ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ АВТОНОМНОЕ ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ
«БЕЛГОРОДСКИЙ ГОСУДАРСТВЕННЫЙ НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ»
(Н И У « Б е л Г У »)

ИНСТИТУТ МЕЖКУЛЬТУРНОЙ КОММУНИКАЦИИ И МЕЖДУНАРОДНЫХ ОТНОШЕНИЙ

КАФЕДРА АНГЛИЙСКОЙ ФИЛОЛОГИИ И МЕЖКУЛЬТУРНОЙ КОММУНИКАЦИИ

ОПТИМИЗАЦИЯ ПРОЦЕССА ПОИСКА ДАННЫХ НА АНГЛИЙСКОМ ЯЗЫКЕ ПРИ ПОМОЩИ АНАЛИЗА СЛОВОСОЧЕТАНИЙ

Выпускная квалификационная работа

обучающегося по направлению подготовки 45.03.04
Интеллектуальные системы в гуманитарной сфере
очной формы обучения,
группы 04001415
Сегеда Алексея Сергеевича

Научный руководитель
к. филол. наук, доцент
Данилова Е.С.

БЕЛГОРОД 2018

ОГЛАВЛЕНИЕ

ВВЕДЕНИЕ

С появлением и развитием вычислительной техники и информационных технологий, интернета и средств коммуникации, у специалистов самых разнообразных профессий и направлений деятельности появилась возможность и, вместе с тем, необходимость взаимодействия с большим количеством информации. Вне зависимости от сферы, платформы и применяемых технологий, ограничения, а вместе с тем и возможности работы с информацией всегда обусловлены двумя факторами: способом её хранения и поиском необходимых данных по всему её объему. Задача совершенствования хранения информации, как правило, решается при помощи увеличения количества доступного места, разработкой новых структур хранения, эволюцией баз данных, в большинстве своём опираясь на развитие физических носителей. А успешность решения задачи поиска необходимых данных зависит, прежде всего, от алгоритмов обработки данных, и, непосредственно, алгоритмов самого поиска.

Особое внимание вопросу поиска данных уделяется, когда речь заходит о всемирной системе объединённых компьютерных сетей для хранения и передачи информации - Интернете. Из формулировки понятно, что Интернет представляет собой множество локальных пространств, которые так или иначе способны взаимодействовать или ссылаться друг на друга. Это могут быть частные, корпоративные, бизнес, государственные и иные ресурсы. Они могут использоваться для обучения, общения, обмена, продаж и покупок, развлечения и досуга, а также огромного количества других целей. Они могут варьироваться в зависимости от размера – от одной до сотен и тысяч веб-страниц. Они могут различаться по количеству посещений – от десятков до сотен миллионов уникальных посетителей в сутки. Они могут содержать контент, на полное ознакомление с которым необходимо от нескольких минут до десятков лет. И, при этом, вопросами

хранения и возможностей поиска информации на ресурсах занимаются непосредственно их владельцы. Но далеко не все из них являются профессиональными разработчиками, способными найти удобные и эффективные решения описанных выше задач.

Таким образом, **актуальность** данной работы обусловлена отсутствием готовых инструментов осуществления поиска по локальном информационным пространствам.

В своё время, **изучением этого вопроса** занимались такие отечественные и зарубежные учёные как Ландэ Д.В., Снарский А. А., Безсуднов И. В, Chu H., Rosenthal M., Risvik K. M., Michelsen R., Tarakeswar M. K., Kavitha M. D. Их работы содержат в себе описания основных подходов к поиску информации и принципов работы поисковых систем.

Объектом исследования выступают поисковые процессы в информационном пространстве, а **предметом исследования –** поиск данных в локальных информационных пространствах.

Цель работы заключается в разработке информационно-поисковой системы локального характера при помощи анализа словосочетаний.

В **задачи** работы входит:

- изучение истории развития информационно-поисковых систем;
- анализ системы английских словосочетаний в современном языкознании, в том числе их структуры, грамматических функций и морфологических характеристик;
- изучение существующих разработок в сфере поиска данных;
- разработка модулей каталогизации, индексации, трансляции;
- разработка алгоритмов каталогизации, индексации и трансляции.

Работа включает в себя введение, две главы, заключение, список использованной литературы и 25 приложений.

ГЛАВА 1. Система словосочетаний в современном английском языке

В Главе 1 рассматривается краткая история развития информационно-поисковых систем и производится анализ английского словосочетания как основной единицы поиска.

1.1 История развития информационно-поисковых систем.

Информационно-поисковая система (далее – ИПС) - это некоторая система для обработки, хранения, сортировки, фильтрации и поиска больших массивов информации. Самой известной на сегодняшний день ИПС является поисковая система «Google». Она обрабатывает более 500 миллионов поисковых запросов ежедневно, и содержит в своей базе сотни миллионов страниц, на которых содержится необходимая пользователю информация. Подобный размах обусловлен, прежде всего, поисковыми потребностями пользователей.

Для решения какой-либо проблемы, в первую очередь необходимое достаточное количество информации. Глобальными поисковыми системами пользуются люди разных возрастов, стран, мировоззрений, религий и статусов. А так как пользователями системы являются практически любые люди, то, получается, система должна быть в состоянии найти любую информацию для удовлетворения их запросов.

Но настолько масштабные поисковые системы требуют вложения огромных ресурсов, и потому на сегодняшний день в мире существует не так много глобальных поисковых систем.

В данной работе будет вестись речь о ИПС меньших масштабов. Ключевым отличием такой системы является локальное поисковое пространство. Самыми близкими примерами могут служить порталы компаний, университетов, администрации, энциклопедии, литературные и иные ресурсы. Т.е. любая совокупность страниц, которая содержит достаточное количество информации для решения каких-либо проблем пользователя данного ресурса.

С момента появления ИПС, для получения ответов, наиболее релевантных запросам пользователей, применялись различные подходы к обработке информации и представлению данных. В эволюции этих методов наблюдается тенденция перехода от чисто статистических методов к смешанным и когнитивным. Статистические методы основаны на анализе частоты встречаемости слов, их расположении и других статистических параметрах. Когнитивные же методы нацелены на понимание поискового запроса пользователя или его псевдо-понимание (имитацию понимания). Смешанные методы содержат в себе элементы обоих перечисленных подходов.

Когда речь заходит о таком термине как «понимание», первое, с чем сталкивается любой исследователь – это когнитивная лингвистика, и лингвистика в общем. А в рамках данной работы именно она позволяет рассматривать слова и словосочетания, являющиеся основой запроса пользователя, не как некоторые статические единицы, но как системы, которые обладают своей структурой, правилами построения и использования.

Лингвистический и статистический подход к анализу слов будет рассматриваться во второй главе данной работы. А в этой главе будет описана современная система словосочетаний.

Прежде всего, нужно определиться с тем, что мы будем понимать под «словосочетанием» в рамках данной работы. В трудах различных авторов можно найти следующие описания.

В большом толковом словаре, словосочетание это - «Сочетание двух или нескольких слов, объединённых грамматически и по смыслу» (Кузнецов, 2007: 156).

В новом словаре методических терминов и понятий словосочетание описывают как «соединение двух или более слов на основе подчинительных связей. С. может быть свободным и несвободным. В свободном С. сохраняются значения входящих в него слов, например: «ходить в школу», «читать книгу». Самостоятельность значения отдельных слов в несвободных С. значительно ослаблена, а иногда и утрачивается: «железная дорога», «бить баклуши». Несвободные С., имеющие одно значение (семантически монолитные), называются фразеологизмами. Фразеологизмы в процессе обучения усваиваются как неразделимые единицы языка» (Азимов, 2010: 254).

В словаре русского языка под «словосочетанием» понимается «сочетание двух или нескольких слов, объединенных подчинительной связью» (Шведова, 1992: 458). В толковом словаре русского языка словосочетание – это «два или несколько слов, связанных между собой грамматически» (Ушаков, 1940: 893).

Согласно большому академическому словарю русского языка, словосочетание – это «простейшая непредикативная, в отличие от предложения, единица речи, которая образуется на основе подчинительной связи (согласования, управления, примыкания) двух и более слов» (Герд, 2013: 279).

Помимо прочего, существует следующее определение - «Словосочетанием (phrase) называется группа синтаксически связанных знаменательных слов в составе предложения, которая сама не является предложением» (Лядова, 2004: 121).

Проанализировав приведённые определения, можно выявить главные черты, характерные для словосочетаний:

Во-первых - наличие более одного слова в своём составе. Но при этом словосочетание не является предложением и в подавляющем большинстве случаев содержит два или три слова.

Во-вторых - наличие связи между словами.

И если с количеством слов в словосочетании всё достаточно прозрачно, то связь между словами требует более внимательного изучения. Для этого обратимся к литературе, посвященной данной тематике. Согласно Ю.В. Лядовой, «по типу синтаксической связи, т.е. по характеру отношений между компонентами, словосочетания можно разделить на подчинительные, сочинительные и предикативные (следуя теории Л.С. Бархударова). Такой принцип деления является довольно распространенным, однако большинство исследователей склоняется к тому, чтобы считать подчинительную связь одним из основных признаков словосочетания. Именно этот тип синтаксической связи признается многими языковедами решающим критерием для отнесения определенного соединения слов к словосочетанию, как к бинарному построению. Если подчинительные соединения слов по самому способу своего построения всегда бинарны, т.е. двучленны, то сочинительные соединения слов далеко не всегда бинарны, т.е. могут включать в себя более двух компонентов.» (Лядова, 2004: 123).

1.2 Средства связи элементов словосочетания

Кроме грамматической, или вернее, синтаксической, сочетаемости слов, есть и иная сочетаемость — фразеологическая. Нужно иметь ввиду, что слова способны взаимодействовать в разных видах связи друг с другом ещё и через их лексическую семантику. К примеру, синтаксически (грамматически), из словоформ «(a) mouse», «(a) book», «(to) catch», «(to) read», возможно сформировать четыре разных словосочетания: «to catch a mouse», «to read a mouse», «to catch a book», «to read a book», но, лексически, два из них не являются корректными и возможными; и причина здесь, заключается, не в некотором специфичном характере этих слов, а в тех отношениях, что наблюдаются в реальном мире между теми процессами и предметами, которые обозначают эти слова (Некрасова, 2017). В итоге, фразеологическая сочетаемость служит своего рода фоном, на котором происходит более строгое, грамматическое комбинирование элементов языка.

Отдельные слова взаимодействуют друг с другом прежде всего, основываясь на их смысле. Подобное соединение слов по смыслу (их лексическому значению) возможно ввиду того, что в сознании человека существуют связи и отношения между соответствующими предметами и явлениями объективной реальности. Те или иные слова взаимодействуют между собой там, где существуют характерные связи между обозначаемыми предметами и явлениями реального мира.

Лексические или, смысловые отношения слов в некоторых случаях могут влиять не только на восприятие смысла даже при отсутствии точного грамматического отображения связи (как, например, в речи иностранцев, плохо владеющих языком), но, в том числе, и для определения и конкретизации самой грамматической конструкции.

Зачастую, лексические значения частей словосочетания или предложения, некоторым способом дополняют грамматические значения использованных способов связи.

Между грамматическими и лексическими отношениями существует однозначная зависимость; в подавляющем числе ситуаций грамматические связи между словами указывают на те отношения, что непосредственно определяются из лексической семантики сочетающихся слов.

В то же время, нельзя упускать из виду, что грамматические и лексические отношения между словами являются двумя самостоятельными типами связи. Ввиду этого, нередко грамматическое и лексическое понимание взаимодействия слов друг с другом может в той или иной мере различаться.

Ввиду вышесказанного, при изучении синтаксического строя языка, имея ввиду различные лексические отношения между словами, стоит, в то же время осуществлять строгое разграничение между ними и связью слов на основе специальных грамматических средств.

1.2.1 Порядок слов как грамматическое средство связи элементов словосочетания

В английском языке, где система словоизменительных форм, как минимум, у существительных, прилагательных и местоимений, не настолько прогрессивна, очередность слов приобретает исключительную значимость. В английском языке распространено контактное расположение – такое, где слова, связанные по смыслу, располагаются рядом. И, соответственно, в случае присутствия нескольких определений к одному и тому же слову, то, что обозначает наиболее весомый признак определяемого и наиболее сильно с ним связано, стоит ближе к определяемому слову, как, например, в

словосочетании «large black eyes» (The BBC). Таким образом, расположение слов в предложении показывает степень их связи друг с другом: чем она сильнее, тем более тесным является взаимное расположение слов.

В английском языке, помимо контактного расположения слов, распространено также и дистантное расположение - такое, где слова, которые связаны по смыслу, не располагаются рядом друг с другом. В этом случае, очередность следования слов в предложении является особо регламентированной. Например, в предложении «That's the man I was speaking of» - связанные по смыслу слова «man» и «of» достаточно отдалены друг от друга, но такое их дистантное расположение возможно лишь благодаря тому, что слово «of» используется не в любом, а в строго определенном месте в предложении — не где-нибудь в начале или середине, а в самом его конце, где невозможно связать это слово с любым другим существительным (Ичкинеева, 2009).

Если говорить о целях использования порядка слов в предложении в общем, то здесь можно выделить три основные функции: грамматическую, заключающаяся в выражении некоторых синтаксических отношений. Экспрессивно-стилистическую. И выражение лексического подлежащего и лексического сказуемого.

В некоторых ситуациях порядок слов преследует несколько целей одновременно, но одна из них при этом всегда выходит на первый план.

В английском же языке, строгие грамматические ограничения изменения порядка слов приводят к тому, что возможности его применения для иной цели, кроме грамматической, ощутимо ограничены. В русском языке для оживления речи и придания ей характера спокойного повествования можно сравнительно свободно изменять порядок слов; в английском языке такого делать практически нельзя, так как существует риск разрыва синтаксических связей между словами. Но все же и в английском языке порядок слов, как описывалось выше, способен преследовать и иные, не грамматические цели. Также стоит отметить, что неосуществимость

свободного порядка слов в английском предложении зачастую крайне преувеличивается.

Возвращаясь к описанию грамматической функции порядка слов, стоит отметить следующее:

Во-первых, нужно обратить внимание на применение порядка слов для разделения между подлежащим и прямым дополнением. Потому что разграничение между именительным и объектными падежами в английском языке осуществляется только у личных местоимений. Но, иногда, даже и у данной категории слов оно не является достаточно четким, так как местоимения «it» и «you» в упомянутых падежах аналогичны по звучанию.

Во-вторых, благодаря твердому порядку слов распознаются, также, прямое и косвенное дополнения, как, например, в предложении «I send the boy a book» (The Daily Mail). Косвенное дополнение, как правило, располагается сразу за глаголом, к которому оно относится, и ввиду того, что прямое дополнение тоже стремится к глаголу, позиция косвенного дополнения становится промежуточной между глаголом и прямым дополнением. Но всё же, возможно нарушение этого соотношения, потому как прямое дополнение способно переходить в самое начало предложения: «That Musk told his engineers» (The Forbes). Вследствие этого, правильнее будет определить позицию косвенного дополнения только по отношению к глаголу.

В-третьих, порядок слов в английском языке, помимо прочего, играет важную роль и в нахождении отношения между определением и определяемым. Подобную роль порядок слов выполняет и в других языках, но характерным именно для английского является то, что в нем в таком случае порядок слов в значительной степени выступает самостоятельно, не будучи осложненным иными факторами и выступая в виде единственного грамматического средства для этого типа отношения. Это обусловлено тем, что согласование в английском языке практически целиком утрачено. Изменение по числам английских указательных местоимений согласно

форме числа определяемого ими существительного («these books» и «those books») служит почти что единственным примером согласования определительных слов.

Экспрессивно-стилистическая функция порядка слов обусловлена существующей потребностью выделения в речи какого-либо слова, чтобы таким образом показать то, на что стоит обратить особое внимание: русск. «Завтра концерт». В таких ситуациях акцентируемое слово выделяется интонационно и при помощи сильного ударения, но, кроме этого, применяется также и изменение порядка слов: подчеркиваемое слово выходит на первое место в предложении. Подобное расположение слов обеспечивает всё высказывание соответствующей экспрессивностью, не искажает его значение.

1.2.2 Морфологические средства связи слов в словосочетании

В современной английской синтаксической системе достаточно распространено также и связывание слов через их формы. Как раз ввиду сравнительной скудности морфологических форм в английском языке, существующие формы несут достаточно значительную нагрузку.

Здесь стоит отметить, что бедность современной английской морфологической системы словоизменения нередко бывает сильно преувеличена. Причина заключается в том, что английский язык не так беден грамматическими формами, как характерен отсутствием разнообразия звучания словоизменительных суффиксов.

Имеется два основных способа связи слов через их формы: согласование и управление. В достаточной степени эти способы связи слов распространены и в современном английском языке.

Как правило, в большинстве лингвистических трудов, под согласованием обычно понимают некоторое соединение слов, в котором одно из них точно повторяет, заимствует форму другого.

Стоит обратить внимание, что так называемые связи между словами в предложении и речи, с одной стороны, основываются на реальных (или рассматриваемых как реальные) отношениях, а с другой, только изображают эти последние.

Согласование нельзя понимать, как элементарное согласование форм слов; согласование является не копированием формой одного слова формы другого слова, а согласное выражение того же самого формами соединенных слов.

В современном английском языке существуют следующие случаи согласования:

Между определением и определяемым. Такой тип согласования располагает достаточно ограниченной областью применения. Прежде всего, здесь исключается согласование по линии рода. В английском языке отсутствует и согласование по линии падежа, а согласование в числе сохраняется только в двух случаях:

Во-первых, между указательными местоимениями и определяемыми ими существительными: «this dog» - «these dogs», «that dog» - «those dogs».

Во-вторых, некоторое подобие согласования существует при соединении неопределенного артикля с существительным, так как во множественном числе неопределенный артикль перед существительным не употребляется: «a dog», но «dogs».

Согласование происходит между подлежащим и сказуемым. В этом случае согласование осуществляется по линии лица и числа;

Стоит обратить внимание, что согласование между подлежащим и сказуемым в принципе, а, тем более в английском языке, является значительно более свободным по сравнению с согласованием между

определением и определяемым. Зачастую обозначаемый подлежащим объект по-разному воспринимается формами подлежащего и формами сказуемого.

Между формами глагола в главном и придаточном предложениях. Особенно в английском языке отличается согласование времен (Sequence of Tenses), являющееся согласованием форм глагола в придаточном предложении с формами глагола в главном предложении: например, «He said he was not guilty» (The Guardian).

Под управлением, как правило, подразумевается использование некоторой падежной формы подчиненного слова, где она зависит не от оформления подчиняющего слова, а от его лексического содержания.

Тем не менее, описанное традиционное определение управления не является полным и требует некоторых дополнений.

Во-первых, стоит отметить, что управление - это не устойчивая, раз и навсегда определенная зависимость формы одного слова от лексического значения другого слова. Эта зависимость способна меняться согласно синтаксической функции управляемого слова. Таким образом, при анализе того или иного случая управления стоит обращать внимание не только на управляющее слово, но и на синтаксическую функцию управляемого слова.

Во-вторых, нужно учитывать, что, кроме управления, существует еще и свободное употребление формы, которое зависит только от её содержания. То есть, ту или иную грамматическую форму возможно употреблять в языке в определенном контексте более или менее свободно. Разница в уровнях свободы применения форм выражается также в том, что, при наличии возможности выбора форм для подчиненного слова, только одна из этих форм имеет сильное связующее значение и характеризует данное слово по линии связи его с управляющим словом, тем временем другая форма, используясь более самостоятельно и независимо, дает слову сравнительно более изолированную характеристику.

Стоит обратить внимание, что управление крепко связано с представлением о падеже. Управление существует только там, где имеется

падежная система. А если говорить об английском языке, то там, ввиду ограниченного количеством падежей, система управления значительно разрушена. Если в русском языке из целого перечня возможностей глагол останавливается на одной, то в английском языке глагол подобных возможностей не имеет. В современном английском языке вопрос о выборе падежа в принципе снимается. И действительно, если личные и некоторые вопросительные местоимения и располагают двумя падежами (именительным и объектным), то один из них (именительный) не способен формировать приглагольное дополнение; в данной ситуации он выражает независимость предмета и потому совсем исключается из системы управления; поэтому, в ситуациях, когда местоимение находится в зависимом положении, оно используется в объектном падеже. В системе существительных также существует только два падежа: общий и притяжательный; но притяжательный падеж существительного не способен выступать в качестве падежа управления, так как он никогда не служит падежом дополнения. И даже в случае, когда притяжательный падеж участвует в формировании формы слова, идущего за глаголом (например, «It's my book and that is my brother's. Take my brother's, it is more interesting»), притяжательный падеж не зависит от глагола; он является полностью самостоятельным и имеет соответствующее ему атрибутивное значение относительно того слова, которое упоминалось ранее (в данном случае — «book»). Во всех ситуациях притяжательный падеж используется в совершенно конкретном значении: для описания расстояния, принадлежности, длительности во времени, меры и т. п. То есть, каждый раз он используется согласно своему значению, в установленной для него семантической сфере. Ввиду этого, нельзя утверждать, что в сочетании «(the) child's toy» слово «toy» управляет притяжательным падежом, так как любая ограниченность или идиоматичность в характере отношения падежной словоформы «child's» со словом «toy» отсутствует: «child's head», «child's voice» и т. п., где абсолютно на тех же правах, что и «toy», используются

слова «head» и «voice». Это значит, что, так как все сочетания подобного рода устанавливаются общим значением притяжательного падежа, в перечисленных сочетаниях мы наблюдаем не управление, а самостоятельное, независимое использование формы, обусловленное только ее смыслом.

Так или иначе, в следующих двух ситуациях присутствует известное подобие того, что, как правило, называется управлением:

Как известно, достаточно много современных английских глаголов вообще не способны сочетаться с зависимым существительным или местоимением. К ним относятся такие, как «stand», «go», «sit» и др., отмечаемые в лингвистической литературе термином «непереходные» или «интранзитивные». Получается, в современном английском языке имеется две формулы:

а) «Verb + noun» в общем падеже или местоимение в объектном падеже.

б) «Verb + null» (т. е. отсутствие существительного или местоимения в каком-либо падеже).

Вопрос об управлении возникает, также, в отношении применения предлогов. Предложное управление есть управление двухэтапное: глаголу нужен известный предлог, а предлог имеет падеж. Но, так как выбора падежа в английском языке после предлога нет, остается только одна ступень — выбор предлога: «to look at the book», «to look for the book», «to look through the book» и т. п. В результате, управление в современном английском языке переходит из области морфологии словоизменения в область служебных слов - предлогов.

1.2.3 Роль служебных слов в установлении связи между элементами словосочетания

В современном английском языке ввиду относительной неразвитости механизма словоизменения, особое положение занимают служебные слова. Наряду с порядком слов, служебные слова служат основным способом связи слов в современном английском языке и делятся на две основные категории: предлоги и союзы.

Предлоги являются промежуточными словами, использующимися для обозначения отношения между предметом и предметом, предметом и признаком или предметом и процессом. Здесь стоит обратить внимание на два момента:

Во-первых, у предлога, служащего связующим словом, есть некоторое грамматическое значение, которое проявляется внешне через специфику его синтаксической сочетаемости. Это грамматическое значение, как уже было сказано выше, заключается в обозначении зависимости предмета от предмета, предмета от процесса или предмета от признака. Например, в предложении «I look at him», «him» является не самостоятельным предметом, а, своего рода, «предмет смотрения», т. е. поставленным в некоторую зависимость от «look», хотя она здесь является сугубо формальной, грамматической. Предлог указывает на отношения между предметами или явлениями, зависимость одного предмета от другого предмета или явления и акцентирует внимание на том, что вводимое предлогом слово и обозначенный им предмет не являются главными в ситуации.

Во-вторых, предлог исполняет роль связующего слова, и, как любое слово, он обладает определенным лексическим значением. Временами оно выступает крайне ярко, как, например, у предлогов перед глаголами, описывающими положение в пространстве: «He sat in a tree» и «He sat under a tree», где от значения предлогов зависит смысл всего предложения: выбор

предлога in или under влияет на осмысление связи между деревом и человеком, при том, что грамматическая конструкция в обоих ситуациях одинаковая.

В некоторых случаях лексическое значение предлога может слегка тускнеть или вовсе становиться почти незаметным. Например, в предложении «It depends on the following» (The Independent) лексическое значение предлога «on» еле уловимо; оно будто затмевается смыслом глагола. Предлог «on» исполняет здесь прежде всего свою связующую функцию: он необходим лишь затем, что глагол «depend» управляет дополнением только при помощи этого предлога. Использование «on» следом за глаголом «depend» стало традиционным и сугубо формальным, хотя когда-то оно и основывалось на полном лексическом смысле предлога.

Таким образом, сочетания глаголов с предлогами в современном английском языке можно подразделить на три основные категории:

- Максимально свободные ситуации вида «to sit on», «to sit under», «to sit over», «to sit in», «to sit behind», «to sit between» и т. п.
- Максимально идиоматичные ситуации вида «to depend on», «to believe in» и т. п.
- Промежуточные ситуации вида «to look at» и т. п., где у предлога раскрывается лексическое значение, соответствующее его сочетаемости с определенной группой семантически близких глаголов.

Стоит обратить внимание и на специфическую для английского языка связь предлога и наречия. В английском языке в большом количестве ситуаций происходит совпадение предлога и наречия, как, например, в «this room», где in - предлог, и в «Come in!», где то же самое «in» исполняет роль наречия. Ввиду этого в английском языке предлоги являются более самостоятельными, а самостоятельность наречий, напротив, уменьшается. Английский предлог является в некотором роде ослабленным наречием — наречием, исполняющим связующую функцию. Такие слова точнее всего называть предложными наречиями. И такие единицы, как «in»-предлог и

«in»-наречие, будут являться одним и тем же словом, одной и той же частью речи, только по-разному функционирующую в различных случаях своего использования.

Таким образом, между функциями наречия и предлога в современном английском языке осуществляется непрерывный взаимопереход, и в итоге одни формы связи могут преобразовываться в другие.

В результате наблюдается переосмысление грамматических конструкций без трансформаций в лексике: в активной конструкции «at» связывается с существительным (или местоимением), образуя предложное дополнение (at him); в пассивной конструкции аналогичное «at» играет роль наречия и формирует с глаголом тесный комплекс (to be laughed at).

Наряду с предлогами, к связующим словам относятся также и союзы. Они используются как для соединения отдельных слов и групп слов в границах предложения, так и для связи предложений. Но с той разницей, что союзы способны связывать разные части речи. По своему смыслу союзы делятся на сочинительные («and», «also», «but» и др.) и подчинительные («since», «as», «till», «because», «if», «than» и др.).

По своей структуре союзы делятся на выступающие в предложении в виде одного отдельного слова, и союзы, выступающие в виде соотносительных пар слов («either ... or», «as ... as» и др.). Вторые по своему смыслу являются сочинительными союзами. Они используются и внутри предложения для соединения его частей и между частями сложных предложений, формируя сложносочиненные предложения. Подчинительные же союзы, как правило, соединяют только части сложноподчиненного предложения, так как подчинительная связь внутри предложения чаще осуществляется с помощью предлогов.

Связующую роль в предложении могут играть также и слова, являющиеся самостоятельными членами предложения. В первую очередь, здесь стоит упомянуть относительные местоимения и наречия. Они, как предлоги и союзы, имеют связующее значение, и, благодаря этому ему,

способствуют созданию связи между частями предложения. Больше всего относительные местоимения и наречия похожи на подчинительные союзы. Эта близость так велика, что во многих ситуациях возможен переход из одной категории слов в другую. Обычно относительные местоимения и наречия используются для введения определительных придаточных предложений

Таким образом, местоимения и наречия можно разделить на следующие категории:

1. Относительные слова, обладающие сугубо атрибутивным значением в предложении. Примером может служить следующий контекст: «He visited his father who lived in London» (The New York Times).

2. Конденсированные относительные слова; в таком случае слово, к которому относится придаточное предложение, опущено, из-за чего относительное местоимение и наречие получают большую степень самостоятельности: «I don't understand what you say» (The Reuters).

3. Связующие местоимения и наречия, располагающиеся ближе к союзу, поскольку вводимое ими предложение не связано ни с каким конкретным словом: «It happened when he came to Moscow» (The Washington Post).

4. Вопросительные местоимения, которые вообще не играют связующую роль: «He asked what I was going to do» (The Reuters).

В английском языке часто в ситуациях, где служебные слова кажутся необходимыми, наблюдается их отсутствие. Такое соединение слов именуется асиндетическим, или бессоюзным. Очень часто наблюдается пропуск союза «that»: как, например, в «Tell Tom that I want to see him» (The Independent) и «Tell Tom I want to see him» - и оба имеют одинаковое значение.

Но при этом отсутствие союза не осложняет понимания смысла предложения. Суть заключается в том, что по-видимому, отсутствие «that» провоцирует внезапное столкновение словоформ, очевидно не сочетаемых

вместе: «Tell Tom I want to see him», где располагаются рядом словоформы «Tom» и «I». Это необычное сочетание словоформ, по-иному не соединяемых друг с другом, является равным союзу «that». Поэтому отсутствие союза является здесь значащим.

Пропуск связующего слова происходит и в ситуации, когда для выражения связи употребляются относительные местоимения. Так, например, наравне с конструкцией «That's the man whom (who) I saw yesterday» (The Guardian), используется также и конструкция «That's the man I saw yesterday» (The BBC).

Отдельным случаем являются предложения с «there is» и «it is». В таких конструкциях возможно отсутствие относительного местоимения, несмотря на то, что оно служит подлежащим придаточного предложения: «It was haste killed him» вместо «It was haste that killed him». Здесь опущение местоимения вызвано обыденностью выражений «there is» и «it is», их особенностью, тем, что они являются в некотором роде застывшими и раз и навсегда данными выражениями».

Выводы по главе 1

Таким образом, в данной главе была рассмотрена краткая история развития информационно-поисковых систем и изучена система английских словосочетаний в современном языкознании, в частности, были рассмотрены средства связи элементов словосочетания – порядок слов, морфологические средства и служебные слова.

Для этого были выбраны и проанализированы дефиниции термина «словосочетание» в таких трудах, как «Новый словарь методических терминов и понятий» Э. Г. Азимова и А. Н. Щукина; «Большой академический словарь русского языка» под редакцией А. С. Герд; «Большой толковый словарь русского языка» под редакцией С. А. Кузнецова; «Толковый словарь русского языка» под редакцией Д. Н. Ушакова; «Толковый словарь русского языка» под редакцией Н. Ю. Шведовой.

В качестве источников словосочетаний и предложений для анализа использовались такие англоязычные издания, как «The Reuters», «The New York Times», «The BBC», «The Forbes», «The Independent», «The Washington Post», «The Daily Mail» и «The Guardian».

ГЛАВА 2. Разработка информационно-поисковой системы локального пространства

В Главе 2 описывается процесс разработки информационно-поисковой системы локального пространства, в частности, модулей каталогизации, индексации и трансляции.

2.1. Информационно-поисковая система локального пространства

Информационно-поисковая система локального пространства – это система, осуществляющая поиск информации по некоторому локальному информационному пространству. Такая система, будучи правильно сконструированной, может легко интегрироваться в работу любого ресурса и решить проблему поиска по его содержимому.

Работа информационно-поисковой системы, изображенная на рис. 2.1, начинается, прежде всего, с пользователя и его запроса.

Рис. 2.1. Процесс работы ИПС

При формулировке запроса к ИПС, желая получить наиболее релевантный ответ, пользователь старается как можно точнее указать интересующий его объект, процесс или явление, а также параметры или характеристики, описывающие его, как на рис. 2.2.

Wild cats
Wild cats online
Wild cats Scotland
Wildcats
Wildcats perth

Рис. 2.2. Пример поисковых запросов пользователей

Далее запрос обрабатывается модулем трансляции ИПС. Модули же каталогизации и индексации осуществляют работу вне запроса пользователя, и делают возможным работу модуля трансляции. Архитектура и алгоритмы работы всех модулей системы будут описаны далее в этой главе.

После обработки запроса пользователя, ему возвращается список страниц, содержащих указанную в запросе информацию.

То есть, по факту, информационно-поисковая система ищет не столько ответ на заданный пользователем вопрос, сколько места, где может содержаться этот ответ. И поэтому пользователь видит не какие-либо точные формулировки и информацию, а список ресурсов, отображаемых в наиболее релевантном запросу порядке.

2.2 Модуль каталогизации

Модуль каталогизации – это первый модуль информационно-поисковой системы локального пространства. Его можно увидеть на рис. 2.3.

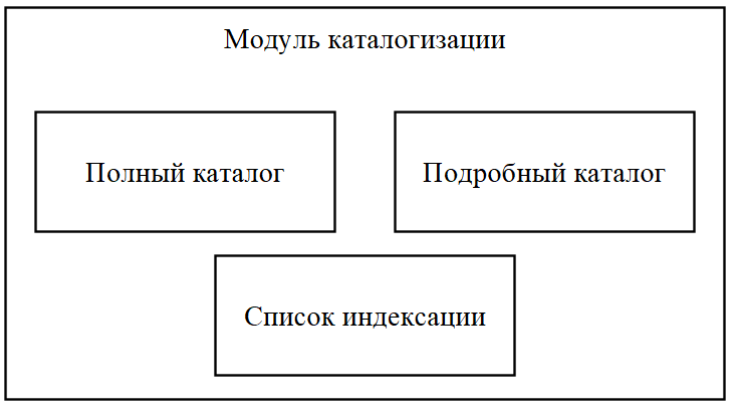


Рис. 2.3. Модуль каталогизации

Цель модуля заключается в осуществлении первого подготовительного этапа работы системы – каталогизации страниц ресурса. То есть, прежде чем система сможет искать информацию по какому-то локальному пространству, ей необходимо «знать» это пространство. «Знание» пространства выражается в наличии, исходя из названия модуля, каталога страниц ресурса, по которым и будет вестись поиск.

Но, прежде чем перейти к правилам и принципам работы каталога, нужно познакомиться с таким термином, как «контрольная сумма».

Контрольная сумма – некоторое значение, рассчитанное по набору данных путем применения определенного алгоритма и традиционно используемое для проверки целостности данных при их передаче и хранении, как на рис. 2.4. То есть, к отправляемым данным применяется какая-либо функция, например, суммирование двоек, возведенных в степени, равные порядковым номерам (в алфавите) всех символов в строке. Та же самая функция применяется к данным при их получении. Два полученных числа сравниваются, и только в случае их равенства операция передачи данных считается успешной.

Рис. 2.4. Сравнение контрольных сумм строки до и после её передачи

Одной из особенностей модуля каталогизации является то, что ключ суммы можно использовать не только традиционным образом, но и как некоторый индикатор изменяемости страницы. Т.е. страница, обладая каким-либо набором данных находится в одном уникальном состоянии, которое можно описать при помощи контрольной суммы. И изменение хотя бы одного символа в коде страницы приводит к изменению и контрольной суммы страницы, означающему переход в новое уникальное состояние. Система, сравнивая одну и ту же страницу в разные моменты времени, сравнивает и контрольные суммы страницы в разные моменты времени, что позволяет ей чётко определить подвергалась ли страница изменениям или нет, что продемонстрировано на рис. 2.5.

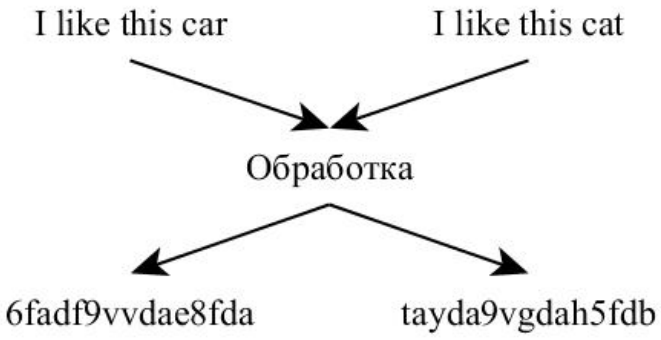


Рис. 2.5. Пример использования контрольной суммы для проверки состояния изменения текста

Помимо контрольной суммы, в раскрытии модуля каталогизации используется ещё три ключевых понятия: подробный каталог, полный каталог и список индексации. Их дальнейшее объяснение, по сути, и является описанием принципа работы всего модуля.

Подробный каталог – каталог страниц портала, содержащий некоторое количество записей, параметрами каждой из которых являются: адрес страницы, контрольная сумма страницы, и ссылки страницы.

Адрес страницы – адрес расположения страницы, по которому можно получить к ней доступ.

Контрольная сумма – контрольная сумма страницы, которая служит индикатором её изменяемости.

Ссылки страницы – это все адреса других страниц, указанные на текущей, по которым можно совершить переход.

Так как поиск начинается с главной страницы портала, она является первым уровнем подробного каталога ресурсов, увидеть который можно в табл. 2.1.

Таблица 2.1. Первый уровень каталога

Первый уровень каталога		
Страница:	Контрольная сумма:	Ссылки:
Root	4d67hfxlun9ddv	A
		B
		C
		D
		E
		…
		Z

Каждый последующий уровень подробного каталога является совокупностью страниц, на которые можно перейти с предыдущего уровня, что показано в табл. 2.2.

Таблица 2.2. Первый и второй уровни подробного каталога

Первый уровень каталога			Второй уровень каталога		
Страница:	Контрольная сумма:	Ссылки:	Сайт:	Ссылки:	Контрольная сумма:
Root	4d67hfxlun9ddv	A	A	AA	ascvva
		B		…	…
		C		AZ	dvzd
		D	B	BA	zxfvf
		E		…	…
		…		BZ	sdf
		Z	C	CA	zsc
				…	…
				CZ	szcv
			…	……	…
				…	…
				……	…
			Z	ZA	dvzd
				…	…
				ZZ	xfv

Подобный способ формирования подробного каталога создаёт некоторую иерархию сайтов ресурса, изображенную в табл. 2.3., в полной мере представляющей искомое пространство для поиска информации. Локальным пространство становится ввиду наложения ограничений на его формирование. Т.е. в формируемую иерархию не включаются ссылки на сторонние ресурсы. Каталог может содержать в себе сколько угодно много уровней, как в табл. 2.3. Их количество зависит от размеров ресурса и возможностей разработчиков системы.

Таблица 2.3. Трёхуровневый каталог ресурса

Первый уровень каталога			Второй уровень каталога			Третий уровень каталога		
Страница:	Контрольная сумма:	Ссылки:	Сайт:	Ссылки:	Контрольная сумма:	Сайт:	Ссылки:	Контрольная сумма:
Root	4d67hfxlun9ddv	A	A	AA	ascvva	AA	AAA	zddv
		B		…	…		…	fbbs
		C		AZ	dvzd		AAZ	adzxc
		D	B	BA	zxfvf	…	…	4sdfd
		E		…	…		…	sdfs
		…		BZ	sdf		…	cxva
		Z	C	CA	zsc	AZ	AZA	zddv
				…	…		…	fbbs
				CZ	szcv		AZZ	adzxc
			…	……	…	BA	BAA	4sdfd
				…	…		…	sdfs
				……	…		BAZ	cxva
			Z	ZA	dvzd	…	…	zddv
				…	…		…	fbbs
				ZZ	xfv		…	adzxc
						BZ	BZA	4sdfd
							…	sdfs
							BZZ	cxva
						…	…	zddv
						ZZ	ZZA	4sdfd
							…	sdfs
							ZZZ	cxva

При сведении подробного каталога в единый список адресов ресурса, получается полный каталог, который можно увидеть в табл. 2.4. Полный каталог формируется таким образом, что сперва идёт упоминание страниц первого уровня, затем второго, третьего и последующих уровней при их наличии. Записи страниц расположены в порядке нахождения ссылок на них на вышеупомянутых страницах.

Таблица 2.4. Полный каталог

ИД:	Адрес:	Контрольная сумма
0	Root	zddv
1	A	ascvva
2	AZ	dvzd
3	BA	zxfvf
4	BZ	sjhhg
5	BA	zsc
6	BZ	szcv
7	ZA	dvzd
8	ZZ	xfv
9	AAA	zddv
10	AAZ	adzxc
11	AZA	zddv
12	AZZ	adzxc
13	BAA	4sdfd
14	BZZ	cxva
…	…	…
210	ZZA	4sddfd
…	…	…
269	ZZZ	cxva

Полный каталог необходим для корректной работы алгоритма каталогизации, приведенного на рис. 2.6.

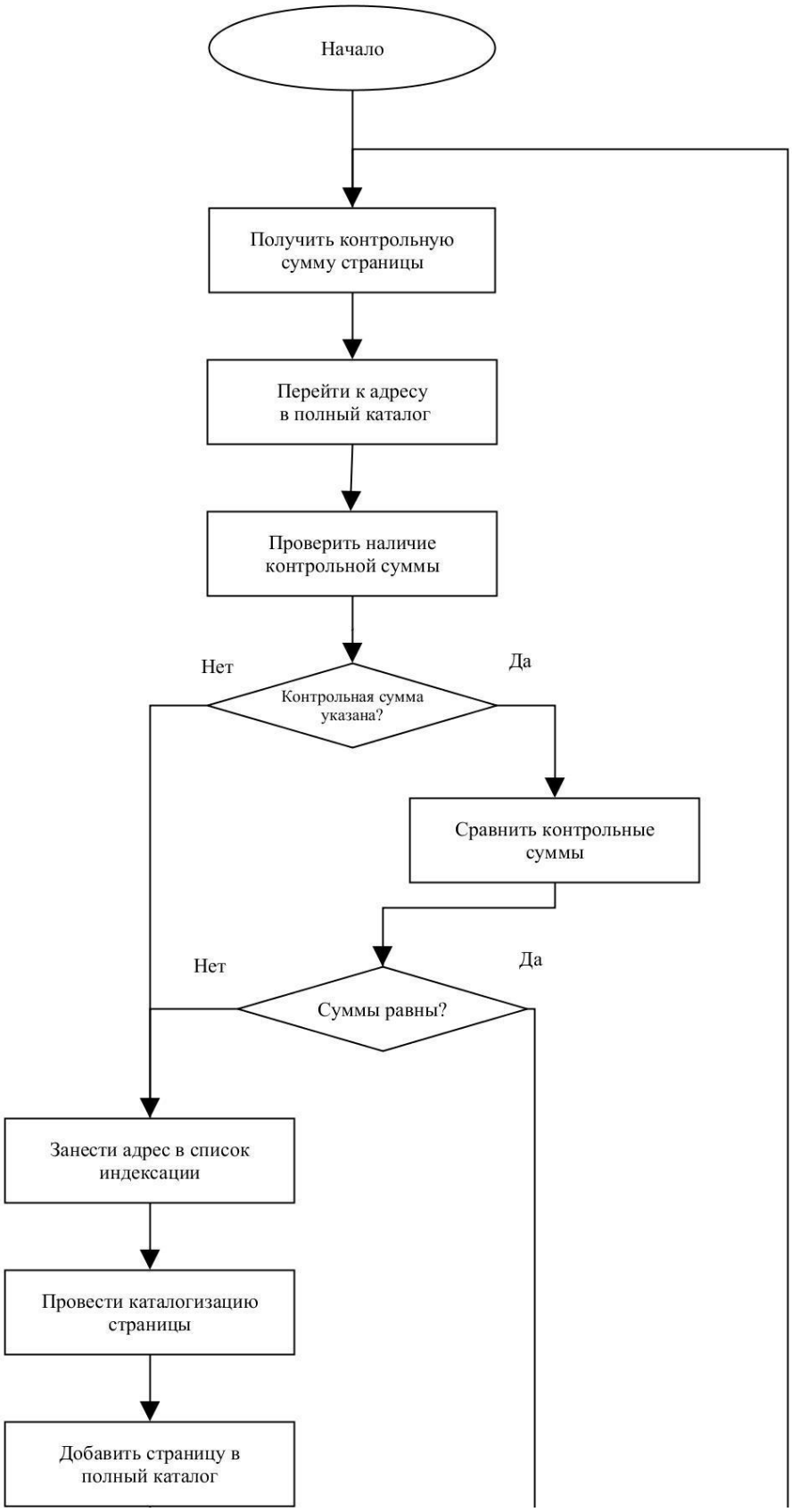


Рис. 2.6. Алгоритм каталогизации

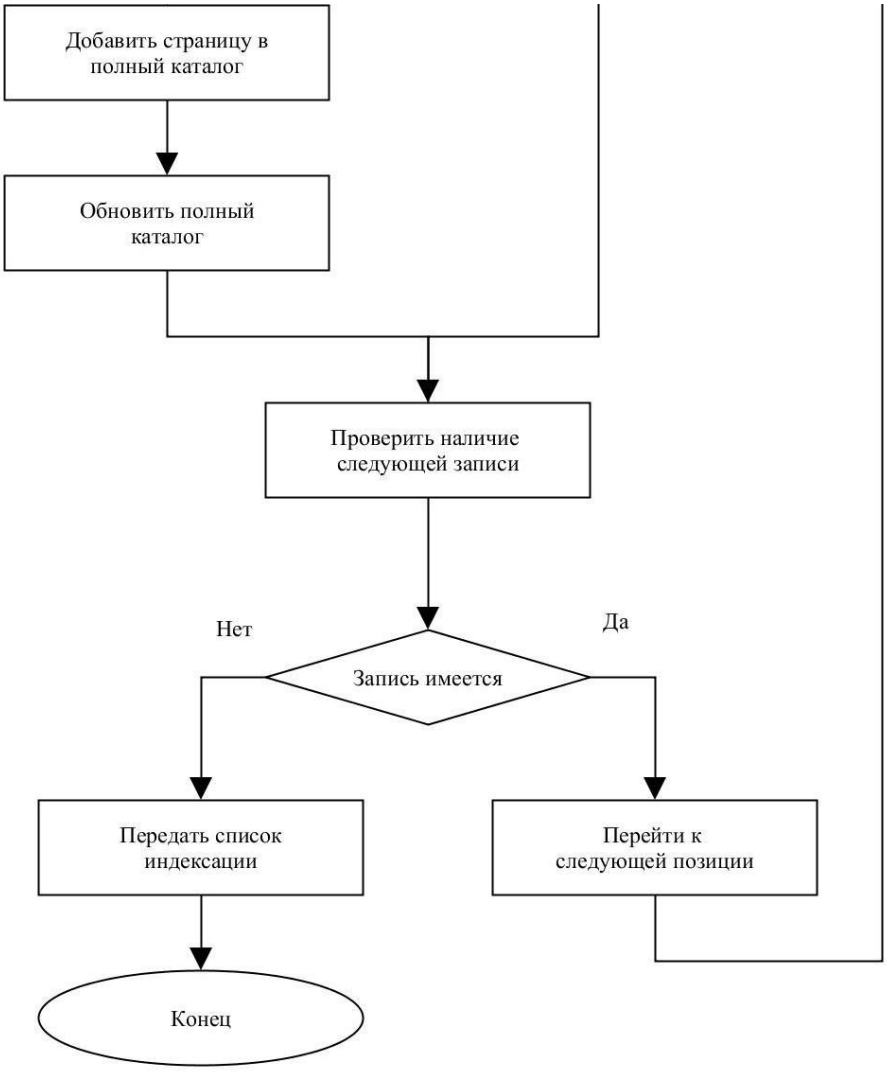


Рис. 2.6. Алгоритм каталогизации (продолжение)

В качестве входных данных алгоритм принимает последнее, актуальное состояние полного каталога. Получая контрольные суммы страниц из каталога, алгоритм сравнивает их с текущими контрольными суммами страниц, и, в случае их различия, добавляет страницы в список индексации. Также, в случае обнаружения страниц, ещё не внесённых в подробный каталог, алгоритм добавляет их, и обновляет полный каталог в соответствии с новой иерархией подробного. Это позволяет обеспечить работу алгоритма с наиболее актуальным состоянием страниц ресурса.

В качестве результата своей работы, помимо поддержания состояния каталогов, алгоритм возвращает список индексации – перечень страниц, которые изменялись с их последней обработки, либо ещё не были

обработаны. Далее список индексации передаётся второму модулю системы – индексатору. Пример списка можно увидеть в табл. 2.5.

Таблица 2.5. Список индексации

ИД:	Адрес:
111	AAA
145	ABZ
178	AZ
179	GZA
202	AZZ
874	D
875	JA
876	LAA
901	TAZ
976	QZ
977	ZZA
999	ZZZ

2.3 Модуль индексации

Модуль индексации – второй модуль информационно поисковой системы локального пространства. Его структура изображена на рис. 2.7.

Рисунок 2.7. Модуль индексации

Его цель заключается в индексировании страниц ресурса. Это является вторым подготовительным этапом работы системы. После того, как мы удостоверились в том, что система «знает» поисковое пространство, необходимо, помимо прочего, привести содержимое этого пространства в такой вид, чтобы система могла с ним взаимодействовать. На практике это означает его обработку до самых простых неделимых элементов – слов.

Получая список индексации, модуль приступает к его обработке. И первое, что модуль делает с каждой из страниц списка – это к выделению её содержимого, как на рис. 2.8.

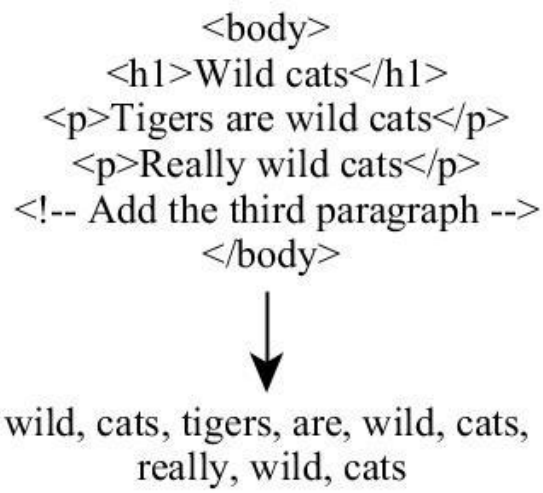


Рис. 2.8. Обработка кода

Современные веб-страницы представляют собой текстовые файлы, оформленные согласно стандартизированному языку разметки документов HTML. Но сама по себе, разметка не является данными, необходимыми для сохранения системой. Поэтому каждая из страниц обрабатывается таким образом, чтобы в результате оставались только единицы, несущие некоторый смысл для пользователя системы – слова.

Как показано на рис. 2.8, система удаляет все теги, содержащиеся в коде, а также слова, определяемые как комментарии кода (слова между <!-- и -->), и получает список слов используемых на странице. В данном случае это девять слов.

При помощи алгоритма индексации страниц, эти слова проходят обработку, и заносятся в базу данных, изображенную на рис. 2.9.

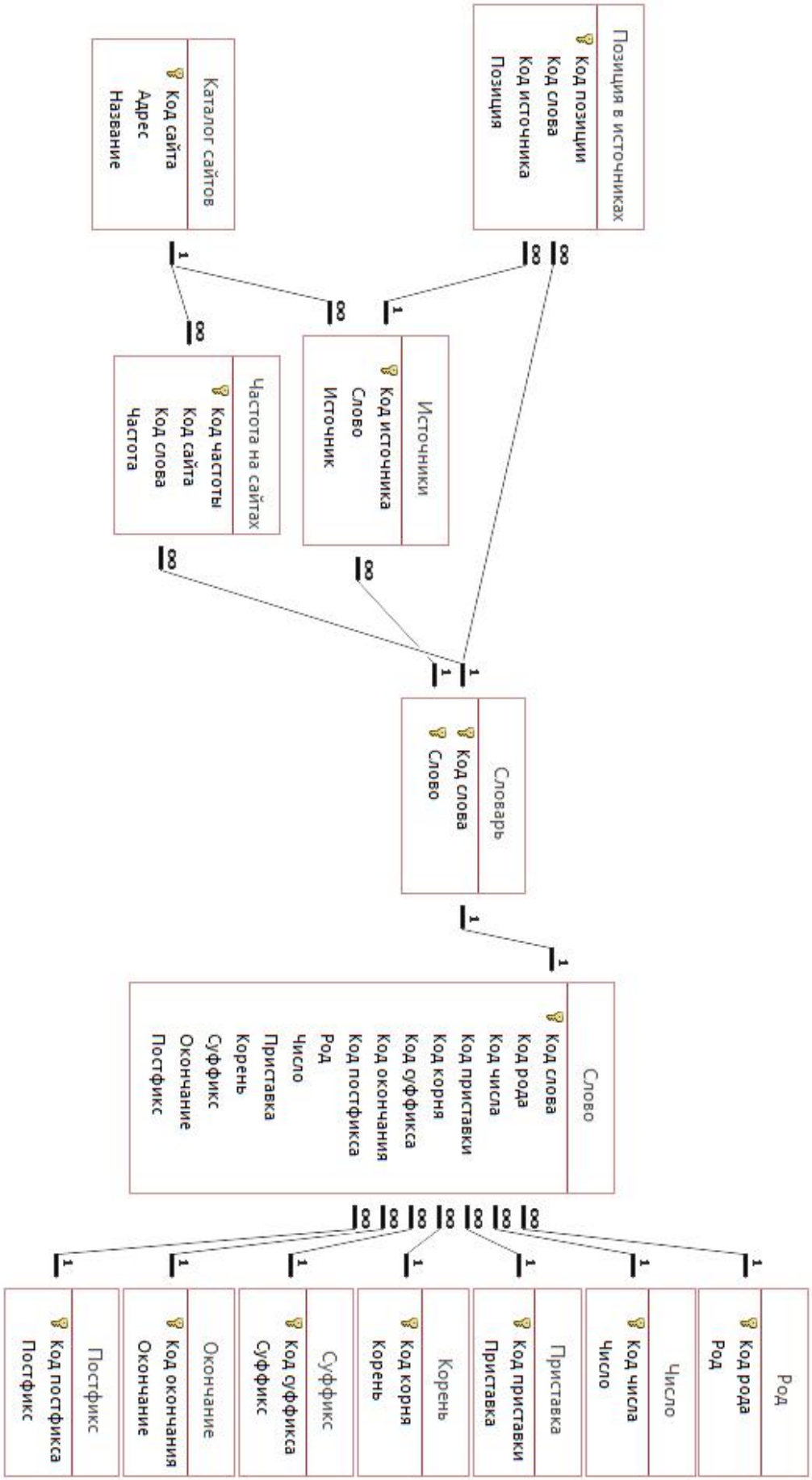


Рис. 2.9. База данных

База данных по одному из определений международных стандартов – это совокупность данных, организованных в соответствии с концептуальной структурой, описывающей характеристики этих данных и взаимоотношения между ними, причём такое собрание данных, которое поддерживает одну или более областей применения. При проектировании базы данных нашей системы, мы, помимо прочего, сталкиваемся с такими базовыми в этой области терминами, как сущность и связь.

Сущность – это некоторый объект, процесс или явление предметной области, который необходимо выделить.

Связь – это природа взаимоотношений двух сущностей предметной области.

Сущности на рис. 2.9 изображены как прямоугольники, обладающие рядом атрибутов, как реальных, так и системных. Т.е. для описания сущности в качестве атрибутов используются реальные характеристики сущности, и некоторые системные параметры, необходимые для правильного хранения данных в системе.

В первой главе затрагивался вопрос разницы статистического и когнитивного подходов к поиску информации. Проектирование базы данных модуля индексации происходит с учётом обоих подходов, и выражается это непосредственно в выделении сущностей и определении связей между ними.

Так, ядровыми сущностями базы данных являются «словарь» и «слово» в табл. 2.6 и 2.7 соответственно.

Таблица 2.6. Сущность «словарь»

Код слова	Слово
1	group
2	schedule
3	enrollee
4	unpredictable
5	reaction
6	application
7	institute
…	…

Таблица 2.7. Сущность «слово»

Код слова	Коды	Префикс	Корень	Суффикс
1	…		group	
2	…		schedule	
3	…		enroll	ee
4	…	un	predict	Able
5	…		react	Ion
…	…	…	…	…

Вторая сущность в качестве атрибутов содержит элементы результата грамматического разбора слова по составу. Первая сущность является, во-первых, элементом, соединяющим статистическую и грамматическую обработку слова (статистическая – сущности слева от сущности «словарь», грамматическая – справа на рис. 2.9), во-вторых, сущностью, согласно названию, содержащей в себе каждое из использованных на портале слов.

Статистическая обработка слова выражается в таких сущностях, как «источники», «позиция в источниках», и «частота на сайтах».

Сущность «каталог сайтов» можно считать обособленной, служащей хранилищем перечня сайтов ресурса, что можно увидеть в табл. 2.8.

Таблица 2.8. Сущность «каталог сайтов»

Код сайта	Адрес	Названия
1	bsu.edu.ru	Главная страница
2	bsu.edu.ru/schedule	Учебное расписание
3	bsu.edu.ru/documents	Документы
…	…	…

Сущность «источники» отображает расположение всех слов системы на конкретных страницах, что показано в табл. 2.9.

Таблица 2.9. Сущность «источники»

Код источника	Слово	Источник
1	1 (group)	2 (bsu.edu.ru/schedule)
2	1 (group)	3 (bsu.edu.ru/documents)
3	2 (schedule)	2 (bsu.edu.ru/schedule)
4	4 (unpredictable)	1 (bsu.edu.ru)
5	4 (unpredictable)	3 (bsu.edu.ru/documents)
6	5 (reaction)	3 (bsu.edu.ru/documents)
…	…	…

Сущность «позиция в источниках» показывает конкретные расположения слов на страницах, что отображено в табл. 2.10.

Таблица 2.10. Сущность «позиция в источниках»

Код позиции	Код слова	Код источника	Позиция
1	4 (unpredictable)	3 (bsu.edu.ru/documents)	56
2	5 (reaction)	3 (bsu.edu.ru/documents)	57
3	1 (group)	2 (bsu.edu.ru/schedule)	23
…	…	…	…

Сущность «частота на сайтах» отображает количество использований слова на сайтах ресурса, что видно в табл. 2.11.

Таблица 2.11. Сущность «частота на сайтах»

Код частоты	Код сайта	Код слова	Частота
1	3 (bsu.edu.ru/documents)	4 (unpredictable)	13
2	3 (bsu.edu.ru/documents)	5 (reaction)	5
3	2 (bsu.edu.ru/schedule)	1 (group)	29
…	…	…	…

Грамматическая обработка слово выражается в сущностях, описывающих состав слова, для английского языка – это сущности «префикс», «корень» и «суффикс». В иных случаях можно прибегать к расширенному анализу слова, по его роду, числу, падежу и другим параметрам.

Результаты грамматической и статистической обработки слов будут раскрыты далее в описании модуля трансляции системы.

Работа модуля индексации происходит согласно алгоритму индексации, изображенному на рис. 2.10.

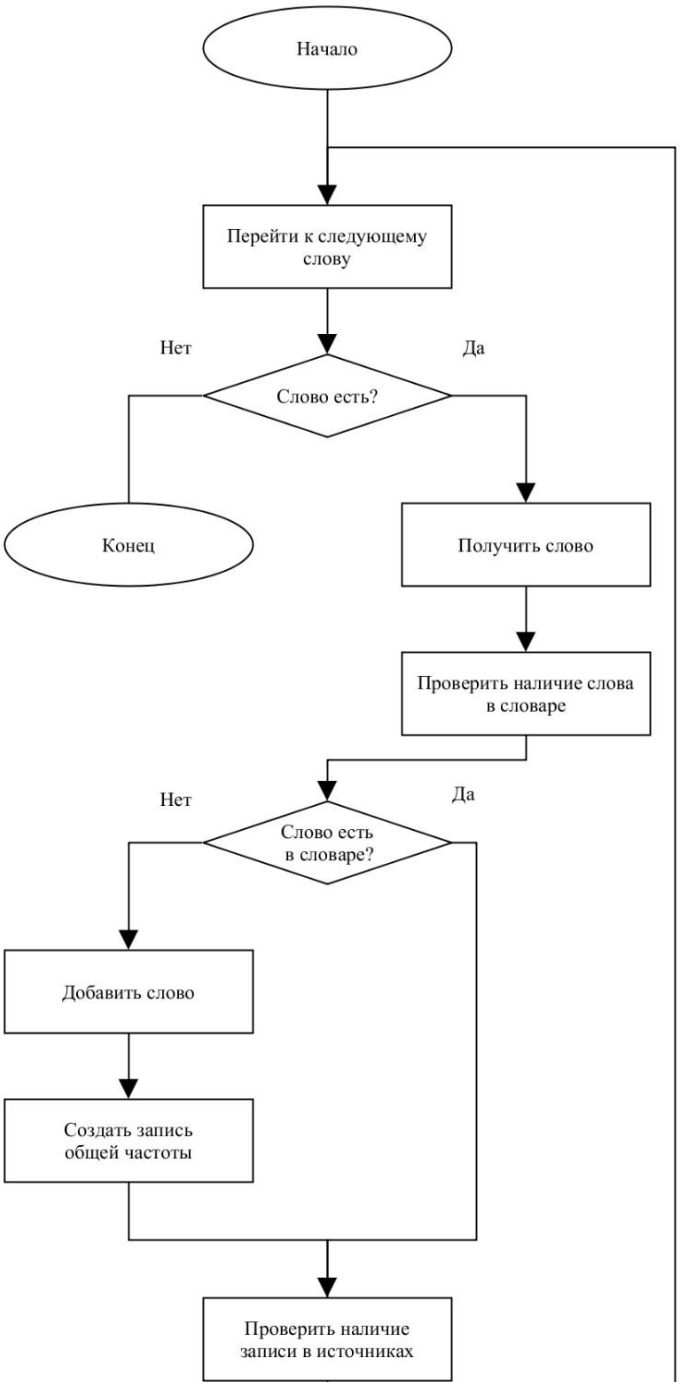


Рис. 2.10. Алгоритм индексации

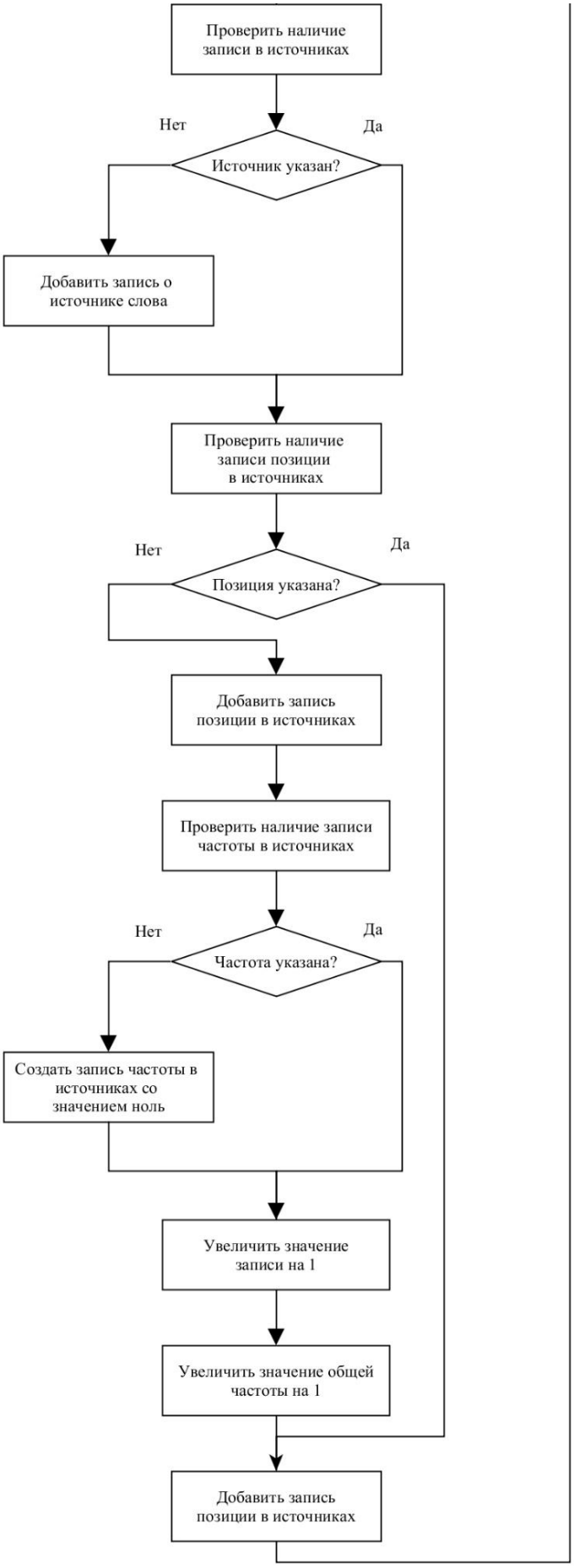


Рис. 2.10. Алгоритм индексации (продолжение)

В качестве входных данных, алгоритм принимает каждое из слов на сайте, далее он проверяет наличие этого слова в словаре, проводит статистическую и грамматическую обработки. По ходу обработки, алгоритм вносит необходимую информацию в базу данных, тем самым приводя информацию в страницах к необходимому для работы с ней виду.

На этом заканчиваются подготовительные этапы работы системы – мы «знаем» поисковое пространство, имеем базу данных всех используемых слов, и можем с ними взаимодействовать.

Запуск модулей каталогизации и индексации инициируется в двух случаях. Во-первых, они работают с некоторой периодичностью – запускаются дважды в сутки, один раз в сутки, несколько дней или неделю – этот параметр устанавливается администраторами системы. Они сами определяют оптимальный промежуток проверки обновлений контента. Во-вторых, он начинают цикл своей работы каждый раз при добавлении/удалении страниц порталы, и, тем самым, поддерживают каталог и базу данных в актуальном состоянии.

2.4 Модуль трансляции

Модуль трансляции – третий модуль информационно поисковой системы локального пространства. Он изображен на рис. 2.11.

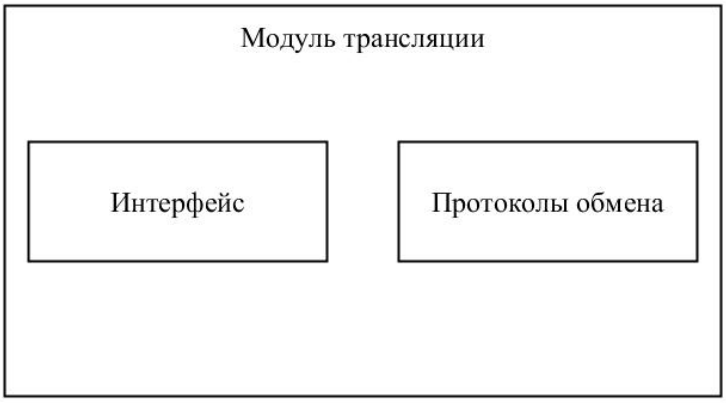


Рис. 2.11. Модуль трансляции

Цель работы модуля трансляции заключается в обеспечении возможностей общения пользователя и информационно-поисковой системы. Модуль принимает поисковой запрос пользователя, обрабатывает его, отправляет в базу данных, получает ответ, и представляет результат в понятном для пользователя виде.

На рис. 2.12 показан алгоритм работы модуля.

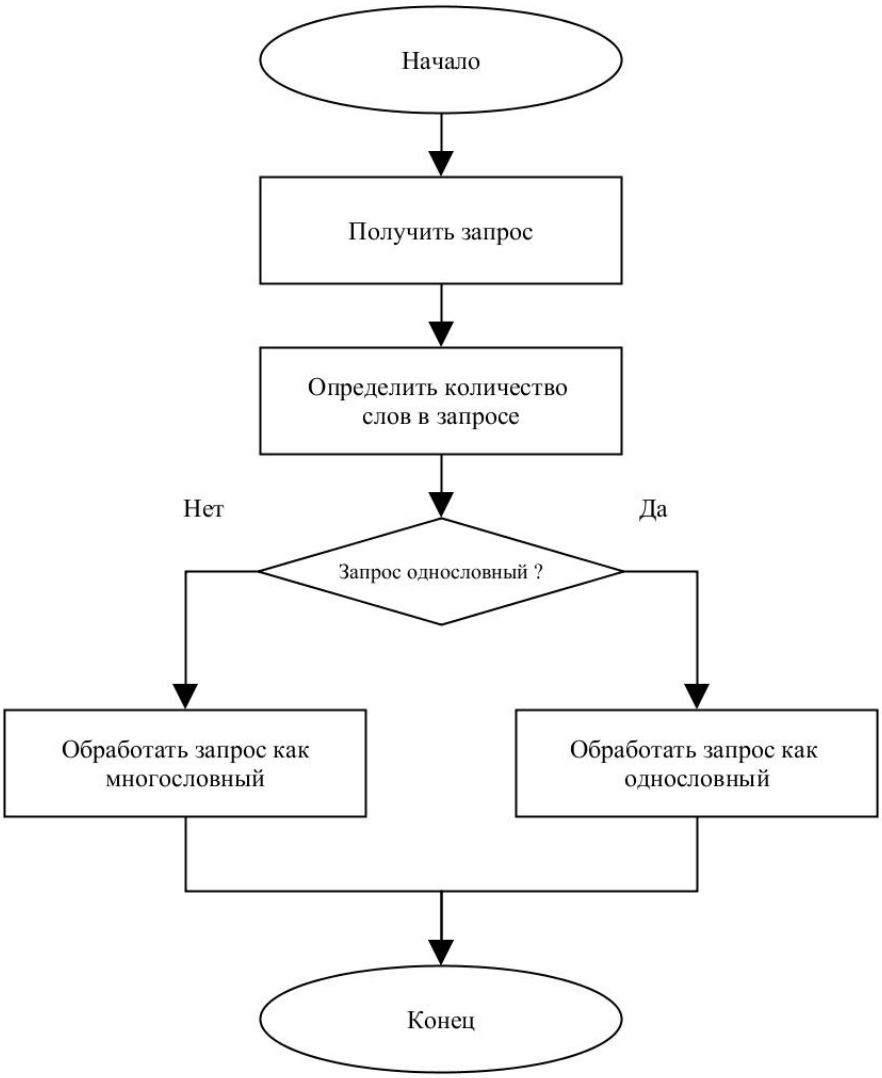


Рис. 2.12. Алгоритм работы модуля трансляции

После получения запроса пользователя, первое, что делает система – это определяет его многословность, так как обработка однословного и многословного запросов немного отличается.

В случае однословного запроса, система обрабатывает его по алгоритму на рис. 2.13.

Рис. 2.13. Алгоритм обработки однословного запроса

Система обрабатывает запрос, отправляет его в базу данных и получает ответ. Так, например, в случае, если пользователь ищет слово «reaction», в базу данных будет отправлен запрос на поиск слов, сформированных в

результате обработки первоначального запроса образом, изображённым на рис. 2.14.

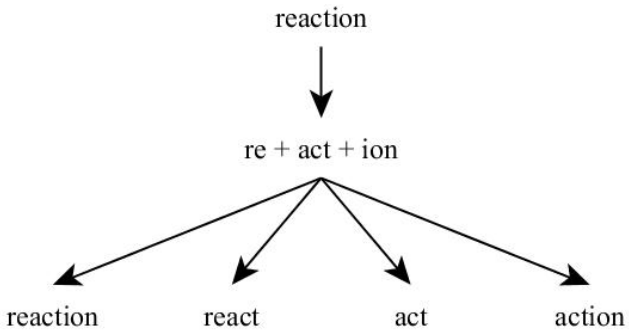


Рис. 2.14. Разбор однословного запроса

Таким образом, система отобразит результаты поиска не только искомого запроса, но и слов, схожих с ним по смыслу.

Очередность отображения результатов однословного запроса показана на рис. 2.15.

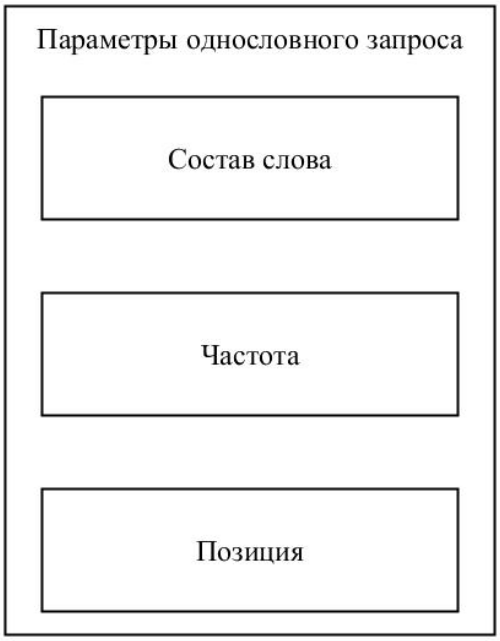


Рис. 2.15. Параметры однословного запроса

Так, прежде всего, пользователю будут показаны места употребления слов, грамматически наиболее приближенных к слову, использованному в запросе. Затем результаты будут отсортированы по частоте использования слова на сайтах – сперва будут отображены сайты, где слово используется чаще всего. И далее результаты будут отсортированы по позиции слова на сайтах, и сперва будут отображены те, где слово находится ближе к началу страницы.

В случае многословного запроса, модуль работает согласно алгоритму, изображенному на рис. 2.16.

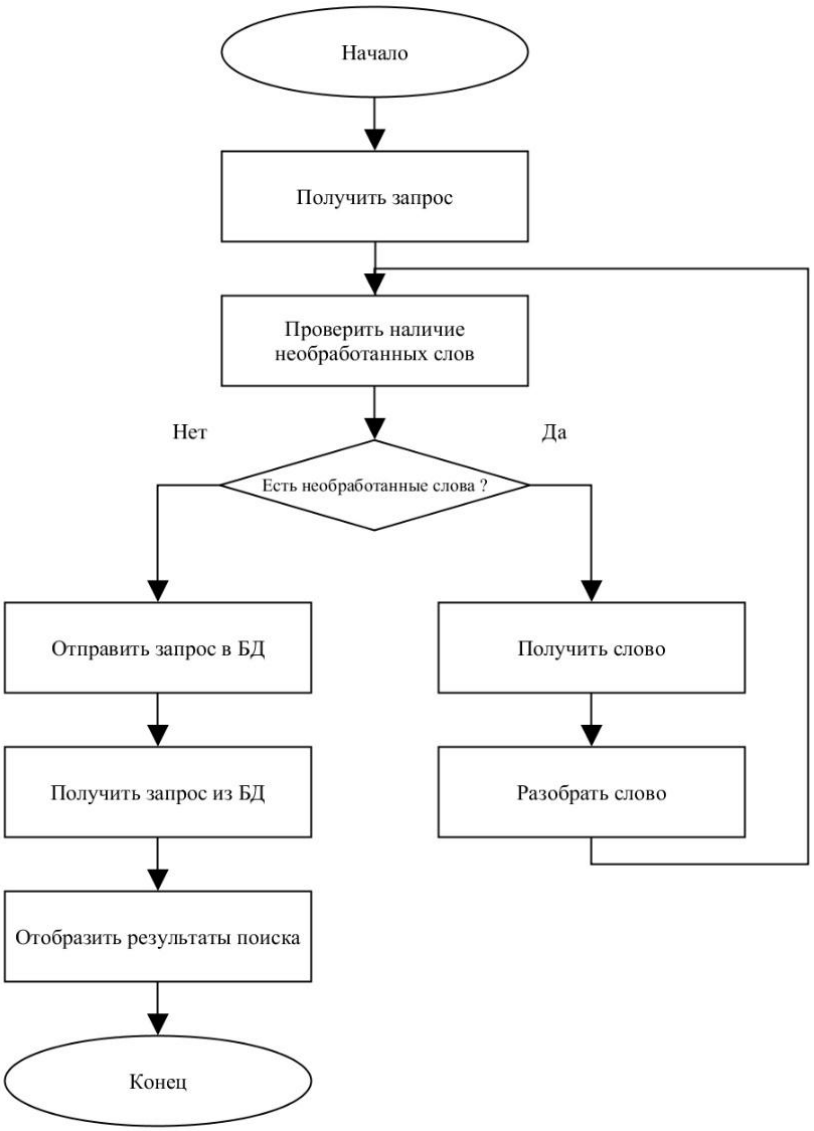


Рис. 2.16. Алгоритм обработки многословного запроса

Системы разбирает слова запроса, и составляет более сложное, по сравнению с однословным запросом, дерево возможных для поиска вариантов. Дерево вариантов изображено на рис. 2.17.

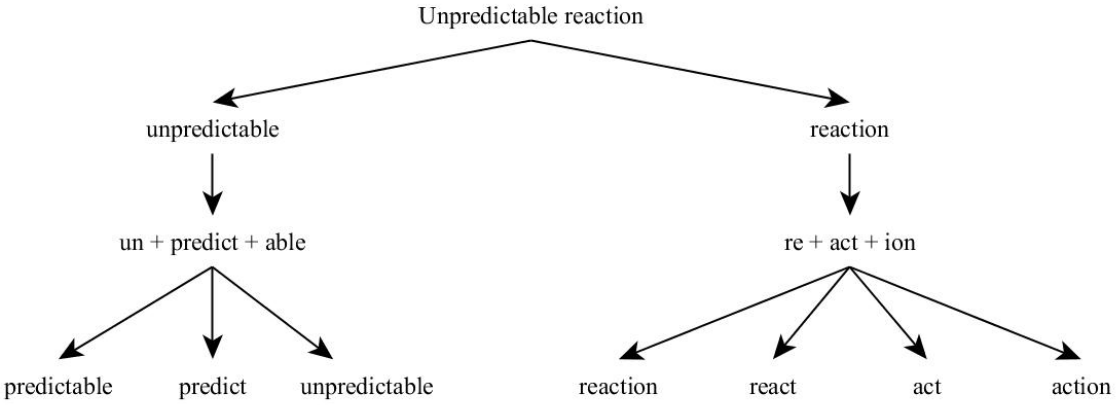


Рис. 2.17. Разбор многословного запроса

После этого модуль формирует запрос к базе данных, пользуясь результатом разбора многословного запроса. Содержание конечного запроса изображено на рис. 2.18, 2.19 и 2.20.

Прежде всего, система указывает слова изначального запроса – «unpredictable» и «reaction», что показано на рис. 2.18.

Затем система помещает в запрос исходные слова одной части запроса, и производные слова другой части запроса, что изображено на рис. 2.19.

И затем система дополняет запрос сочетанием производных слов всех частей запроса, что отображено на рис. 2.20.

На рис. 2.21 показан результат разбора многословного запроса.

unpredictable - unpredictable

predict predict

unpredictable reaction

reaction

react

unpredictable react unpredictable act unpredictable action predictable reaction predict reaction

predictable predictable

act

predictable react predictable act predictable action predict react predict act predict action

action

Рис. 2.18. Слова оригинал

Рис. 2.19. Рис.2.21. Результат разбора многословного запроса

слу

чае

мн

ого

сло

вн

ого

зап

роса, сортировка отображения результатов осуществляется в очередности параметров, изображенной на рис. 2.22.

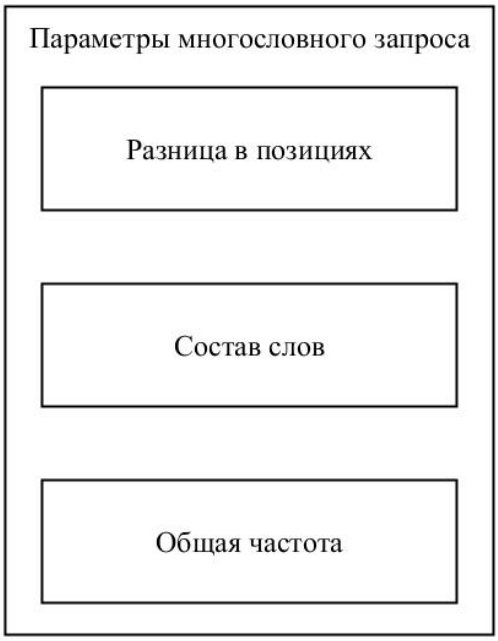


Рис. 2.22. Параметры многословного запроса

В первую очередь, пользователю отображаются страницы с наименьшей разницей в позициях слов. Далее результаты сортируются по степени идентичности изначальному запросу, в приведённом примере – это сперва «unpredictable reaction», затем группа слов «unpredictable react, unpredictable act, unpredictable action, predictable reaction, predict reaction», и потом «predictable react, predictable act, predictable action, predict react, predict act, predict action». После этого результаты сортируются по общей частоте использования слов в запросах.

В случае, если система не смогла найти ни одного результата, каждое из слов многословного запроса обрабатывается по алгоритму однословного запроса, и пользователю отображаются места использования слов в отдельности. Интерфейсы ввода и отображения результатов показаны на рис. 2.23 и 2.24.

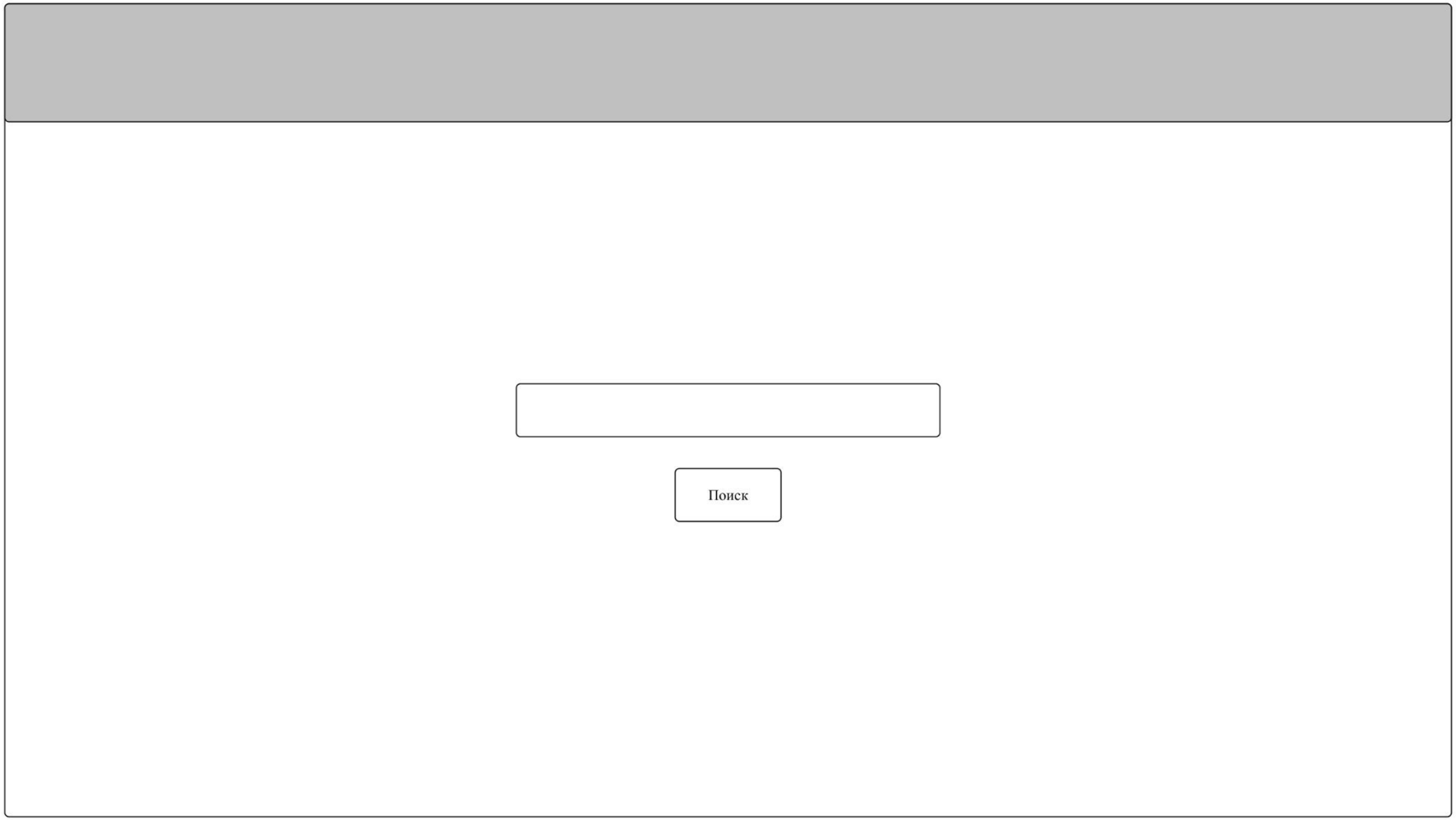


Рис. 2.23. Интерфейс ввода запроса

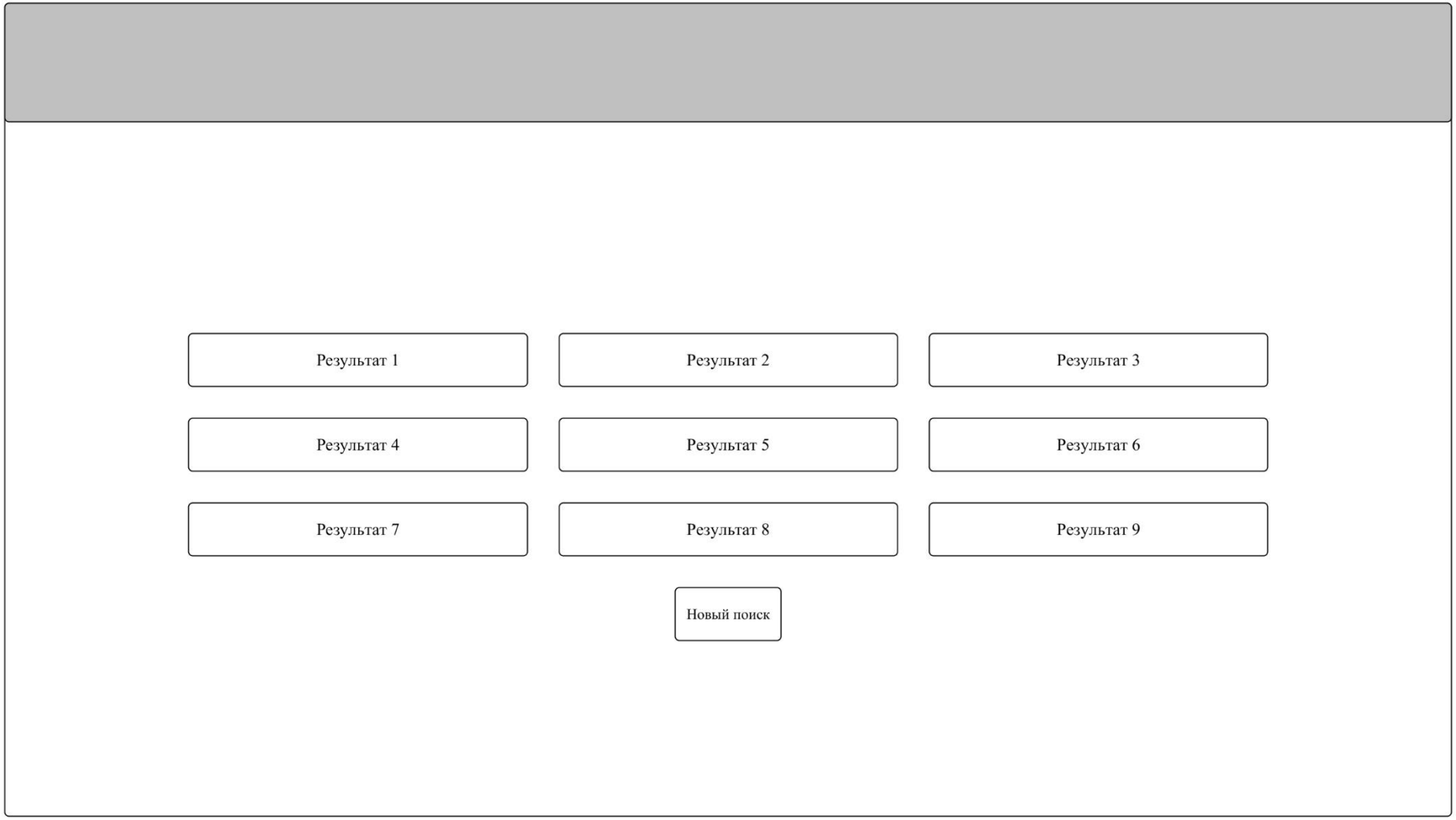


Рис. 2.23. Интерфейс отображения результатов

Выводы по Главе 2

Таким образом, в Главе 2 был показан процесс разработки информационно-поисковой системы локального пространства.

Были спроектированы модули каталогизации, индексации и трансляции, а также разработаны алгоритмы работы модулей системы.

При разработке модулей системы особое внимание было уделено использованию контрольной суммы страниц и различным видам каталога.

Была разработана база данных, служащая для индексации поискового пространства ресурса. Это позволило в дальнейшем реализовать статистический и когнитивный подходы к поиску информации.

В рамках модуля трансляции были разработаны алгоритмы обработки однословного и многословного запросов пользователей.

ЗАКЛЮЧЕНИЕ

В процессе выполнения данной работы были исследованы поисковые процессы в информационном пространстве и рассмотрена история развития информационно-поисковых систем, проанализирована структура словосочетаний в современном английском языке.

Была разработана информационно-поисковая система локального пространства на основании анализа английских словосочетаний.

Были проанализированы существующие на сегодняшний день наработки в сфере поиска информации.

Были разработаны модули каталогизации, индексации и трансляции. А в частности была спроектирована их архитектура и разработаны соответствующие алгоритмы.

Результатом данной работы стало нахождение способа решения проблемы поиска данных в локальных информационных пространствах, сочетающее в себе статистический и когнитивный подходы.

СПИСОК ИСПОЛЬЗОВАННОЙ ЛИТЕРАТУРЫ

1. Азимов, Э. Г. Новый словарь методических терминов и понятий (теория и практика обучения языкам) / Э. Г. Азимов, А. Н. Щукин. - М.: Издательство ИКАР, 2009. - 448 с.

2. Аракин, В. Д. Сравнительная типология английского и русского языков / В. Д. Аракин. - Л.: Просвещение, 1979. – 324 с.

3. Ахманова, О. С. Словарь лингвистических терминов / О. С. Ахманова. - 6-е изд. - М.: URSS, 2012. - 576 с.

4. Бархударов, Л. С. Структура простого предложения современного английского языка / Л. С. Бархударов. - М.: Просвещение, 1966. - 256 с.

5. Бегг, К. Базы данных. Проектирование, реализация и сопровождение. Теория и практика / К. Бегг, Т. Коннолли. - М.: Вильямс, 2003. - 1436 с.

6. Белоусов, К. И. Синергетика текста: от структуры к форме : монография / К. И. Белоусов. - М.: Эдито-риал УРСС, 2008. - 248 с.

7. Бенвенист, Э. Общая лингвистика / Э. Бенвенист. - М.: Едиториал УРСС, 2002. - 448 с.

8. Борисова, Е. Г. Слово в тексте. Словарь коллокаций (устойчивых сочетаний) русского языка с англо-русским словарём ключевых слов / Е. Г. Борисова. - М.; ФАРГУС, 1995. - 148 с.

9. Большой академический словарь русского языка: словарь / Гл. ред. А. С. Герд. - М; СПб: Наука, 2013. - 744 с.

10. Большой толковый словарь русского языка: словарь / Гл. ред. С. А. Кузнецов. - СПб.: Норинт, 2007. - 960 с.

11. Гуревич, В. В. Теоретическая грамматика английского языка / В. В. Гуревич. - М.: ЭКСМО, 2003. - 218 с.

12. Иванова, И. П., Теоретическая грамматика современного английского языка / И. П. Иванова, Г. Г. Почепцов. - М.: АГРАФ, 1981. - 325 с.

13. Иофик, Л. Л. Хрестоматия по теоретической грамматике английского языка / Л. Л. Иофик, Л. П. Чахоян. - Л.: Просвещение, 1972. - 276 с.

14. Ичкинеева, Д. А. Дистантные и контактные связи как способ реализации категорий дискретности и континуальности структуры текста / Д. А. Ичкинеева // Вестник Челябинского государственного университета. - 2009 - №39. - С. 53-57.

15. Лядова, Ю.В. Проблема словосочетания в современной лингвистике / Ю. В. Лядова // Филологические науки в МГИМО: Сб. научных трудов / Ю. В. Лядова. - М.: МГИМО, 2004. - С.120- 124.

16. Некрасова, О. А. Лингвистические коллокации английского языка и факторы, влияющие на процесс их образования / О. А. Некрасова // Вестник Московского государственного областного университета. Серия: Лингвистика. - 2017. - №195. - С. 120-123.

17. Оголева, Л. Н. Реинжиниринг производства: учебное пособие / Л. Н. Оголева, Е. В. Чернецова, В. М. Радиковский. - М.: КНОРУС. - 2005. - 304 с.

18. Ойхман, Е. Г. Реинжиниринг бизнеса: реинжиниринг организаций и информационные технологии / Е. Г. Ойхман, Э. В. Попов. - М.: Финансы и статистика, 1997. - 336 с.

19. Плоткин, В. Л. Строй английского языка / В. Л. Плоткин. - М.: ЭКСМО, 1989. - 345 с.

20. Смирницкий, А. И. Синтаксис английского языка / А. И. Смирницкий. - М.: ЭКСМО, 2007. - 421 с.

21. Стройков, С. А. Стилистика английского языка / С. А. Стройков. - С.: Поволжская государственная социально - гуманитарная академия, 2009. - 85 с.

22. Толковый словарь русского языка: словарь / Под ред. Д.Н. Ушакова. - М.: Гос. ин-т "Сов. энцикл."; ОГИЗ; Гос. изд-во иностр. и нац. слов., 1935-1940. (4 т.)

23. Толковый словарь русского языка: словарь / Совместно с Н. Ю. Шведовой. - М.: Азъ, 1992. - 944 с.

24. Хаммер, М. Реинжиниринг корпорации / М. Хаммер, Д. Чашпи. - Манн: Иванов и Фербер, 2011. – 288 с.

25. Conrad J., Viescas J. Microsoft Office Access 2007 Inside Out. - Pearson Education, 2007.

26. Ellis M. A., Stroustrup B. The annotated C++ reference manual. - Addison-Wesley Longman Publishing Co., Inc., 1990.

27. Kruglinski D. J., Wingo S., Sheperd G. W. Programming Microsoft Visual C++. - Microsoft press, 1998.

28. Swart B. Borland C++ Builder 6 Developer's Guide. - Sams Publishing, 2003.

29. The BBC [Электронный ресурс]. Режим доступа: http://www.bbc.com (дата обращения: 21.03.2018).

30. The Daily Mail [Электронный ресурс]. Режим доступа: http://www.dailymail.co.uk. (дата обращения: 21.03.2018).

31. The Forbes [Электронный ресурс]. Режим доступа: https://www.forbes.com (дата обращения: 27.03.2018).

32. The Guardian [Электронный ресурс]. Режим доступа: https://www.theguardian.com (дата обращения: 16.04.2018).

33. The Independent [Электронный ресурс]. Режим доступа: https://www.independent.co.uk (дата обращения: 12.04.2018).

34. The New York Times [Электронный ресурс]. Режим доступа: https://www.nytimes.com (дата обращения: 19.04.2018).

35. The Reuters [Электронный ресурс]. Режим доступа: https://www.reuters.com (дата обращения: 15.05.2018).

36. The Washington Post [Электронный ресурс]. Режим доступа: https://www.washingtonpost.com (дата обращения: 21.05.2018).

ПРИЛОЖЕНИЯ

Приложение 1

Рис. 2.1. Процесс работы ИПС

Wild cats
Wild cats online
Wild cats Scotland
Wildcats
Wildcats perth

Рис. 2.2. Пример поисковых запросов пользователей

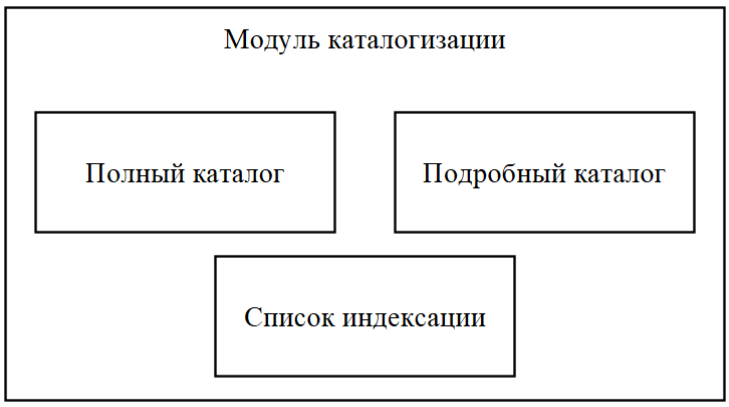


Рис. 2.3. Модуль каталогизации

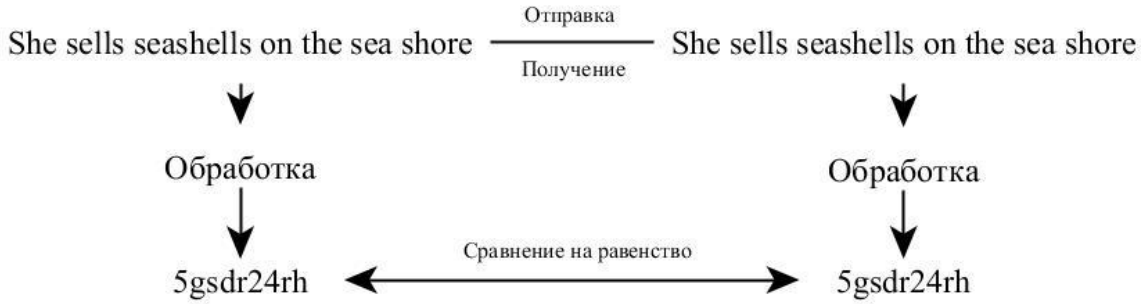


Рис. 2.4. Сравнение контрольных сумм строки до и после её передачи

Приложение 2

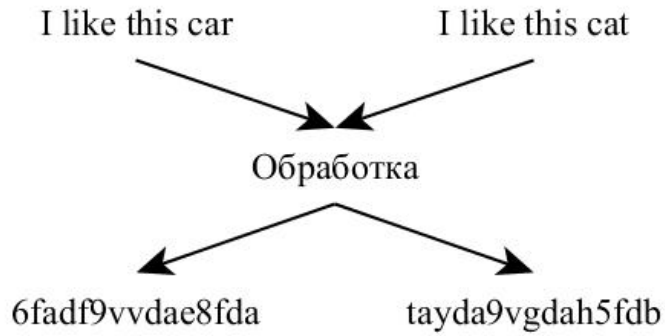


Рис. 2.5. Пример использования контрольной суммы для проверки состояния изменения текста

Таблица 2.1. Первый уровень каталога

Первый уровень каталога		
Страница:	Контрольная сумма:	Ссылки:
Главная страница	zddv	А
		Б
		В
		Г
		Д
		…
		Я

Приложение 3

Таблица 2.2. Первый и второй уровни подробного каталога

Первый уровень каталога			Второй уровень каталога		
Страница:	Контрольная сумма:	Ссылки:	Сайт:	Ссылки:	Контрольная сумма:
Главная страница	zddv	А	А	АА	ascvva
		Б		…	…
		В		АЯ	dvzd
		Г	Б	БА	zxfvf
		Д		…	…
		…		БЯ	s
		Я	В	ВА	zsc
				…	…
				ВЯ	szcv
			…	……	…
				…	…
				……	…
			Я	ЯА	dvzd
				…	…
				ЯЯ	xfv

Таблица 2.3. Трёхуровневый каталог ресурса

Первый уровень каталога			Второй уровень каталога			Третий уровень каталога		
Страница:	Контрольная сумма:	Ссылки:	Сайт:	Ссылки:	Контрольная сумма:	Сайт:	Ссылки:	Контрольная сумма:
Root	4d67hfxlun9ddv	A	A	AA	ascvva	AA	AAA	zddv
		B		…	…		…	fbbs
		C		AZ	dvzd		AAZ	adzxc
		D	B	BA	zxfvf	…	…	4sdfd
		E		…	…		…	sdfs
		…		BZ	sdf		…	cxva
		Z	C	CA	zsc	AZ	AZA	zddv
				…	…		…	fbbs
				CZ	szcv		AZZ	adzxc
			…	……	…	BA	BAA	4sdfd
				…	…		…	sdfs
				……	…		BAZ	cxva
			Z	ZA	dvzd	…	…	zddv
				…	…		…	fbbs
				ZZ	xfv		…	adzxc
						BZ	BZA	4sdfd
							…	sdfs
							BZZ	cxva
						…	…	zddv
						ZZ	ZZA	4sdfd
							…	sdfs
							ZZZ	cxva

Приложение 5

Таблица 2.4. Полный каталог

ИД:	Адрес:	Контрольная сумма
0	Главная страница	zddv
1	АА	ascvva
2	АЯ	dvzd
3	БА	zxfvf
4	БЯ	sjhhg
5	ВА	zsc
6	ВЯ	szcv
7	ЯА	dvzd
8	ЯЯ	xfv
9	ААА	zddv
10	ААЯ	adzxc
11	АЯА	zddv
12	АЯЯ	adzxc
13	ВАА	4sdfd
14	ВАЯ	cxva
15	ВЯА	4sdfd
16	ВЯЯ	cxva
17	ЯЯА	4sdfd
18	ЯЯЯ	cxva

Приложение 6

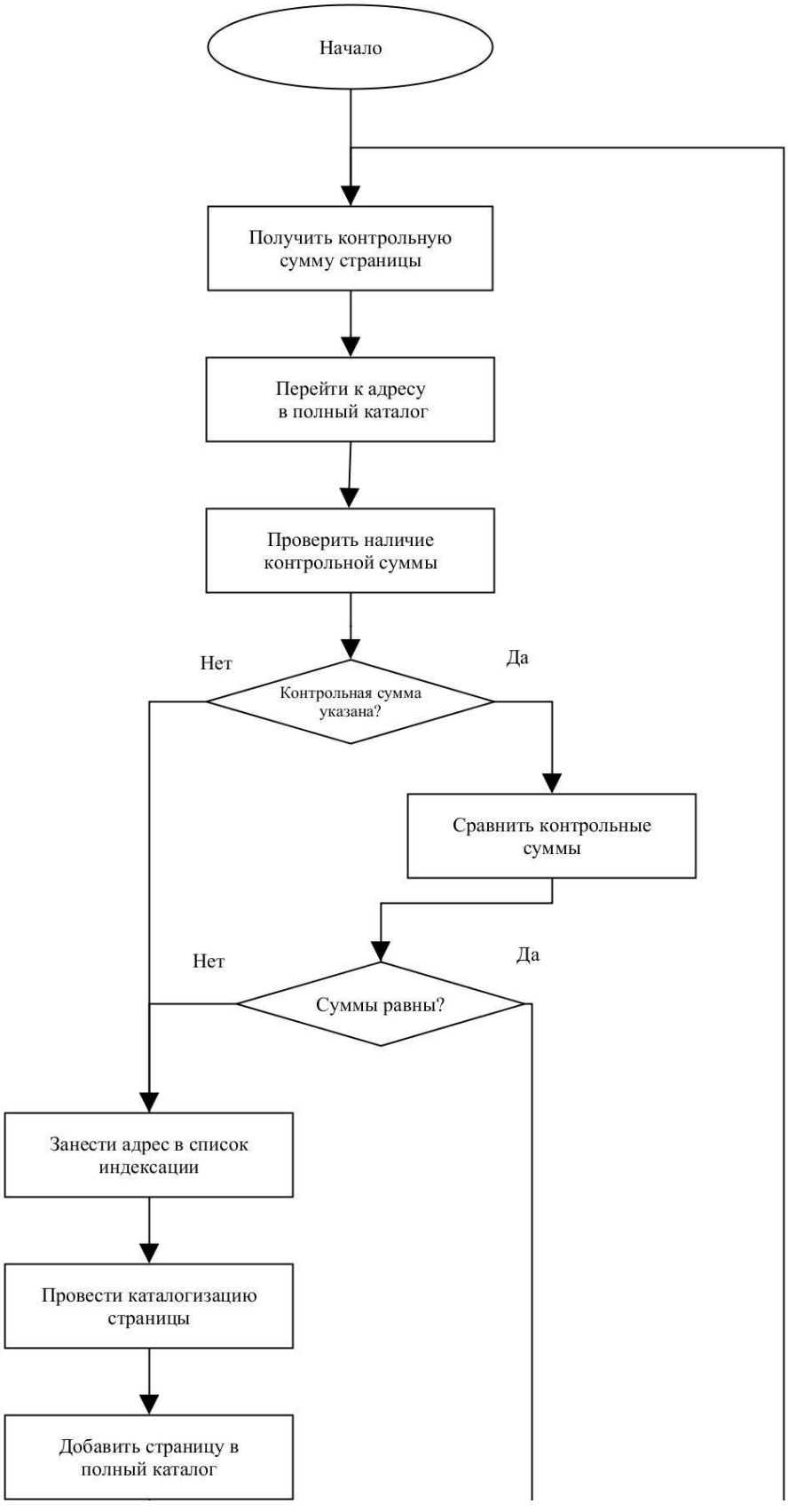


Рис. 2.6. Алгоритм каталогизации

Приложение 7

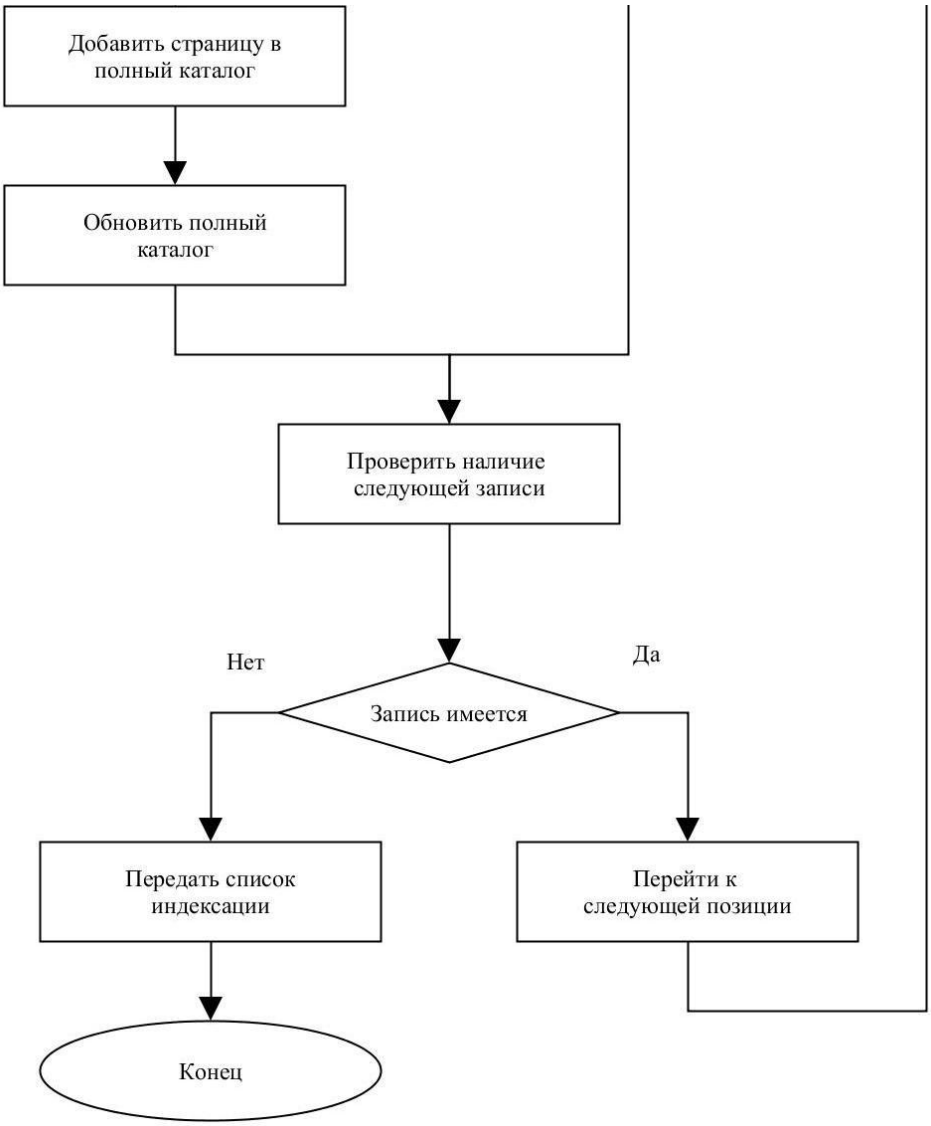


Рис. 2.6. Алгоритм каталогизации (продолжение)

Приложение 8

Таблица 2.5. Список индексации

ИД:	Адрес:
111	ААА
145	ААЯ
178	АЯ
179	АЯА
202	АЯЯ
874	Я
875	ЯА
876	ЯАА
901	ЯАЯ
976	ЯЯ
977	ЯЯА
999	ЯЯЯ

Рисунок 2.7. Модуль индексации

Приложение 9

```
<body>
<h1>Wild cats</h1>
<p>Tigers are wild cats</p>
<p>Really wild cats</p>
<!-- Add the third paragraph -->
</body>
```

↓

wild, cats, tigers, are, wild, cats,
really, wild, cats

Рис. 2.8. Обработка кода

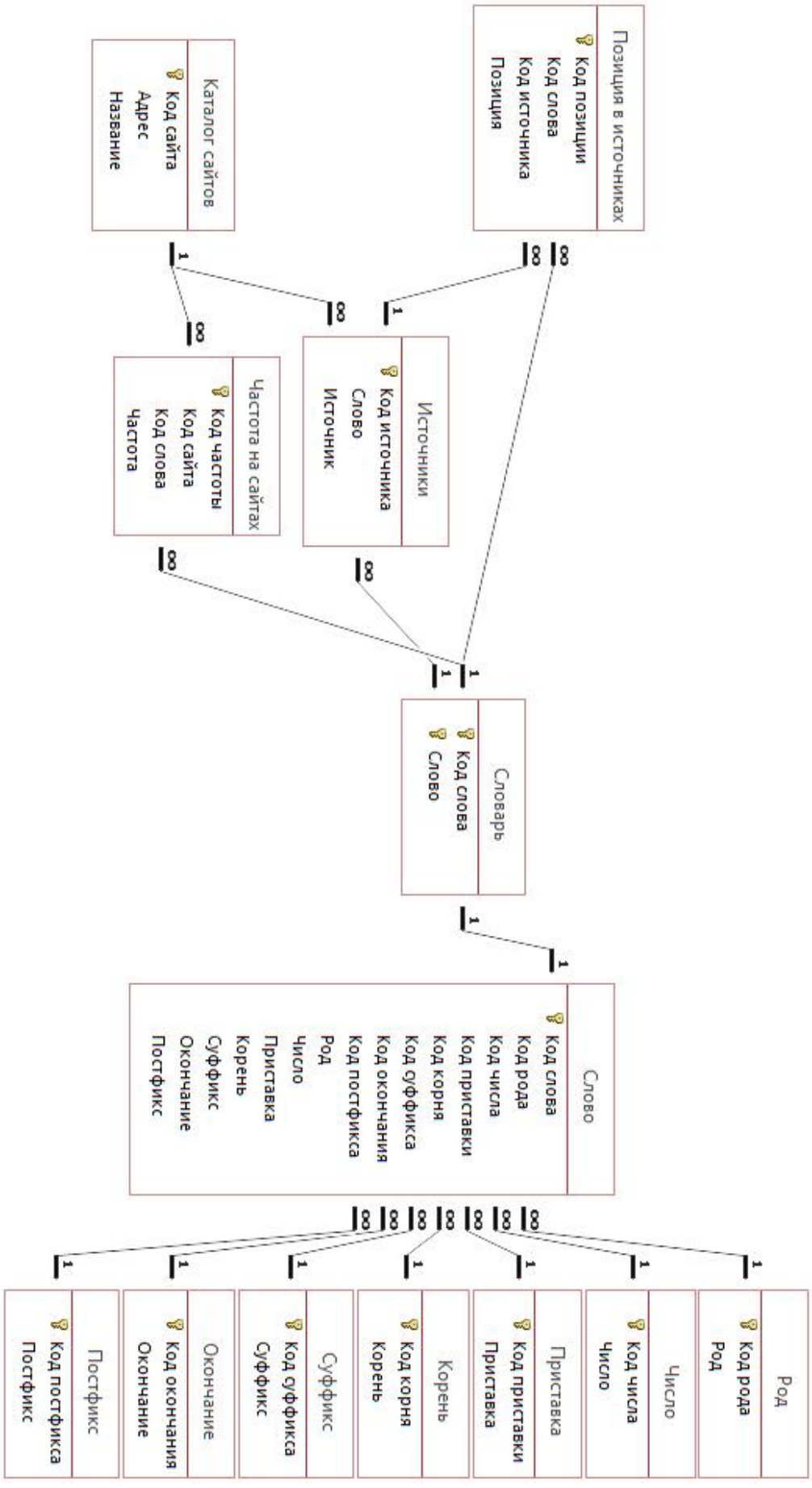


Рис. 2.9. База данных

Приложение 10

Приложение 11

Таблица 2.6. Сущность «словарь»

Код слова	Слово
1	group
2	schedule
3	enrollee
4	unpredictable
5	reaction
6	application
7	institute
…	…

Таблица 2.7. Сущность «слово»

Код слова	Коды	Префикс	Корень	Суффикс
1	…		group	
2	…		schedule	
3	…		enroll	ee
4	…	Un	predict	able
5	…		react	ion
…	…	…	…	…

Таблица 2.8. Сущность «каталог сайтов»

Код сайта	Адрес	Названия
1	bsu.edu.ru	Главная страница
2	bsu.edu.ru/schedule	Учебное расписание
3	bsu.edu.ru/documents	Документы
…	…	…

Приложение 12

Таблица 2.9. Сущность «источники»

Код источника	Слово	Источник
1	1 (group)	2 (bsu.edu.ru/schedule)
2	1 (group)	3 (bsu.edu.ru/documents)
3	2 (schedule)	2 (bsu.edu.ru/schedule)
4	4 (unpredictable)	1 (bsu.edu.ru)
5	4 (unpredictable)	3 (bsu.edu.ru/documents)
6	5 (reaction)	3 (bsu.edu.ru/documents)
…	…	…

Таблица 2.10. Сущность «позиция в источниках»

Код позиции	Код слова	Код источника	Позиция
1	4 (unpredictable)	3 (bsu.edu.ru/documents)	56
2	5 (reaction)	3 (bsu.edu.ru/documents)	57
3	1 (group)	2 (bsu.edu.ru/schedule)	23
…	…	…	…

Таблица 2.11. Сущность «частота на сайтах»

Код частоты	Код сайта	Код слова	Частота
1	3 (bsu.edu.ru/documents)	4 (unpredictable)	13
2	3 (bsu.edu.ru/documents)	5 (reaction)	5
3	2 (bsu.edu.ru/schedule)	1 (group)	29
…	…	…	…

Приложение 13

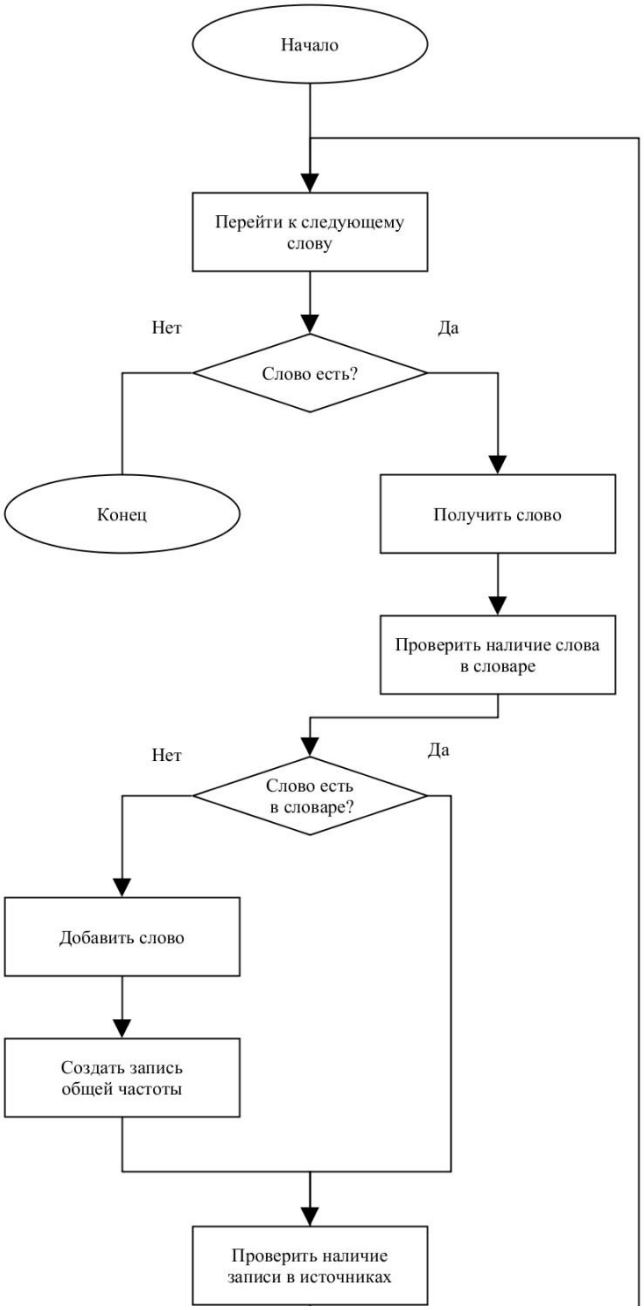
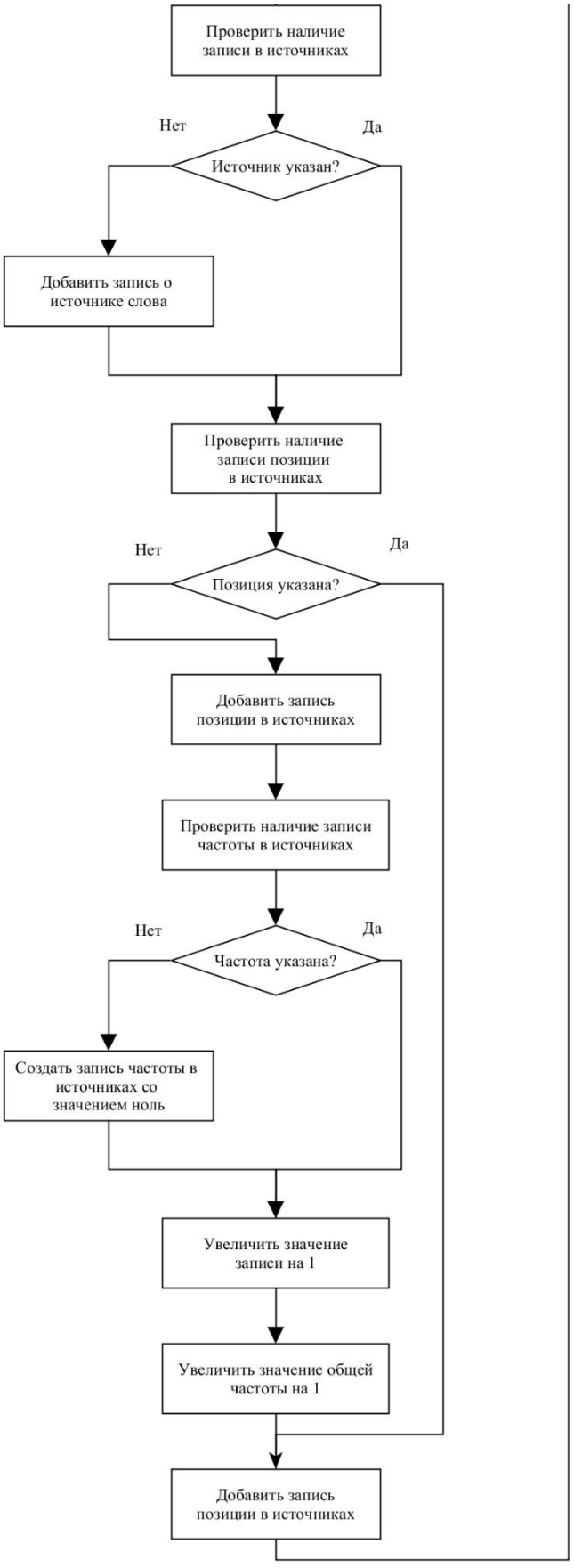


Рис. 2.10. Алгоритм индексации



Приложение 14

Рис. 2.10. Алгоритм индексации (продолжение)

Приложение 15

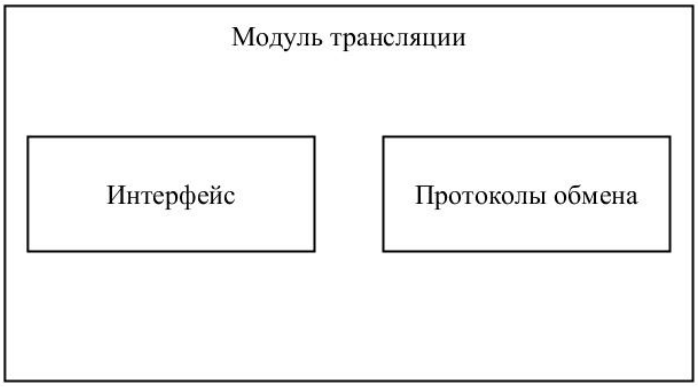


Рис. 2.11. Модуль трансляции

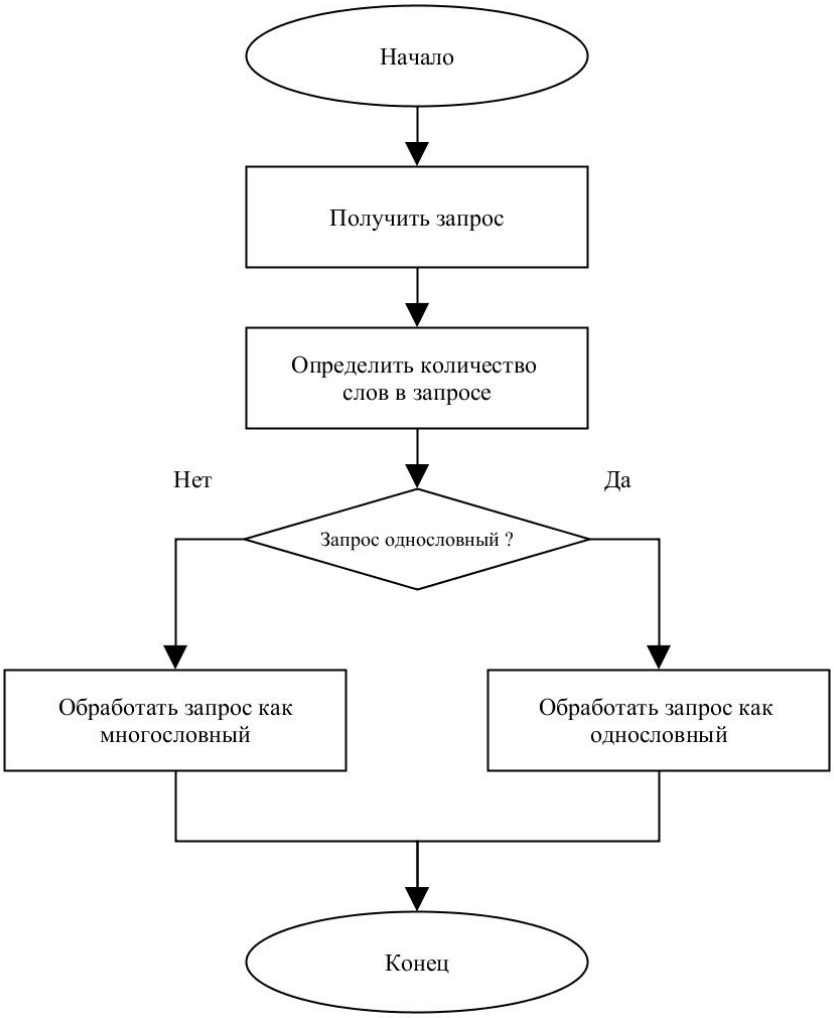
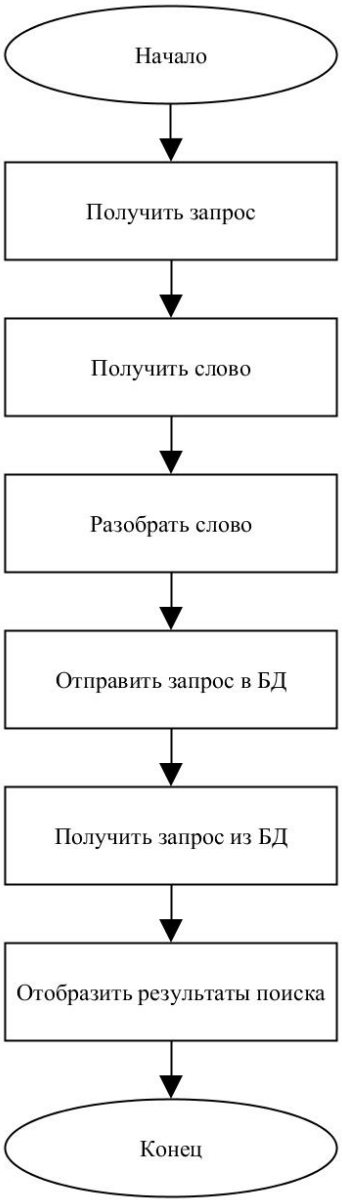


Рис. 2.12. Алгоритм работы модуля трансляции

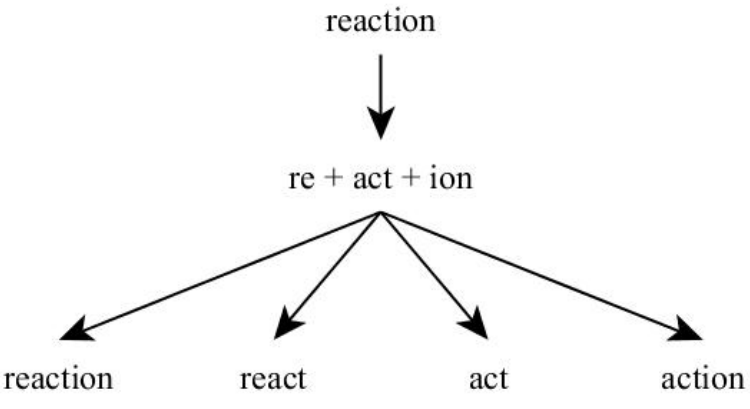


Приложение 16

Рис. 2.13. Алгоритм обработки однословного запроса

Рис. 2.14. Разбор однословного запроса

Приложение 17



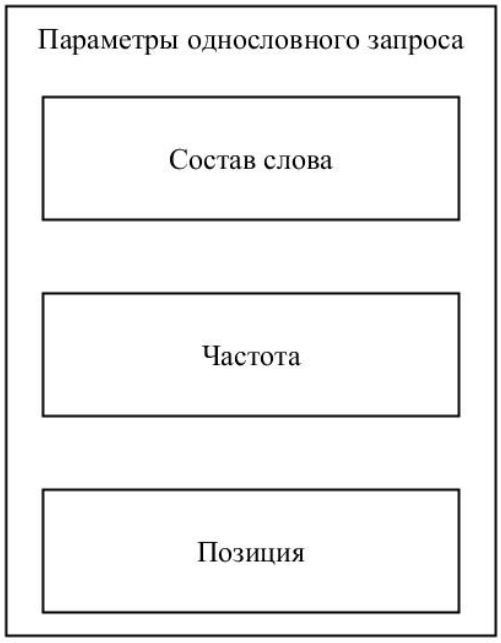


Рис. 2.15. Параметры однословного запроса

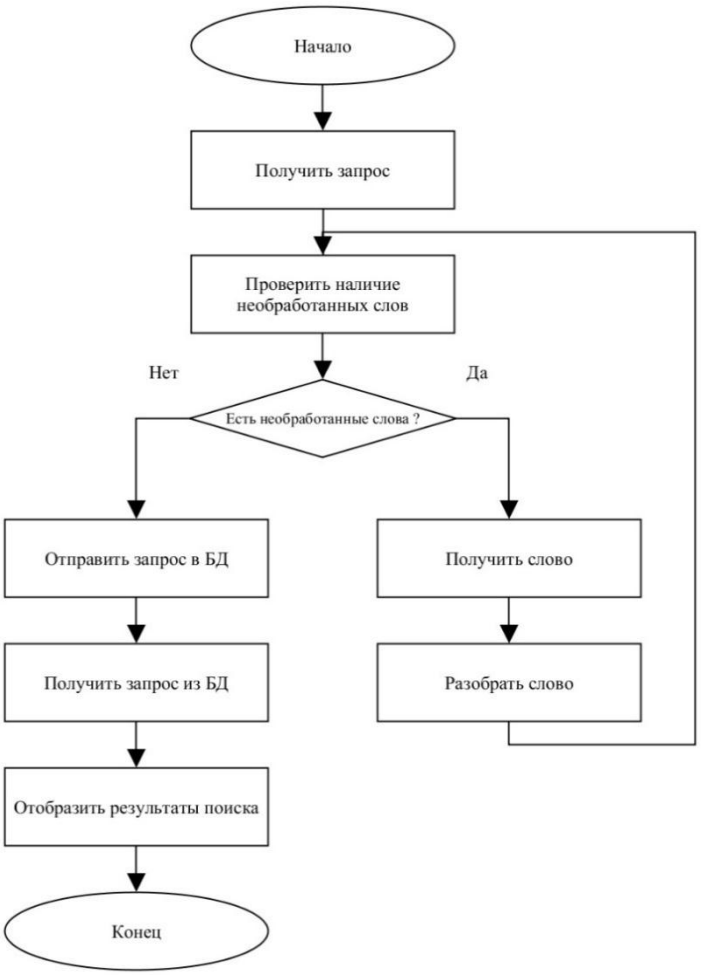


Рис. 2.16. Алгоритм обработки многословного запроса

Приложение 18

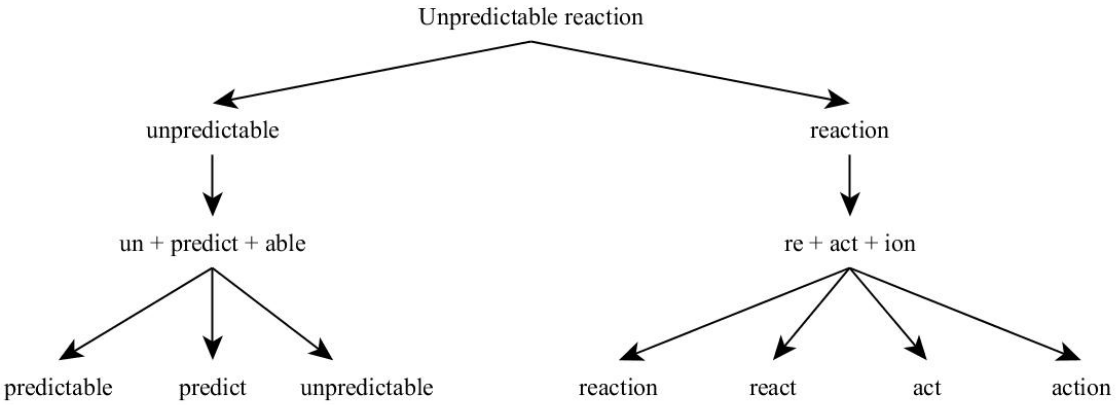


Рис. 2.17. Разбор многословного запроса

Приложение 19

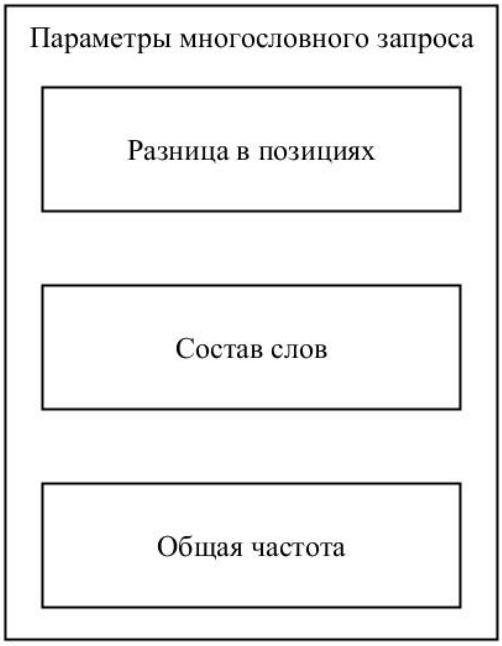


Рис. 2.22. Параметры многословного запроса

Приложение 20

Рис. 2.18. Слова оригинального запроса

Приложение 21

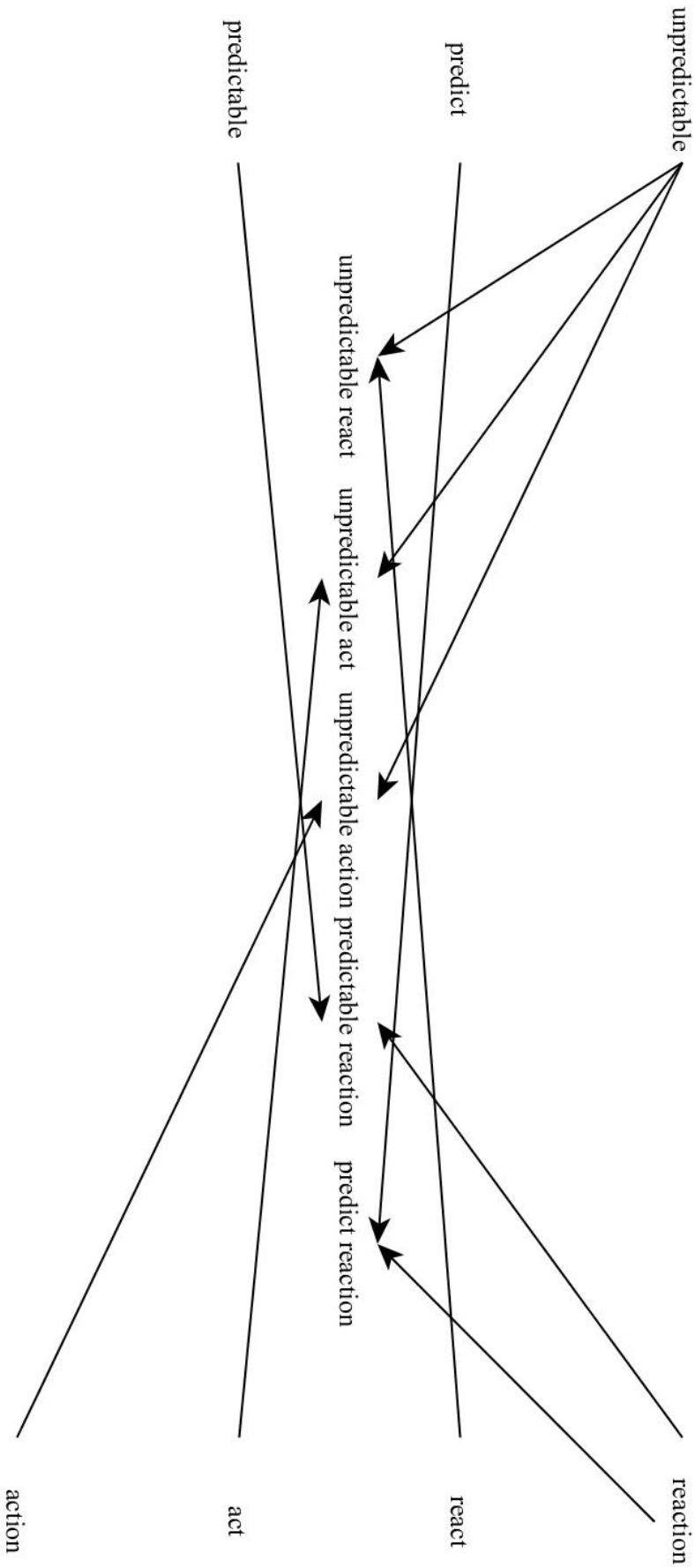


Рис. 2.19. Слова оригинального запроса и производные слова

Приложение 22

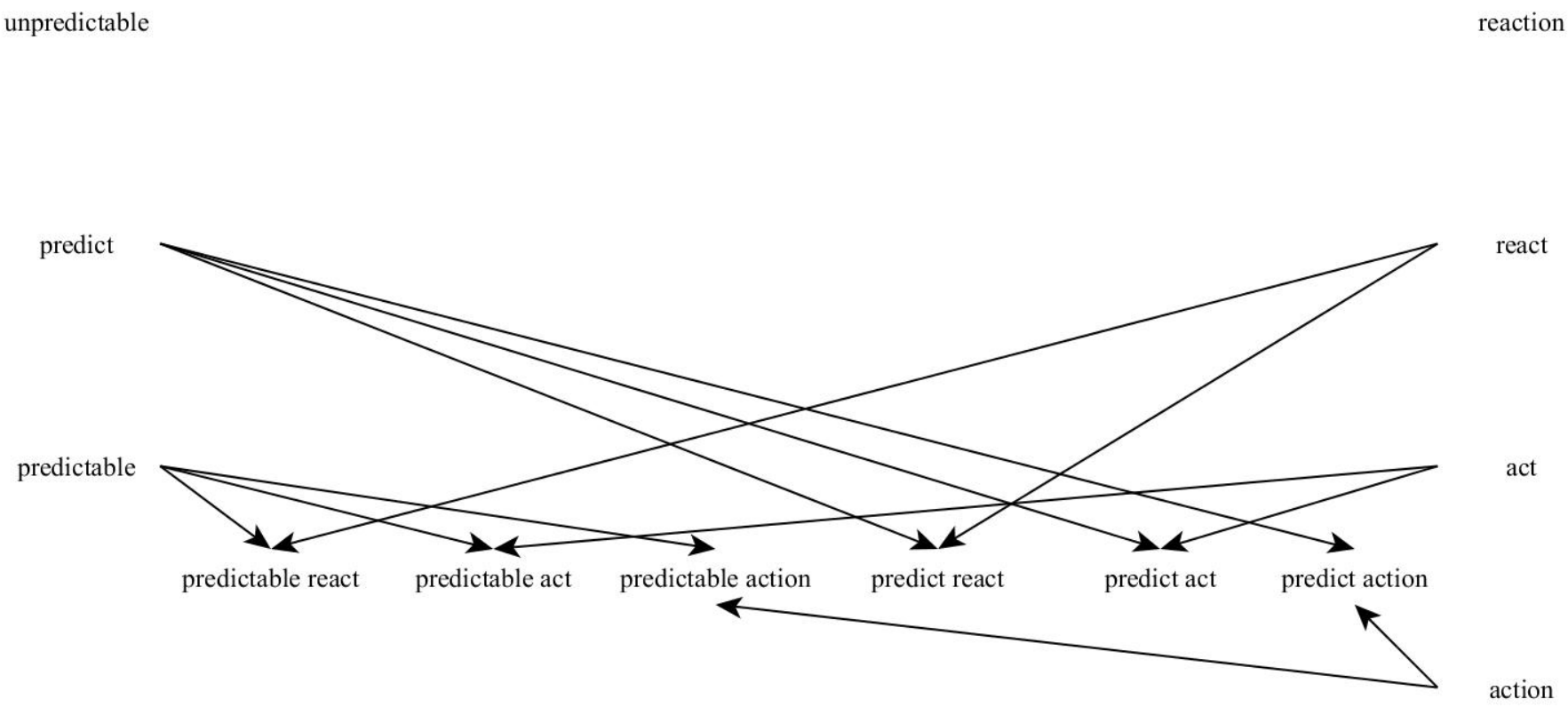


Рис. 2.20. Производные слова

Приложение 23

unpredictable reaction

predict react

unpredictable react unpredictable act unpredictable action predictable reaction predict reaction

predictable act

predictable react predictable act predictable action predict react predict act predict action

action

Рис.2.21. Результат разбора многословного запроса

Приложение 24

Рис. 2.23. Интерфейс ввода запроса

Приложение 25

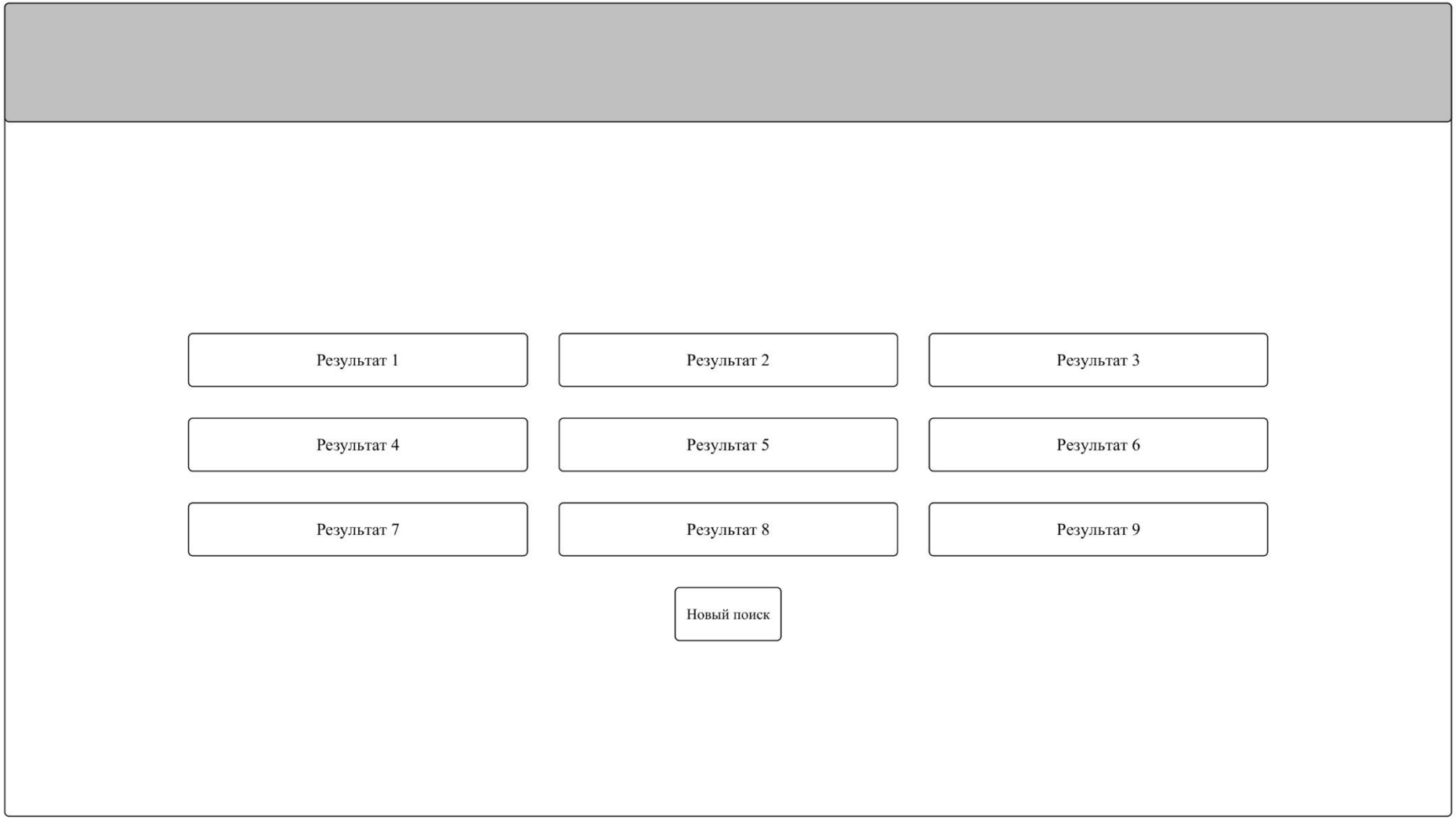


Рис. 2.23. Интерфейс отображения результатов