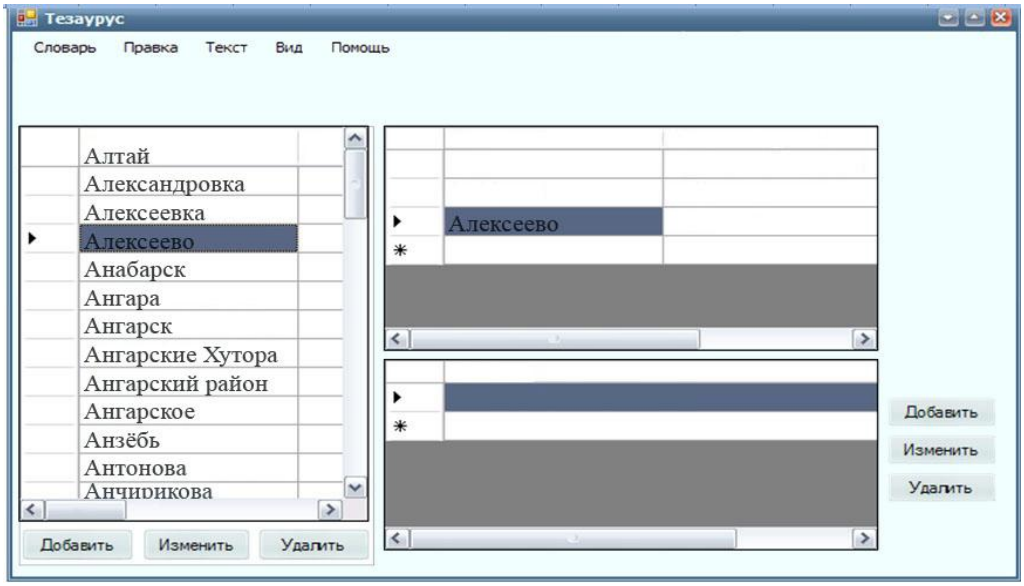


Рисунок 5-Графический интерфейс редактора тезауруса.

Как мы видим интерфейс редактора тезауруса очень удобный. На рисунке 6 показано как можно добавлять новые названия не имеющиеся в словаре. Тезаурус можно дополнить также синонимами официальных названий: например, название города Санкт-Петербург в разных источниках может быть употреблен как Петербург , Питер, северная столица или же город на Неве.

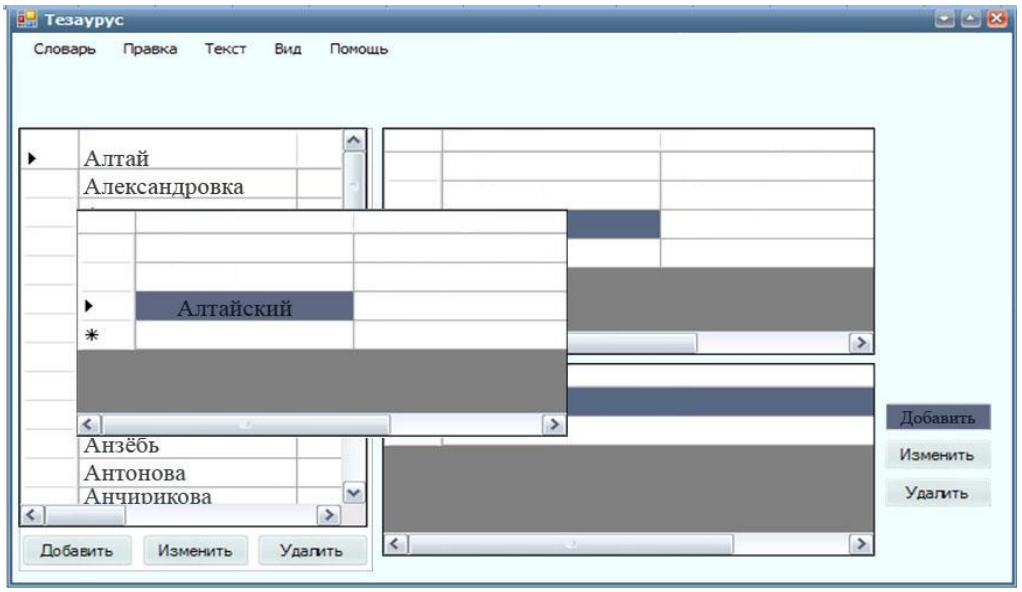


Рисунок 6-Добавление новых названий в тезаурус.

3. ПОПОЛНЕНИЕ ЛЕКСИЧЕСКОГО СЛОВАРЯ СЛОВОФОРМАМИ ГЕОГРАФИЧЕСКИХ НАЗВАНИЙ

3.1 Основные понятия морфологического анализа текста

В системах орфографического контроля русских текстов, системах автоматического индексирования документов используется принцип построения алгоритма морфологического анализа текстов. Он основан на принципе аналогии, который описывается в работе Белоногова Г. Г. [18].

При автоматической обработке текста возникает проблема «новых» слов. Для синтаксического анализа и синтеза необходимо знать грамматические характеристики слов. Если слова в словаре нет, то морфологический анализ не может быть выполнен, а, следовательно, не могут быть определены грамматические характеристики слова. Для того чтобы определить грамматические характеристики слов без словаря, Белоногов предложил принцип аналогии. Он основан на том, что существует сильная корреляционная связь между грамматическими характеристиками слов и буквенным составом их концов.

Все слова были разбиты на флективные классы (типы словоизменения), каждому из которых ставилась в соответствие система окончаний всех словоформ слова-представителя. А значит, теоретически, существует алгоритм, который каждому слову, для которого известен флективный класс, сопоставляет всевозможные формы этого слова путем присоединения определенных окончаний к основе.

Для задач извлечения из текстов географических наименований необходимо иметь все их словоформы. Склонение столь большой выборки слов вручную – не эффективно, и не дает гарантий безошибочного ввода данных. Мы использовали веб-приложение на основе алгоритма Белоногова, позволяющую автоматически генерировать словоформы и разработать интерфейс, удобный и достаточно простой для использования.

В основе работы веб-приложения лежит алгоритм Г.Г.Белоногова [17], использующий разбиение слов языка на флективные классы, т.е. типы словоизменения, каждому из которых ставилась в соответствие система окончаний всех словоформ слова-представителя (основа существительных и прилагательных, как правило, остается неизменной). Всего Г.Г.Белоноговым выделено для существительных 66 флективных классов, для прилагательных –12, каждому из которых поставлен в соответствие полный набор окончаний.

В случае добавления в словарь географических названий число возможных флективных классов для существительных значительно уменьшается.

Это происходит, во-первых, за счет классов, относящихся к одушевленным существительным как мужского, так и женского рода (таких классов соответственно 19 и 8). Возможные совпадения названий географических объектов с одушевленными нарицательными существительными (Орёл, Горняк, Чуваши и т.п.) не нуждаются в специальном анализе, поскольку такая омонимия выявляется заранее в процессе предварительной работы с тезаурусом при составлении списка географических названий, имеющих «негеографические» омонимы, и «многозначных» названий (подробнее об этом см. в следующем разделе), а это означает, что образец склонения слова-омонима уже имеется в лексическом словаре. Что же касается совпадения названий географических объектов с одушевленными собственными существительными (русскими фамилиями), то, как известно, соответствующие географические названия склоняются по неодушевленному образцу: с С.М.Кировым, но с городом Кировом.

Во-вторых, флективные классы для неодушевленных существительных зачастую различаются типом склонения лишь во множественном числе, однако для тех географических названий, которые соответствуют форме единственного числа, нет необходимости генерировать словоформы множественного числа.

Наибольшее количество возможных флективных классов, из которых приходится делать выбор, возникает при генерации словоформ географических названий, изначально стоящих во множественном числе: Печоры, Спас-Клепики, Выгоничи и Ливны относятся к разным флективным классам. Впрочем, при омонимии географических названий с неодушевленными нарицательными существительными образец склонения слова-омонима также имеется в лексическом словаре, что, в частности, исключает необходимость генерации словоформ географических названий, совпадающих со множественным числом нарицательных неодушевленных нарицательных существительных.

Что же касается прилагательных, входящих в географические названия, то в лексический словарь не входят либо притяжательные прилагательные, относящиеся к географическим названиям (Болотнинский район), либо диалектные, простонародные и т.п. слова (Верхнекокшенгский Погост), либо прилагательные, выступающие в качестве имен существительных (Новокручининский). При этом большинство подобных слов относится к одному флективному классу.

Таким образом, работа с веб-приложением заключается в следующем. Обрабатывая новое слово, эксперт устанавливает при необходимости его начальную форму и

указывает его тип: независимое существительное, прилагательное или зависимое слово-дополнение в родительном падеже. Зависимое слово сразу добавляется в словарь, так как единственной формой слова (применительно к соответствующему контексту) является оно само (море Лаптевых). При выборе независимого существительного на следующем шаге необходимо указать его род и число. Для прилагательного дополнительные характеристики не указываются.

Для уменьшения размеров надклассов, на которые разбиты флективные классы, применяется модификация алгоритма Г.Г.Белоногова, описанная в работе [11], состоящая в автоматическом анализе окончаний нормализованной словоформы внутри каждого надкласса, что приводит к значительному уменьшению количества элементов, из которых предстоит сделать выбор. Тем самым программа автоматически проводит предварительный анализ окончания слова, отсеивая те классы, к которым данное слово заведомо принадлежать не может. После этого нужно выбрать флективный класс, которому соответствует слово. Для выбора предоставляется таблица возможных флективных классов, которые определяются словом-представителем и его несколькими характерными словоформами.

Мысленно просклоняв данное слово по указанным формам и сравнив полученные окончания с окончаниями из таблицы, можно однозначно определить его флективный класс. После этого программа генерирует все формы слова, отображая их в виде таблицы, в которой они распределены по падежам и родам (если это прилагательное). Выводится список уникальных словоформ, так как обычно слово может иметь одинаковые окончания в разных формах. На основании этого списка эксперт принимает решение о занесении словоформ в словарь или, в исключительных случаях, когда сгенерированные словоформы оказываются неверными (например, у слова оказалась изменяемая основа), о переходе в ручной режим работы.

Веб-приложение для автоматизированной генерации словоформ географических названий позволяет пользователю отнести слово (существительное или прилагательное) к определенному классу склонения в соответствии с известным алгоритмом Г.Г.Белоногова [18], после чего генерирует все словоформы, опуская дубликаты.

Иллюстрация работы:

1. Пользователю предоставляется возможность ввести географическое название. На этом же шаге необходимо выбрать часть речи (существительное или прилагательное)

2. При выборе существительного на следующем шаге необходимо указать род или выбрать к какому типу это слово относится.
3. Далее, для существительного нужно выбрать номер флективного класса, которому соответствует слово. Появляется таблица с номерами флективных классов, которые определяются словом-представителем и его несколькими характерными словоформами, то есть, какие именно окончания имеет данное слово в определенной форме . А для сложносоставного географического названия появляется страница с примерными географическими названиями, которые склоняется так же как вводимое сложносоставное географическое название.
4. Просклоняв данное слово по указанным формам, и сравнив полученные окончания с окончаниями из таблицы, можно однозначно определить номер флективного класса. Здесь для удобства выбора предоставляется возможность сортировки по каждому полю таблицы. Кроме того, с учетом окончания анализируемого слова программа самостоятельно отбрасывает ненужные флективные классы, что так же облегчает выбор.

Всевозможные окончания слов хранятся в таблице, которая в свою очередь связана с морфологической таблицей, содержащей информацию о флективном анализе слова. С помощью нее программа генерирует все формы слова, отображая их в виде таблицы, в которой они распределены по падежам, числам и родам, если это прилагательное.

Программа выводит список словоформ, так как обычно слово может иметь одинаковые окончания в разных формах.

3.2 Пример работы веб-приложения автоматизированной генерации словоформ.

Итак, продемонстрируем, как работает веб-приложение. Пользователю предоставляется возможность ввести географическое название (рис. 7). На этом же шаге необходимо выбрать часть речи (существительное или прилагательное).

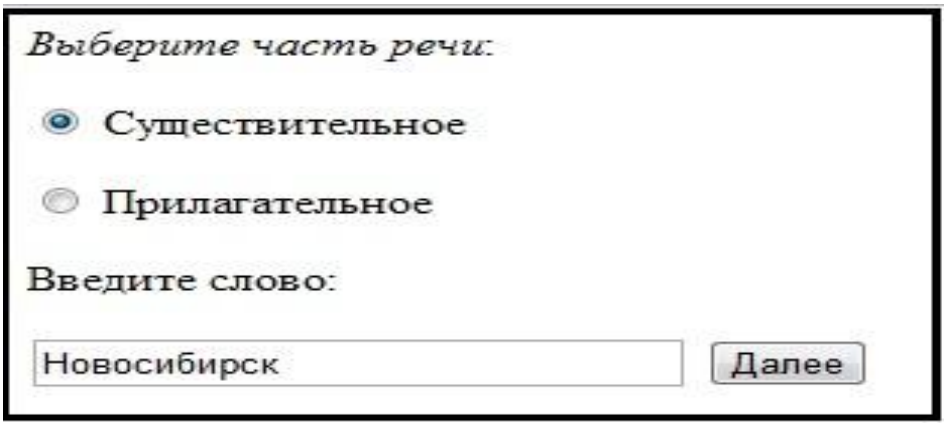


Рисунок 7 – Выбор части речи.

При выборе существительного на следующем шаге необходимо указать род или выбрать к какому типу это слово относится (рис.8).

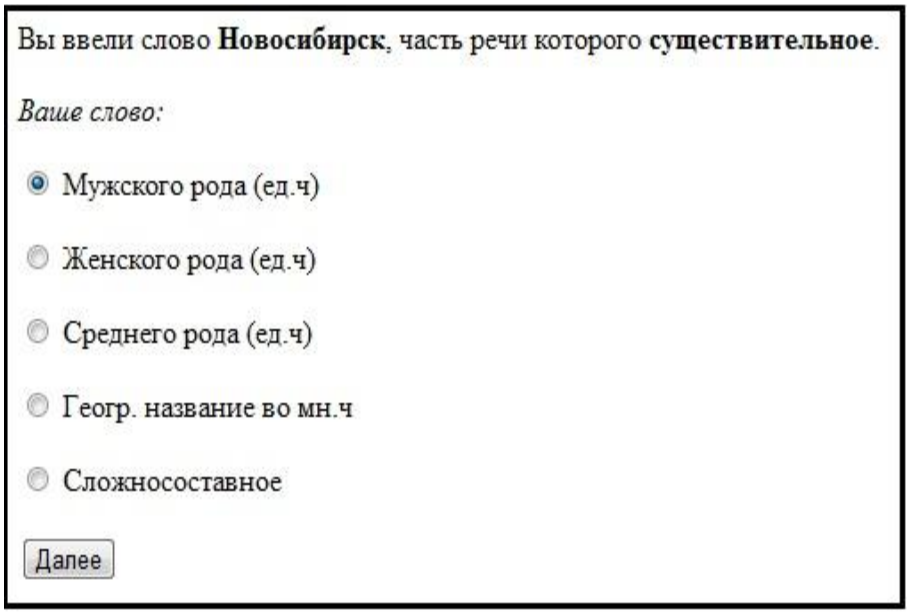
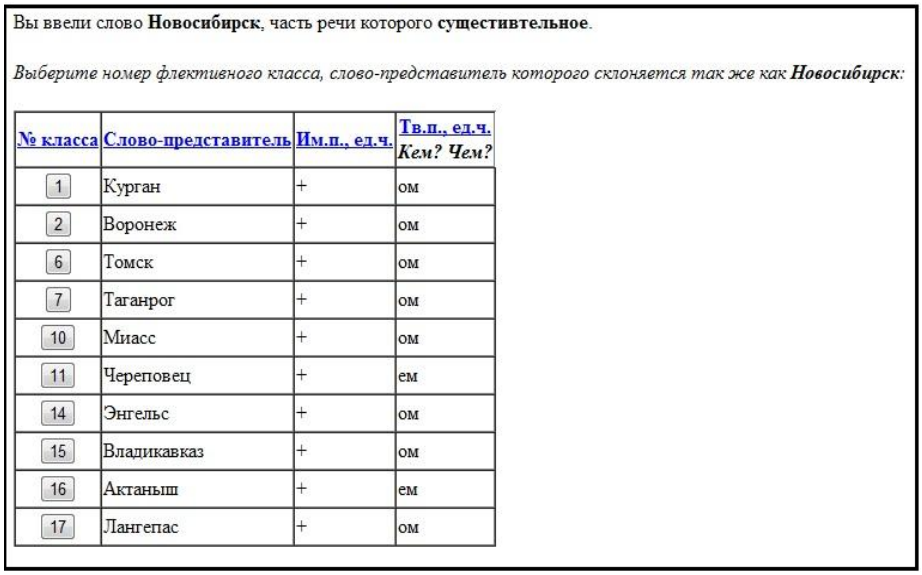


Рисунок 8 - Выбор рода слова «Новосибирск».

Далее, для существительного нужно выбрать номер флективного класса, которому соответствует слово. Появляется таблица с номерами флективных классов, которые определяются словом-представителем и его несколькими характерными словоформами, то есть, какие именно окончания имеет данное слово в определенной форме (рис. 9).



Вы ввели слово **Новосибирск**, часть речи которого **существительное**.

Выберите номер флективного класса, слово-представитель которого склоняется так же как ***Новосибирск****:*

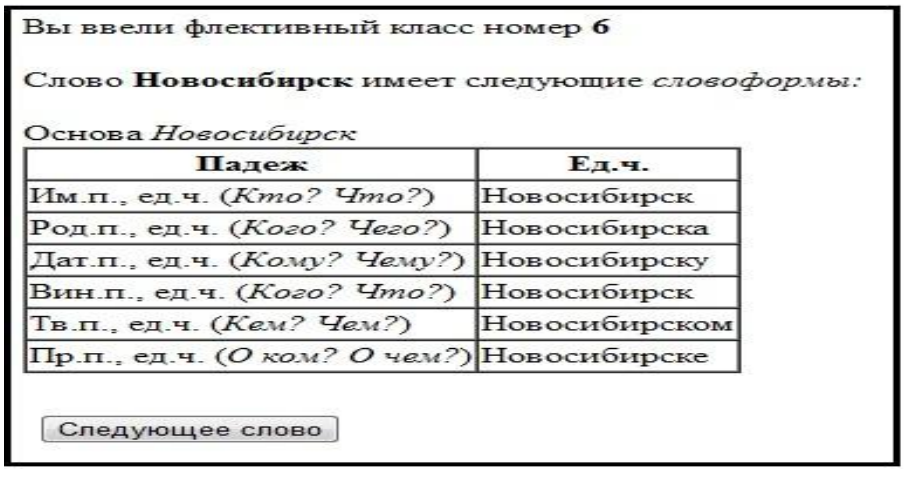
№ класса	Слово-представитель	Им.п., ед.ч.	Тв.п., ед.ч. *Кем? Чем?*
1	Курган	+	ом
2	Воронеж	+	ом
6	Томск	+	ом
7	Таганрог	+	ом
10	Миасс	+	ом
11	Череповец	+	ем
14	Энгельс	+	ом
15	Владикавказ	+	ом
16	Актаныш	+	ем
17	Лангепас	+	ом

Рисунок 9 – Выбор флективного класса слова «Новосибирск».

Просклоняв данное слово по указанным формам, и сравнив полученные окончания с окончаниями из таблицы, можно однозначно определить номер флективного класса. Здесь для удобства выбора предоставляется возможность сортировки по каждому полю таблицы. Кроме того, с учетом окончания анализируемого слова программа самостоятельно отбрасывает ненужные флективные классы, что так же облегчает выбор.

Всевозможные окончания слов хранятся в таблице, которая в свою очередь связана с морфологической таблицей, содержащей информацию о флективном анализе слова. С помощью нее программа генерирует все формы слова, отображая их в виде таблицы, в которой они распределены по падежам, числам и родам, если это прилагательное.

Программа выводит список словоформ, так как обычно слово может иметь одинаковые окончания в разных формах (рис. 10, рис. 11).



Вы ввели флективный класс номер **6**

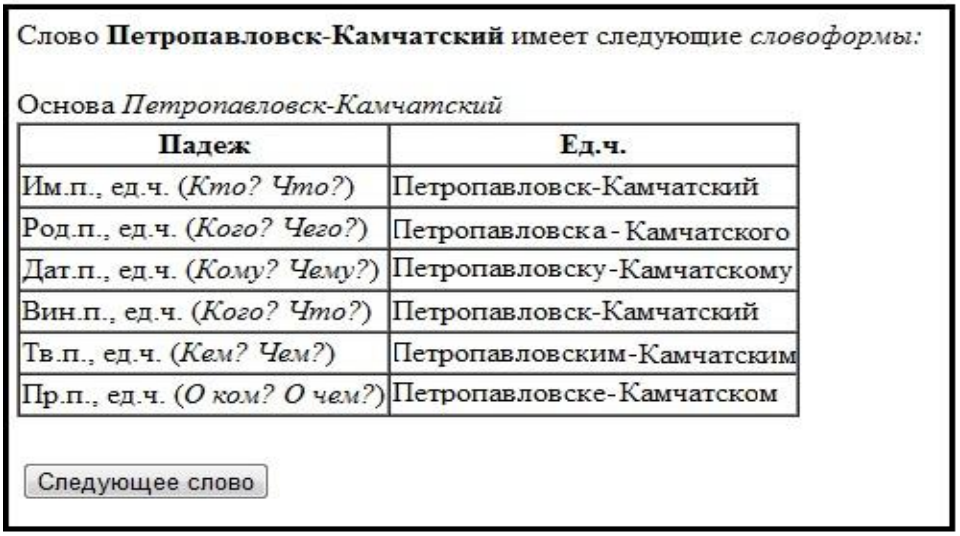
Слово **Новосибирск** имеет следующие *словоформы:*

Основа *Новосибирск*

Падеж	Ед.ч.
Им.п., ед.ч. (*Кто? Что?*)	Новосибирск
Род.п., ед.ч. (*Кого? Чего?*)	Новосибирска
Дат.п., ед.ч. (*Кому? Чему?*)	Новосибирску
Вин.п., ед.ч. (*Кого? Что?*)	Новосибирск
Тв.п., ед.ч. (*Кем? Чем?*)	Новосибирском
Пр.п., ед.ч. (*О ком? О чем?*)	Новосибирске

Следующее слово

Рисунок 10 - Вывод списка словоформ «Новосибирск».



Слово **Петропавловск-Камчатский** имеет следующие *словоформы:*

Основа *Петропавловск-Камчатский*

Падеж	Ед.ч.
Им.п., ед.ч. (*Кто? Что?*)	Петропавловск-Камчатский
Род.п., ед.ч. (*Кого? Чего?*)	Петропавловска-Камчатского
Дат.п., ед.ч. (*Кому? Чему?*)	Петропавловску-Камчатскому
Вин.п., ед.ч. (*Кого? Что?*)	Петропавловск-Камчатский
Тв.п., ед.ч. (*Кем? Чем?*)	Петропавловским-Камчатским
Пр.п., ед.ч. (*О ком? О чем?*)	Петропавловске-Камчатском

Следующее слово

Рисунок 11 - Вывод списка словоформ «Петропавловск-Камчатский».

При нажатии на кнопку «Следующее слово» приложение возвращается на первый шаг.

3.3 Выявление омонимов и определение названий, отражающих содержание документа

К сожалению, простой подсчет количества вхождений в документ слов и словосочетаний, содержащихся в соответствующем тезаурусе, не является удовлетворительным решением задачи извлечения из текста документа географических названий. Дело в том, что географические названия бывают омонимичны другим словам, являющимися именами как нарицательными: *Орёл, Белая* и т.п., так и собственными: *Киров, Кострома* и т.п. Кроме того, нередко одно и то же название носят сразу несколько различных географических объектов. Возникает необходимость отсеять из полученного набора слов омонимы географических названий, таковыми не являющиеся, а также установить, к какому конкретно географическому объекту относится найденное в документе «многозначное» название.

Важно подчеркнуть, что при решении этой задачи, как и при решении рассматриваемой далее проблемы, выявления названий, действительно отражающих содержание текста, целесообразно использовать следующий подход к определению более нежелательной ошибки («ошибки первого рода»). Отсутствие того или иного конкретного документа вряд ли будет замечено пользователем системы, ищущим документы,

привязанные к некоему географическому объекту. Напротив, обнаружив в результатах запроса документ, к интересующему объекту явно не относящийся (а особенно несколько подобных документов), пользователь с большой вероятностью утратит доверие к такой информационно-поисковой системе. Именно поэтому механизм разрешения коллизий должен быть достаточно строгим, чтобы обеспечить отсев посторонних документов.

Итак, для выявления в тексте омонимов географических названий, таковыми не являющихся, а также для конкретизации значения «многозначных» названий, необходимо:

1. Заранее в процессе работы с тезаурусом составить список географических названий, имеющих такие омонимы, и «многозначных» названий. Если «многозначные» названия в тезаурусе выявляются достаточно просто, путем его непосредственного анализа (а также путем сравнения тезауруса российских географических названий с иностранными аналогами: например, населенные пункты *Николаевка*, *Павловка* и т.п. имеются как в России, так и на Украине), то выявление омонимов «общего плана» – задача более сложная.
2. Омонимы имен нарицательных выявляются в процессе сравнения тезауруса географических названий с общим лексическим словарем, используемым программной библиотеке.
3. Омонимы имен собственных могут быть обнаружены при сравнении тезауруса с биографическими, мифологическими и т.п. словарями. При этом, разумеется, никогда нельзя быть уверенным в достаточно полном выявлении омонимов указанного типа, поскольку небольшие населенные пункты могут носить имена деятелей местного масштаба, чьи имена не встречаются в сколько-нибудь распространенных биографических словарях (но, вместе с тем, имена этих деятелей могут встретиться, например, в документах, посвященных истории соответствующего региона).

Кратко изложим подходы к выявлению омонимов и конкретизации «многозначных» названий применительно к конкретному документу. Наиболее общим приемом выявления нарицательных омонимов является учет регистра первой буквы слова. Этот прием может оказаться неэффективным, если омонимичное слово является первым словом в предложении, а также если заголовок документа набран прописными буквами. В случае неоднократного вхождения такого слова в текст почти наверняка удастся выявить его смысл путем анализа регистра первой буквы всех его вхождений. Если же омонимичное слово встречается только раз и притом в качестве первого слова в предложении, то относить его к географическим названиям вряд ли целесообразно хотя бы потому, что

географические названия зачастую употребляются с предлогом указания места или направления (т.е. не выступают в качестве первого слова предложения), а в случае возможной омонимии – и с указанием типа географического объекта (*город Орёл*, *река Белая* и т.п.).

Последнее соображение учитывается и при выявлении омонимов – имен собственных (не относящихся к географическим названиям): «географический» омоним будут характеризовать предлоги или (и) указание типа географического объекта, а «персональный» могут характеризовать, например, инициалы или имя персоны (все подобные сведения должны храниться в специальном дополнительном словаре, предназначенном для идентификации омонимов).

Наиболее сложна задача о конкретизации «многозначных» географических названий. Для ее решения следует учитывать возможное указание типа географического объекта (*река Москва*), вхождение в текст названия региона, к которому может принадлежать объект, вхождение в текст других географических названий, относящихся к этому региону (если последний явно не упомянут) и т.п. При этом следует учитывать сравнительную значимость объектов: например, при отсутствии дополнительных сведений о принадлежности к региону населенного пункта *Киров* речь почти наверняка идет об областном центре, а не о городе в Калужской области.

4.ОПИСАНИЕ РАБОТЫ АЛГОРИТМА

4.1 Этапы алгоритма

I. Извлечение из текста документа всех географических названий, входящих в тезаурус.

2. Создание предварительного индекса «номер названия» – «позиция слова в названии» – «название в символьном представлении».

3. Добавление встретившихся неизвестных слов в лексический словарь библиотеки, где им присваиваются идентификационные номера.

4. Переработка индекса в формат «номер названия» – «позиция в тексте» – «номер названия из лексического словаря».

5. Сбор статистики о длинах названий для реализации поиска и идентификации составных названий (т.е. названий, состоящих более чем из одного слова).

6. Сбор статистики о количестве вхождений отдельных слов для оптимизации поиска путем исключения из рассмотрения терминов, заведомо отсутствующих в тексте.

II. Заключительная стадия алгоритма– подсчет количества вхождений названий в текст (тексты). Ее этапы:

1. Подсчет возможных комбинаций «текст» – «название», основанный на статистике вхождения отдельных слов.

2. Нахождение всех потенциально возможных мест вхождения каждого названия в текст (тексты) на основе наличия хотя бы одного общего слова из лексического словаря. Позиция каждого потенциально возможного вхождения фиксируется.

4. Рассмотрение каждого из возможных мест вхождений с точки зрения соответствия названию в целом. Актуальность вхождения определяется наличием рядом с соответствующей позицией других слов, входящих в термин.

5. Исключение учета вхождений, поглощаемых более длинными вхождениями.

Отметим, что при решении задачи извлечения географических названий этапы 4 и 5 актуальны довольно редко, но все-таки их нельзя полностью исключить: например, практически равно употребительны термины *Новосибирский район* и *Новосибирский сельский район*, обозначающие один и тот же географический объект.

4.2 Реализация алгоритма

Для автоматизации работы эксперта построено приложение на языке PHP, автоматически извлекающий из текста слова, имеющиеся в соответствующем тезаурусе. Работа с приложением осуществляется через веб-интерфейс. Создана база данных на MySql где хранятся публикации и статьи виде текстовой информации.

Первым этапом решения поставленной задачи является извлечение из текста документа всех географических названий, входящих в тезаурус. Сразу оговоримся, что вхождение в текст документа притяжательного прилагательного, соответствующего тому или иному географическому названию (***новосибирский*** *метрополитен*, ***омский*** *спортсмен* и т.п.), будет приравниваться к вхождению в текст самого географического названия. Разумеется, для большей точности работы алгоритма тезаурус следует дополнить синонимами официальных названий: например, *Санкт-Петербург – Петербург – Питер – северная столица – город на Неве* и т.д., а также их производными, но мы отдаем себе отчет в трудоемкости и слабой формализуемости решения этой задачи.

Итак, фактически, мы имеем дело с задачей координатного индексирования текста терминами, входящими в заданный словарь, при этом термины могут состоять не только из одного, но и из нескольких (как правило, двух) слов, например, *Новосибирская область, Белое море, Северная Двина* и т.п. Ввиду того, что в русском языке имена существительные и прилагательные при склонении изменяют свою форму, разработка эффективного алгоритма автоматизации извлечения из текста ключевых терминов, в том числе и географических названий, представляет нетривиальную задачу, ибо необходимо учитывать и те случаи, когда слова, образующие термин, находятся не только в именительном (как они занесены в тезаурус, за редкими исключениями типа *море* ***Лаптевых***), но и в косвенных падежах.

Проиллюстрируем работу приложения на ряде записей из текстовой базы данных различных статьи. Из всей структуры записи нас будут интересовать несколько полей, а именно:

- заголовок статьи
- аннотация к статье
- источник
- место публикации

Заголовок статьи и аннотацию отнесем к контенту, источник и место публикации – к контексту.

Выберем из базы данных статью случайным образом (рис. 12).

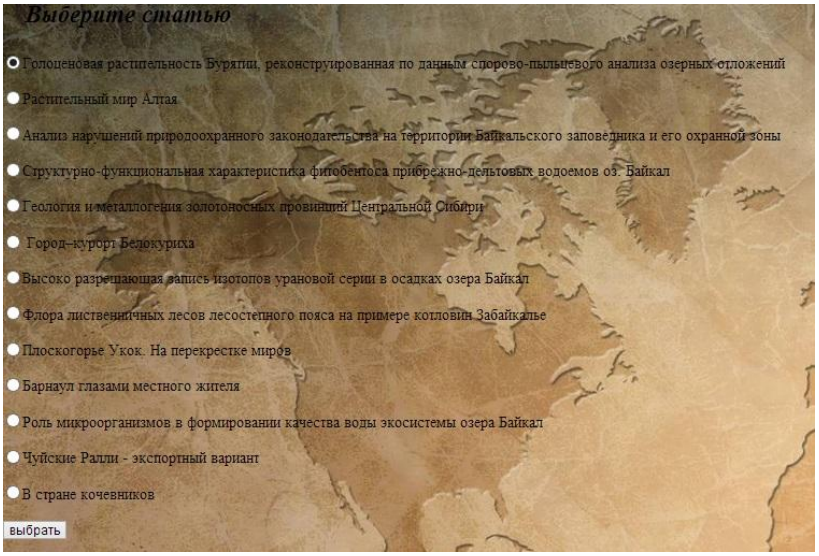


Рисунок 12 – Выбор статьи №1.

1). Заголовок: “Голоценовая растительность Бурятии, реконструированная по данным спорово-пыльцевого анализа озерных отложений”.

Источник: “Актуальные проблемы палинол. на рубеже 3-го тысячелетия”

Место публикации: “Москва”.

Аннотация: “В непосредственной близости к оз. Байкал растительность лесотундрового облика существовала 10000-10600 л. н. Ей на смену пришли холодные степи, широко распространившиеся 9000-10000 л. н. в условиях повышения летних температур и некоторого иссушения климата. После 9000 л. н. в регионе стала быстро распространяться древесная растительность. Период между 6000 и 8500 л. н. характеризуется максимальным за голоцен распространением березовых и еловых лесов, что предполагает значительный рост увлажненности и смягчение континентальности климата. Состав пыльцевых спектров свидетельствует о том, что участие пихты в составе темнохвойных таежных лесов достигло максимума в интервале 3000-6000 л. н. Пихта является наиболее требовательной к климатическим условиям древесной породой региона, что позволяет считать время ее распространения оптимальным с точки зрения соотношения тепла и влаги. На протяжении последних 3000 лет сосновые и лиственничные леса играют ключевую роль в растительном покрове, отражая ухудшение условий увлажненности и усиление континентальности климата.”

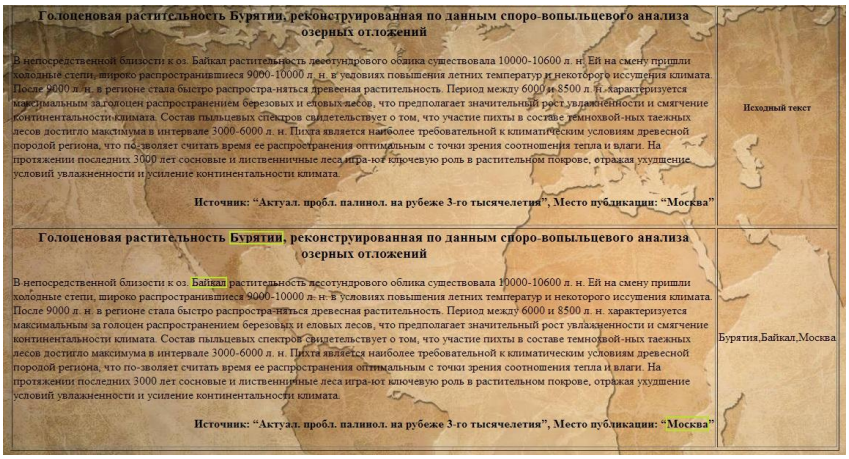


Рисунок 13 – Результаты обработки статьи №1.

Названия, относящиеся к контенту (заголовок, аннотация): Бурятия, Байкал.

Названия, относящиеся к контексту (источник, место публикации): Москва. (рис. 13).

Далее таким образом выбираем другие статьи (рис.14).

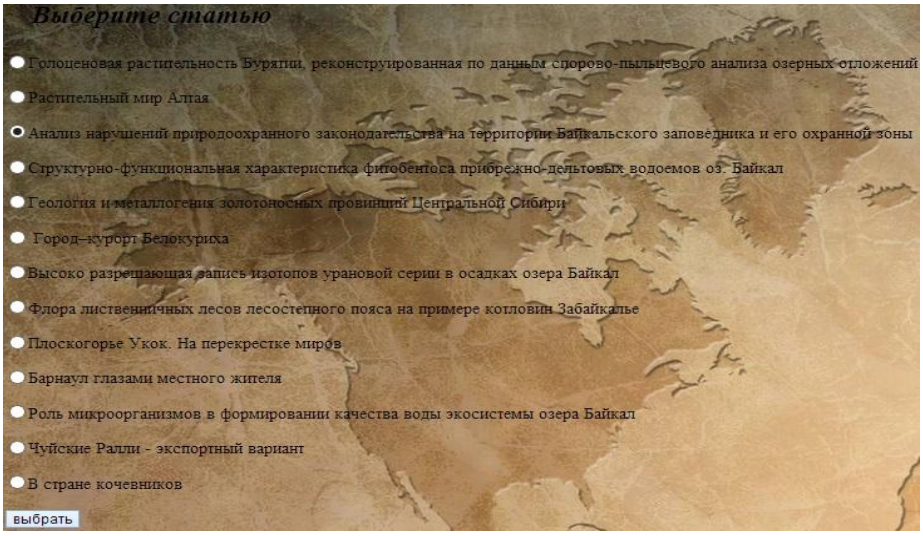


Рисунок 14 – Выбор статьи №2.

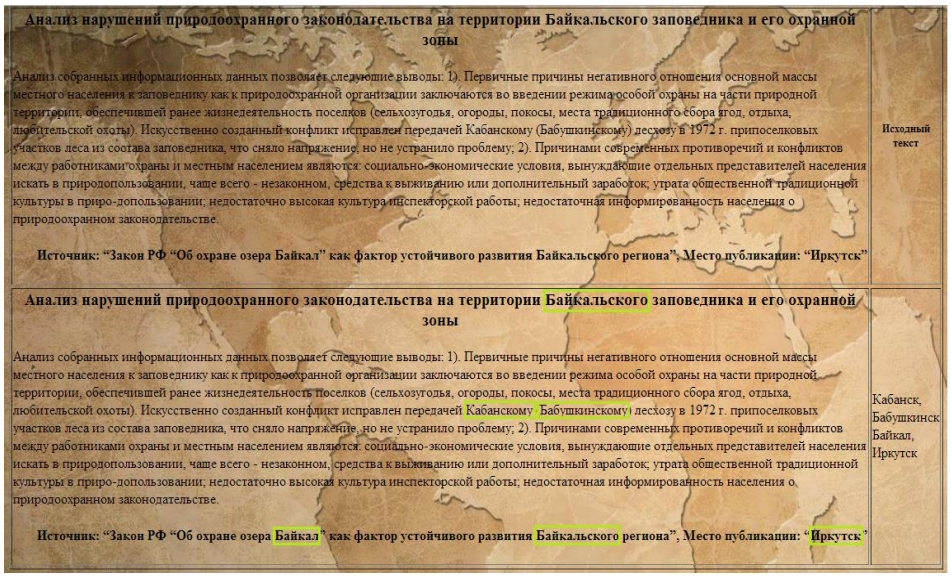
2). Заголовок: "Анализ нарушений природоохранного законодательства на территории Байкальского заповедника и его охранной зоны"

Источник: "Закон РФ "Об охране озера Байкал" как фактор устойчивого развития Байкальского региона".

Место публикации: "Иркутск".

Аннотация: "Анализ собранных информационных данных позволяет следующие выводы: 1). Первичные причины негативного отношения основной массы местного населения к заповеднику как к природоохранной организации заключаются во введении режима особой охраны на части природной территории, обеспечившей ранее жизнедеятельность поселков (сельхозугодья, огороды, покосы, места традиционного

сбора ягод, отдыха, любительской охоты). Искусственно созданный конфликт исправлен передачей Кабанскому (Бабушкинскому) лесхозу в 1972 г. припоселковых участков леса из состава заповедника, что сняло напряжение, но не устранило проблему; 2). Причинами современных противоречий и конфликтов между работниками охраны и местным населением являются: социально-экономические условия, вынуждающие отдельных представителей населения искать в природопользовании, чаще всего - незаконном, средства к выживанию или дополнительный заработок; утрата общественной традиционной культуры в природопользовании; недостаточно высокая культура инспекторской работы; недостаточная информированность населения о природоохранном законодательстве.”



Анализ нарушений природоохранного законодательства на территории Байкальского заповедника и его охранной зоны Анализ собранных информационных данных позволяет следующие выводы: 1). Первичные причины негативного отношения основной массы местного населения к заповеднику как к природоохранной организации заключаются во введении режима особой охраны на части природной территории, обеспечившей ранее жизнедеятельность поселков (сельхозугодья, огороды, покосы, места традиционного сбора ягод, отдыха, любительской охоты). Искусственно созданный конфликт исправлен передачей Кабанскому (Бабушкинскому) лесхозу в 1972 г. припоселковых участков леса из состава заповедника, что сняло напряжение, но не устранило проблему; 2). Причинами современных противоречий и конфликтов между работниками охраны и местным населением являются: социально-экономические условия, вынуждающие отдельных представителей населения искать в природопользовании, чаще всего - незаконном, средства к выживанию или дополнительный заработок; утрата общественной традиционной культуры в приро-дополь зовании; недостаточно высокая культура инспекторской работы; недостаточная информированность населения о природоохранном законодательстве. **Источник: “Закон РФ “Об охране озера Байкал” как фактор устойчивого развития Байкальского региона”, Место публикации: “Иркутск”**	**Исходный текст**
Анализ нарушений природоохранного законодательства на территории Байкальского заповедника и его охранной зоны Анализ собранных информационных данных позволяет следующие выводы: 1). Первичные причины негативного отношения основной массы местного населения к заповеднику как к природоохранной организации заключаются во введении режима особой охраны на части природной территории, обеспечившей ранее жизнедеятельность поселков (сельхозугодья, огороды, покосы, места традиционного сбора ягод, отдыха, любительской охоты). Искусственно созданный конфликт исправлен передачей Кабанскому Бабушкинскому) лесхозу в 1972 г. припоселковых участков леса из состава заповедника, что сняло напряжение, но не устранило проблему; 2). Причинами современных противоречий и конфликтов между работниками охраны и местным населением являются: социально-экономические условия, вынуждающие отдельных представителей населения искать в природопользовании, чаще всего - незаконном, средства к выживанию или дополнительный заработок; утрата общественной традиционной культуры в приро-дополь зовании; недостаточно высокая культура инспекторской работы; недостаточная информированность населения о природоохранном законодательстве. **Источник: “Закон РФ “Об охране озера Байкал” как фактор устойчивого развития Байкальского региона”, Место публикации: “Иркутск”**	Кабанск, Бабушкинск Байкал, Иркутск

Рисунок 15 – Результаты обработки статьи №2.

Названия, относящиеся к контенту (заголовок, аннотация): Байкал, Кабанск, Бабушкинск.
Названия, относящиеся к контексту (источник, место публикации): Байкал, Иркутск. (рис. 15).

3). Заголовок: “Роль микроорганизмов в формировании качества воды экосистемы озера Байкал”.

Источник: “Экология – 2007”.

Место публикации: “Архангельск”.

Аннотация: “Рациональное использование природных ресурсов озера Байкал, а также разработка водоохранных мероприятий невозможна без выявления роли микроорганизмов в процессах превращения веществ антропогенной природы, поступающих в озеро.

Микроорганизмы являются хорошими индикаторами при выявлении антропогенного фактора на экосистему озера. Необходимо срочно решать вопрос о строительстве очистных сооружений, поскольку в связи с развитием новой экономической зоны в п. Листвянка, нагрузка на экосистему озера будет все увеличиваться и эта ситуация может оказаться непредсказуемой. Озеро Байкал - внутриконтинентальный водоем с низкими скоростями водообмена, низкими температурами и слабыми микробиальными процессами деструкции.”

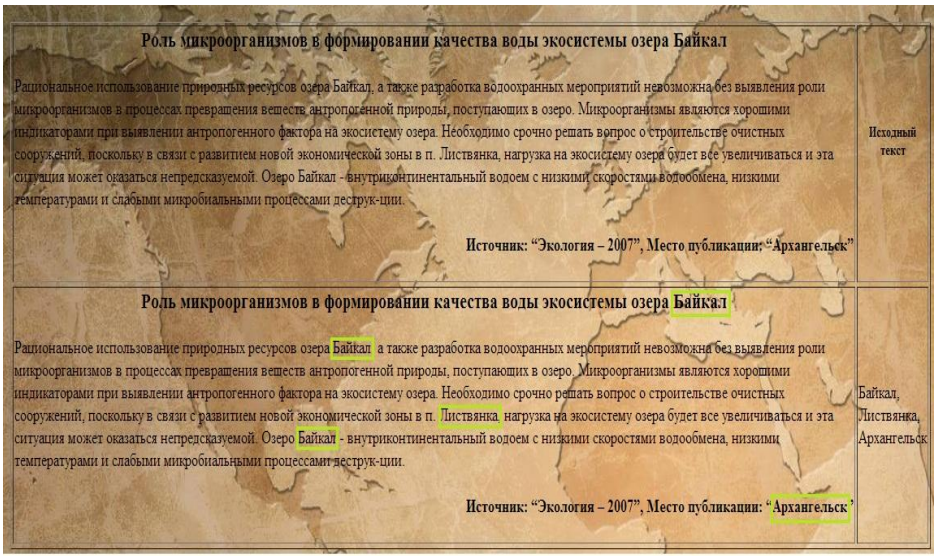


Рисунок 16 – Результаты обработки статьи №3.

Названия, относящиеся к контенту (заголовок, аннотация): Байкал, Листвянка.
Названия, относящиеся к контексту (источник, место публикации): Архангельск. (рис. 16).

После того, как из текста документа выделены все входящие в него географические названия и конкретизированы «многозначные» названия, неизбежно встает главный вопрос: какие именно названия отражают содержание документа?

По-видимому, наиболее простым вариантом, не требующим привлечения сложных алгоритмов семантического анализа текста, является проверка вхождения наименования в метаданные (атрибуты библиографического описания) документа.

Если географическое название встретилось в названии документа, его аннотации, ключевых словах, то такое название следует считать отражающим содержание документа.

Если же перечисленные атрибуты не содержат географических названий, то проводится анализ вхождения названий непосредственно в текст документа, при этом отражающими содержание документа признаются название, имеющее наибольшее число вхождений.

ЗАКЛЮЧЕНИЕ

Из полученных результатов можно сделать следующие выводы: алгоритм извлекает географические названия с достаточной точностью, кроме сокращенных, омономичных названий по причине сложностей с их определением.

В целом, работоспособность общего алгоритма была проверена при решении ряда практических задач.

Работа может быть использована в задачах, где требуется поиск и извлечение ключевых слов с текстов, с применением тезауруса предметной области, а также в геоинформационных системах.

ЛИТЕРАТУРА

1. *Жижимов О.Л., Мазов Н.А.* География и стандарты метаданных для электронных библиотек: содержание, применение, проблемы [Электронный ресурс] // Электронные библиотеки.
2009. Т.12. №1 http://www.elbib.ru/index.phtml?page=elbib/rus/journal/2009/part1/ZM

2. *Жижимов О.Л., Мазов Н.А.* Об использовании географических координат при поиске библиографической информации // Научные и технические библиотеки. 2009. № 1. С.54-60.

3. *Жижимов О.Л., Мазов Н.А.* Проблемы географической привязки цифровых объектов в электронных библиотеках // Труды Двенадцатой Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» (RCDL'2010). – Казань, 13-17 октября 2010 г. С. 207-214.

4. *Скачков Д.М., Жижимов О.Л.* О профиле доступа к данным тезауруса для ретроспективного геокодирования и географического поиска в электронных библиотеках // Восемнадцатая международная Конференция «Крым 2011», Судак, 4–12 июня 2011 г. http://www.gpntb.ru/win/inter-events/crimea2011/disk/059.pdf

5. *Скачков Д.М., Жижимов О.Л.* Об использовании ретроспективного геокодирования для географического поиска в электронных библиотеках // Труды Тринадцатой Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» (RCDL'2011). – Воронеж, 19-22 октября 2011 г. С. 30-37.

6. *The Zthes* specifications for thesaurus representation, access and navigation. http://zthes.z3950.org/

7. *Getty* Thesaurus of Geographic Names Online. http://www.getty.edu/research/tools/vocabularies/tgn/index.html

8. *Лаврёнова О.А.* Многоязычный доступ к данным на основе тезауруса географических названий // Сборник тезисов постерных докладов Девятой Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» (RCDL'2007).. Переславль-Залесский, 15-18 октября 2007 г. С. 57-62.

9. *Тезаурус* РГБ. – http://aleph.rsl.ru/F/?func=file&file_name=find-b&local_base=tst11

10. *Барахнин В.Б., Жижимов О.Л., Скачков Д.М.* Проблема извлечения из текстовых документов географических названий, отражающих содержание // Сборник трудов XI

Всероссийской конференции с участием иностранных ученых «Проблемы мониторинга окружающей среды» (ЕМ-2011). Кемерово, 24-28 октября 2011 г. С. 285-290.

11. *Шокин Ю.И., Федотов А.М., Барахнин В.Б.* Проблемы поиска информации. Новосибирск: Наука, 2010.

12. *Барахнин В.Б., Куперштох А.А.* Алгоритм координатного индексирования электронных научных документов // Труды международной конференции «Вычислительные и информационные технологии в науке, технике и образовании». Казахстан, Павлодар, 20-22 сентября 2006 г. Т. I. С.228-232.

13.*Библиотека* морфологического анализа phpMorphy. – http://phpmorphy.sourceforge.net

14. *Барахнин В.Б.*, *Нехаева В.А.* Технология создания тезауруса предметной области на основе предметного указателя энциклопедии // Вычислительные технологии. 2007. Т. 12. Специальный выпуск 2. С.3-9.

15. *Белоногов Г.Г., Новоселов А.П.* Автоматизация процессов накопления, поиска и обобщения информации. М.: Наука, 1979.

16. Г. Г. Белоногов, А. П. Новоселов Автоматизация процессов накопления, поиска и обобщения информации, Издательство «Наука», Москва, 1979.

17. В. Барахнин, Ж.С Ыдырыс, А.Б.Альменова, А.А.Енсегенова «Сборник трудов Международной научной конференции «Информационно-вычислительные технологии и математическое моделирование», г.Кемерово. 2013 г.

18. Гендина Н. И., Информационно-поисковые тезаурусы: основные виды и области применения.