**Введение**

Современный этап развития цивилизации характеризуется переходом наиболее развитой части человечества от индустриального общества к информационному. Одним из наиболее ярких явлений этого процесса является возникновение и развития глобальной информационной компьютерной сети.

В 90-е гг. ХХ в. человечество получило доступ к огромному массиву информации в самых разных предметных областях – миллионам связанных документов, которые расположены на компьютерах по всему миру. Интернет сегодня – это неструктурированное распределенное хранилище динамичной информации громадного объёма, в котором постоянно добавляются новые документы, а устаревшие редактируются или удаляются. Колоссальное и непрерывно растущее количество информации приводит к тому, что обеспечение поиска в Интернете становится критически важной задачей. Однако без соответствующих поисковых средств проблема поиска в сети необходимой информации становится в настоящее время практически неразрешимой. Появление поисковых систем сети Интернет стало новым этапом в развитии информационного поиска. Сегодня информационно-поисковые системы(ИПС) сети Интернет эффективно применяют разработанные за много лет методы поиска информации. И важнейшей структурной составляющей ИПС является информационно поисковый язык.[5]

Быстрый рост и динамическое развитие информационных ресурсов обязывает каждого иметь разнообразные навыки сложного поиска. Уметь быстро и правильно искать в Интернете значит экономить время, владеть достоверной и актуальной информацией, а значит делать верные выводы и принимать правильные решения.

Все это свидетельствует об ***актуальности*** более глубокогоизучения информационно-поисковых языков в сети Интернет, что и определило выбор темы нашего исследования.

В этой связи ***целью*** исследования является рассмотрение в теории и практике возможности использования информационно-поисковых языков в различных поисковых системах сети

***Объектом*** исследования является интернет как единая информационная среда.

***Предметом*** являются информационно-поисковые языки сети Интернет.

Для достижения поставленной цели необходимо решить следующие ***задачи***:

* провести анализ теоретической и научно-методической литературы по данной теме;
* описать основные типы информационно-поисковых языков;
* рассмотреть различные ИПС и провести сравнительный анализ применения в них информационно-поисковых языков.

*Глава I. Обзор информационно-поисковых языков сети Интернет*

**§1.Информационно-поисковые языки: классификация**

Главная задача информационно-поисковой системы - это поиск информации релевантной информационным потребностям пользователя. Слово релевантность означает соответствие между желаемой и действительно получаемой информацией. Релевантность можно еще представить как меру близости между реально полученными документами и тем, что следовало бы получить из системы. Естественно, что здесь возникает две задачи, которые следует решить: представление информации в системе и формулирование информационных потребностей пользователя. Эти две проблемы тесно связаны друг с другом.

Наиболее распространенными моделями представления документов в информационно-поисковой системе являются различные вариации на тему векторной модели, когда документ представляется как набор терминов. Но это не весь текст документа, а только небольшой набор терминов, который отражает его содержание. Базируясь на таком представлении о документе и рассмотрим различные информационно-поисковые языки (ИПЯ). [4]

Информационно поисковый язык — искусственный язык, предназначенный для формализованного описания смыслового содержания документов, данных, отдельных понятий или терминов и обеспечения последующего их поиска в информационно-поисковых массивах.

Формализация лексики и создание различных ИПЯ вызвано необходимостью устранения "избыточности" и "недостаточности" естественного языка для целей информационного поиска.

Как и в естественных языках, в информационно-поисковых языках используется некая лексика, то есть свой словарный состав. Лексические единицы (слова) ИПЯ могут быть простыми (элементарными) или сложными, то есть ИПЯ может иметь или не иметь свои средства «словообразования» – формирования более сложных лексических единиц из простых.

Основными элементами ИПЯ являются алфавит, лексика и грамматика. Алфавит ИПЯ система знаков, используемых для записи слов и выражений ИПЯ. Лексика, или словарный состав, ИПЯ совокупность слов, словосочетаний и выражений, используемых для построения текстов ИПЯ.[1]

По области или по сфере применения информационно-поисковых языков можно выделить:

1. Коммуникативные (общесистемные) ИПЯ - предназначенные для обеспечения взаимодействия между различными (информационными, библиотечными и др.) системами (в том числе распределенными по государственной, ведомственной или территориальной принадлежности);
2. Локальные (внутренние) ИПЯ - предназначенные для использования в рамках отдельной системы;
3. Внешние ИПЯ - используемые в других системах и предназначенные для взаимодействия только с ними.

Различают языки описания (декларативные языки), которые в свою очередь подразделяются на языки предкоординатные (классификационные) и посткоординатные (дескрипторные), а также процедурные языки - языки запросов и манипулирования данными .

Классификационные информационно-поисковые языки

Информационно-поисковые каталоги, основанные на классификации сведений по определенной предметной области, были первыми системами информационного поиска документов.

Первоначальные подходы к классификации тематики документов основывались на формировании списка предметных аналогов, располагаемых в алфавитном порядке. Каждая предметная рубрика получала определенный цифровой или буквенно-цифровой код. Содержание документа индексировалось перечислением кодов тех рубрик, которые отражали темы документа. Это перечислительная классификация.

Особенностью систем перечислительной классификации является возможность индексирования документов любым количеством рубрик, отражающих содержание документ. Для осуществления поиска необходимых документов по классификатору определяются коды интересующих пользователя рубрик и далее отбираются из хранилища те документы, которые проиндексированы соответствующими кодами. Отсутствие систематизированных связей и отношений между предметными рубриками является основным недостатком перечислительной классификации.

При систематизированной классификации список предметных рубрик строится как иерархическая структура, в виде перевернутого дерева. Вся предметная область разбивается на ряд взаимоисключающих (не пересекающихся) рубрик. Каждая рубрика, в свою очередь, может включать несколько подрубрик. Таким образом, при систематизированной классификации учитываются уже некоторые семантические основы предметной области, выражаемые в родовидовых отношениях основных категорий, понятий и классов.

Содержание документа индексируется кодами соответствующих рубрик, однако при этом отпадает необходимость в явном указании более общих рубрик, к которым относятся выделенные полрубрики. В результате индексирование и поиск документов на основе иерархической классификации позволяют более адекватно отражать содержание документов и обеспечивают большую точность поиска.

Перечислительный и иерархический подходы к классификации используются в алфавитно-предметных каталогах библиотек. Недостатком как перечислительной, так и иерархической классификации является принципиальная невозможность заранее перечислить все возможные темы документов.

Дескрипторные информационно-поисковые языки(ДИПЯ)

В основе построения дескрипторных информационно-поисковых языков лежит принцип координатного индексирования, который предполагает, что основное смысловое содержание документа может быть выражено списком ключевых слов. К ключевым словам относятся так называемые полнозначные слова — существительные, прилагательные, глаголы, наречия, числительные, местоимения. Ключевыми словами не могут быть предлоги, союзы, связки, частицы.

Основными элементами ДИПЯ являются:

* словарь лексических единиц;
* правила применения ИПЯ (грамматика), определяющие процедуру перевода текстов документов и запросов с естественного языка на ИПЯ;
* правила построения ИПЯ.

Словари лексических единиц делятся на две группы:

* основные лексические словари, составляющие лексику ИПЯ;
* морфологические словари, обеспечивающие морфологический анализ и нормализацию слов.

В качестве лексических единиц основных словарей используются ключевые слова, словосочетания и дескрипторы.

Дескриптор — понятие, обозначающее группу эквивалентных или близких по смыслу ключевых слов. Дескриптор - это имя класса синонимов. В качестве дескрипторов могут быть использованы код, слово или словосочетание.

Разработка дескрипторного языка фактически сводится к разработке информационно-поискового тезауруса (ИПТ).

Тезаурус (от греч. «хранилище», «сокровищница») в узком смысле представляет собой специальный словарь-справочник, в котором перечислены ключевые слова — дескрипторы определенной предметной области, указаны их синонимы, установлены способы устранения синонимии, омонимии, полисемии, определены родовидовые и ассоциативные связи дескрипторов.

Наиболее важными парадигматическими отношениями ИПТ являются:

* соподчинение;
* род-вид;
* часть—целое;
* причина-следствие;
* функциональное сходство.

Обобщенная структура ИПТ включает как минимум три составляющих: словарную часть, семантическую карту, руководство по использованию. Словарная часть — алфавитный список дескрипторов с их словарными статьями. Семантическая карта — система тематических классов дескрипторов, представленная в виде графической схемы или таблицы. Руководство по использованию ИПТ содержит правила перевода ключевых слов и словосочетаний на ИПЯ, правила лексикографического контроля и редактирования поискового образа документа и поискового образа запроса, а также правила ведения ИПТ.

Отличием информационно-поисковых тезаурусов от информационно-поисковых каталогов на основе предметной иерархической рубрикации является то, что в тезаурусах, помимо классификационной схемы, присутствуют сами ключевые слова и дескрипторы, объединяемые под названием классов, рубрик и т. д. В каталогах же присутствуют только лишь обозначения (названия) классов.

Главная идея информационно-поисковых тезаурусов заключается в повышении эффективности индексирования документов в рамках дескриптивного подхода. Однако в процессе индексирования учитываются семантические отношения между дескрипторами, что, в конечном счете, обеспечивает более адекватный содержанию документа поисковый образ и повышает эффективность поиска документов.

В настоящее время происходит расширение сфер применения автоматических тезаурусов. При этом тезаурусы выступают составной частью современных систем подготовки текстов, осуществляя лингвистическую поддержку процесса подготовки и обработки текстов на естественном языке.

Среди наиболее перспективных направлений развития автоматических тезаурусов можно указать следующие:

1. Получение справки по используемому слову. Указав слово, в качестве ключа для запроса, пользователь в ответ получает соответствующий фрагмент словаря, содержащий лингвистическую информацию о данном слове. Например, автоматический тезаурус получает от пользователя некоторое существительное и в ответ выдает совокупность устойчиво сочетающихся с ним глаголов или все наиболее часто сопровождающие его определения. При этом автоматически выполняется процедура нормализации входного слова (т.е. приведение существительного к именительному падежу).
2. Контекстные замены по требованию пользователей. В данном случае тезаурус не только подбирает вместо одного словосочетания другое, которое пользователь счел более соответствующим контексту по смысловым или стилистическим соображениям, но и автоматически переоформляет параметры слов (например, род прилагательного) в соответствии с контекстом. Это означает, что синтаксические операции, производимые тезаурусом, существенно усложняются.
3. Автоматическая оценка стиля. Если слова и словосочетания в тезаурусе снабдить стилистическими пометками, то он может использоваться для стилистической оценки текста с выделением слов и словосочетаний, выпадающих, из общего стиля документа.[6,1]

**§2. Типы информационно-поисковых языков**

Булевый поиск

Наиболее распространенным ИПЯ является язык, позволяющий составить логические выражения из набора терминов. При этом используются булевые операторы AND, OR, NOT. Запрос при этом может выглядеть следующим образом: ((информационная and система ) or ИПС) not СУБД

В данном случае эта фраза означает: "Найди все документы, которые содержат одновременно слова "информационная" и "система", либо слово "ИПС", но не содержат слова "СУБД"".

Запрос можно рассматривать как и реальный документ из базы данных. В нашем случае, фактически, мы имеем дело с двумя запросами: информационная and система not СУБД и ИПС not СУБД каждый из которых подразумевает как бы два действия: сначала найти все документы, содержащие необходимые пользователю термины, а потом отсеять те, которые содержат термин "СУБД".

Такая схема достаточно проста, и поэтому наиболее широко применяется в современных информационно-поисковых системах. Но еще 20 лет тому назад были хорошо известны и ее недостатки.

Булевый поиск плохо масштабирует выдачу. Оператор AND может очень сильно сократить число документов, которые выдаются на запрос. При этом все будет очень сильно зависеть от того, насколько типичными для базы данных являются поисковые термины. Оператор OR напротив может привести к неоправданно широкому запросу, в котором полезная информация затеряется за информационным шумом. Для успешного применения этого ИПЯ следует хорошо знать лексику системы и ее тематическую направленность. Как правило, для системы с таким ИПЯ создаются специальные документально лексические базы данных со сложными словарями, которые называются тезаурусами и содержат информацию о связи терминов словаря друг с другом.

Модификацией булевого поиска является взвешенный булевый поиск. Идея такого поиска достаточно проста. Считается, что термин описывает содержание документа с какой-то точностью, и эту точность выражают в виде веса термина. При этом взвешивать можно как термины документа, так и термины запроса. Запрос может формулироваться на ИПЯ, описанном выше, но выдача документов при этом будет ранжироваться в зависимости от степени близости запроса и документа. При этом измерение близости строится таким образом, чтобы обычный булевый поиск был бы частным случаем взвешенного булевого поиска.

Языки типа "Like this"

При внимательном рассмотрении взвешенного поиска закрадывается естественное желание вообще обойтись без логических коннекторов и измерять близость документа и запроса какими-либо другими критериями. Наиболее простой моделью этого типа является линейная модель индексирования и поиска, когда близость документа и запроса рассматривается как угол между ними. В этом случае высчитывается sin угла, который получают как скалярное произведение двух векторов. В соответствии со значением меры близости происходит ранжирование документов при выдаче ссылок на них пользователю. Вообще говоря, скалярное произведение не очень хорошо подходит для информационно-поисковых систем Интернет, так как длина запроса обычно невелика. Это в традиционных системах существовали специальные службы, которые отлаживали длинные запросы, а в Интернет такие службы только нарождаются. Поэтому реально применяются другие меры близости, но принцип остается тот же: сначала вычисляется мера, а потом происходит ранжирование.

Рассмотренный подход дает возможность более мягкого расширения и уточнения запросов, но он также не гарантирует высоких показателей релевантности, в случае выбора неудачной лексики.

Поиск в нечетких множествах

При этом типе поиска весь массив документов описывается как набор нечетких множеств терминов. Каждый термин определяет некую монотонную функцию принадлежности документам документального массива. Когда запрашивается AND, то это интерпретируется как минимум из двух функций, соответствующих терминам запросов, OR - как максимум, NOT - как 1-<значение функции>. В соответствии с полученными значениями результат поиска также ранжируется, как и в случае с поиском по мерам близости.

Следует сразу сказать, что этот метод поиска используется только в исследовательских системах и распространен крайне ограничено.

Пороговые модели

Как было видно из предыдущего изложения, на конечном этапе поиска выборка найденных документов ранжируется. Но, совершенно очевидно, что меры близости или поиск в нечетких множествах приводит к ранжированию всего массива документов в базе данных. Современные информационно-поисковые системы Интернет имеют базы данных только индексов, занимающие терабайты. Ранжировать целиком такие массивы - это просто безумная затея. Поэтому применяются пороговые модели, которые задают пороговые значения для документов, выдаваемых пользователю.

Кластерная модель и вероятностная модель информационного поиска

В кластерной модели может использоваться два подхода. Первый заключается в том, что массив заранее разбивается на подмножества документов и при поиске высчитывается близость запроса некоторому подмножеству. В другом подходе кластер "накручивается" вокруг запроса и ближайших к нему терминов. Наиболее часто эта модель применяется в системах, уточняющих запрос по релевантности найденных документов.

При вероятностной модели вычисляется вероятность принадлежности документа классу релевантных запросу документов. При этом используется вероятность принадлежности терминов запроса каждому из документов базы данных.

Коррекция запроса по релевантности

Многие системы применяют механизм коррекции запроса по релевантности. Это означает, что процедура поиска носит интерактивный и итеративный характер. После проведения первичного поиска пользователь отмечает из всего списка найденных документов релевантные. На следующие итерации система расширяет, уточняет запрос пользователя терминами из этих документов и снова выполняет поиск. Так продолжается до тех пор пока пользователь не сочтет, что лучшего результата, чем он уже имеет добиться не удастся. Коррекция запроса по релевантности - это достаточно широко внедренный способ уточнения запросов. В некоторых системах пользователь может и не знать, о том, что эта процедура применяется.[8]

*Глава II. Применение ИПЯ в различных поисковых системах*

**§1. Анализ популярных поисковых систем сети Интернет**

При описании и классификации информационно-поисковых систем ставилась задача проанализировать наиболее популярные и наиболее типичные системы, которыми пользуются в Сети.

Lycos

Как и большинство систем, Lycos дает возможность использовать простой запрос и более изощренный метод поиска. В простом запросе в качестве поискового критерия вводится предложение на естественном языке. Lycos производит нормализацию запроса, удаляя из него так называемые stop-слова, и только после этого приступает к его выполнению. Почти сразу выдается информация о числе документов на каждое слово, а уже позже и список ссылок на формально релевантные документы. В списке напротив каждого документа указывается его мера близости запросу, число слов из запроса, которые попали в документ и оценочная мера близости, которая может быть больше или меньше формально вычисленной. На апрель 1996 года в Lycos не был реализован булевый поиск, такие планы были анонсированы. Последнее предложение подразумевает только то, что нельзя вводить эти операторы в строке вместе с терминами, но использовать логику через систему меню Lycos позволяет. Последнее относится к расширенной форме запроса, который предназначен для использования искушенными пользователями системы, которые уже научились пользоваться этим механизмом.

Таким образом мы видим, что Lycos относится к системе с языком запросов типа "Like this", но предполагается его расширения и на другие способы организации поисковых предписаний.

AltaVista

Наиболее интересным с точки зрения информационно-поискового языка в AltaVista является возможность расширенного поиска. Здесь стоит сразу выделить, что в отличии от многих систем AltaVista поддерживает одноместный оператор NOT. Кроме этого есть еще и оператор NEAR, который реализует возможность контекстного поиска, когда термины должны располагаться рядом в тексте документа. AltaVista разрешает поиск по ключевым фразам, при этом она имеет довольно большой словарь этих фраз. Кроме всего прочего, при поиске в АltaVista можно задать имя поля где должно встретиться слово. Это может быть гипертекстовая ссылка, applet, название образа, заголовок и ряд других полей. К сожалению, подробно процедура ранжирования в документации по системе не описана, но сказано, что ранжирование применяется как при простом поиске, так и при расширенном запросе. Реально эту систему можно отнести к системе с расширенным булевым поиском.

Yahoo

Данная система появилась в сети одной из первых, и поэтому говорить будем о сегодняшнем состоянии Yahoo, а не о состоянии годовой давности. В настоящее время Yahoo сотрудничает со многими производителями средств информационного поиска и на различных ее серверах используется различное программное обеспечение. На наш взгляд, это самая незатейливая информационная служба, которая сосредоточилась на информации о Web как таковой. ИПЯ Yahoo достаточно прост: все слова следует вводить через пробел и они соединяются либо AND, либо OR. При выдаче не выдается степени соответствия документа запросу, а только подчеркиваются слова из запроса, которые встретились в документе. При этом не производится нормализация лексики и не проводится анализ на "общие" слова. Хорошие результаты поиска получаются только тогда, когда пользователь знает, что информация в базе данных Yahoo точно есть. Ранжирование производится по числу терминов запроса в документе. Yahoo относится к классу простых традиционных систем с ограниченными возможностями поиска.

OpenText

Информационная система OpenText представляет из себя самый коммерциализированный информационный продукт в сети. Все описания больше напоминают рекламу, чем реальное руководство по работе. Система позволяет провести поиск с использованием логических коннекторов, размер запроса ограничен тремя терминами или фразами. В данном случае речь идет о расширенном поиске. При выдаче результатов поиска сообщается степень соответствия документа запросу и размер документа. Система позволяет также улучшить результаты поиска в стиле традиционного булевого поиска. OpenText можно было бы отнести без сомнения к разряду традиционных информационно-поисковых систем, если бы не механизм ранжирования.

InfoSeek

Система InfoSeek обладает довольно развитым информационно-поисковым языком, который позволяет не просто указывать какие термины должны встречаться в документах, но и своеобразно взвешивать их. Достигается это при помощи специальных знаков "+" - термин обязан быть в документе, "-" - термин обязан отсутствовать в документе. Кроме этого InfoSeek позволяет проводит то, что называется контекстным поиском. Это значит, что используя специальную форму запроса можно потребовать последовательной совместной встречаемости слов. Кроме этого можно указать, что некоторые слова должны совместно встречаться не только в одном документе, а даже в отдельном параграфе или заголовке. Есть возможность и указания ключевых фраз. Ключевая фраза от последовательной встречаемости отличается тем, что фраза всегда ищется как единое целое, а при последовательной встречаемости слова могут стоять рядом, но в произвольном порядке. Ранжирование при выдаче осуществляется по числу терминов запроса в документе, по числу фраз запроса в документе, за вычетом общих слов. Все эти факторы используются как вложенные процедуры. Подводя краткое резюме можно сказать, что InfoSeek относится к традиционным системам с элементом взвешивания терминов при поиске.

WAIS

WAIS является одной из наиболее изощренных поисковых систем Интернет. В отличии от многих поисковых машин, ИПЯ системы позволяет строить не только вложенные булевые запросы, считать формальную релевантность по различным мерам близости, взвешивать термины запроса и документа, но и осуществлять коррекцию запроса по релевантности. Система также позволяет использовать усечение терминов, разбиение документов на поля и ведение распределенных индексов. Не случайно именно эта система была выбрана в качестве основной поисковой машины для реализации энциклопедии "Британика" на Интернет. [3]

Применение языков на практике

Рассмотрим теперь небольшой сравнительный пример использования описанных выше поисковых машин. В качестве запроса использовалась фраза:

"Best on the Web"

Подразумевалось, что следует найти документ, связанный с конкурсами "Лучший на Сети". Эта фраза задавалась в качестве набора слов и при этом получались следующие результаты.

AltaVista - после нормализации лексики от запроса осталось только Best. Естественно, что при этом качество поиска было неудовлетворительным. Однако, использование поиска по фразе как по единому целому, поставило требуемый документ на первое место в списке найденных.

Lycos - здесь отсеялись "on the" и документ был указан только в конце списка. Поиск по фразе улучшения результатов не дал.

InfoSeek - при расширенном поиске нужный документ был найден третьим в списке из десяти документов. Уточнение поиска привело только к миграции документа вглубь списка.

OpenText - документ занимает пятую строчку в списке из десяти документов. Как и в случае с InfoSeek уточнение запроса результатов не дало.

Yahoo - документ попал в список найденных и занял третье место (ошибка в запросе: вместо "on the" следовало указывать "of the"). Но здесь следует заметить, что основное место хранения этого документа база данных Yahoo, т.е. запрос точно совпадает с тематикой базы данных. [2]

В завершении хотелось бы обратить внимание еще на один аспект выбора информационно-поисковой системы. Это профиль ее баз данных. Можно возразить, что все системы индексируют одно и тоже - массив документов Интернет. Однако делают они это по-разному. Очень важен профиль системы, который задается разбиением документов по темам и словарем индексирования, а также способом его поддержания. Определенным ориентиром здесь могут служить виртуальные библиотеки.

**Заключение**

Итоги теоретического исследования позволили нам выявить достоинства и недостатки различных ИПС, базирующихся на том или ином информационно-поисковым языке.

Так как каждая поисковая система предоставляет различные возможности поиска, из различных баз данных, поэтому информационный поиск на базе ИПС представляет собой достаточно сложный процесс познавательно-практической деятельности, требующий от поисковых субъектов априорной подготовки.

Анализ ИПЯ сети интернет поможет провести свой собственный выбор наиболее подходящего средства поиска, которое обеспечивало актуальность, быстроту и точность результатов.

**Список литературы**

1. Гост 7.25-2001 СИБИД. Информационно-поисковые языки. Термины и определения - М.: Изд-во стандартов,2001.-38 с.
2. Егоров А.Б. Поиск в Интернете – Санкт – Петербург.: НиТ, 2007г.
3. Захаров В.П. Информационно-поисковые системы: Учебно-методическое пособие.- Спб.,2005 г,48 с.
4. Кузьмин А.В. Золотарева Н.Н. Поиск в Интернете - СПб.: Изд-во НиТ,2006г.
5. Лозовский Л.Ш., Ратновский Л.А. Интернет – это интересно. – М.: ИНФРА-М, 2007. – 128 с.
6. Н.А. Гайдамакин .Автоматизированные информационные системы, базы и банки данных, М.: «Гелиос», 2006.
7. Симонович С.В., Мураховский В.И. Интернет у вас дома: Полное руководство для начинающего пользователя. – М.: АСТ-ПРЕСС, 2008. – 432 с.
8. Храмцов П.И. Информационно-поисковые языки, М.: «Гелиос», 2008