Содержание:

Введение

С каждым годом количество произведённой человечеством информации увеличивается. На 2018 г. объём данных составлял более 1,8 зеттабайт (1,8 трлн Гб). Как считают в международной исследовательской и консалтинговой компании International Data Corporation (IDC)[1], количество данных будет удваиваться каждые два года до 2020 г., а количество полезной информации будет составлять только 35% от общего объёма данных. Однако и это очень много.

Для облегчения поиска и ориентирования в таком объёме информации создаются различные поисковые средства. Большинство пользователей Интернет сообщества начинают свой рабочий день с поисковых систем, где пытаются найти столь необходимую им информацию и решить свои проблемы. К сожалению, поисковые системы часто не способны точно и справедливо интерпретировать ресурсы. Как результат, на первых позициях поиска зачастую оказываются сайты «далекие» от решаемого вопроса.

Причина такого положения проста и кроется в технологии получения и представления результатов поисковыми системами. При этом надо понимать, что главная проблема заключается в отсутствии четких правил, доступных и открытых для всех желающих. Чем больше неопределенности в алгоритмах формирования поисковых индексов (некий черный ящик), тем меньше поисковые системы отражают процесс формирования реальной информации. И соответственно, тем меньше будет уровень доверия к результатам поиска поисковых систем.

Но это вина не поисковых систем, поскольку они обязаны скрывать правила построения поисковых индексов. Это вина самой технологии при организации поиска. По своей сути технология поисковых систем направлена на пассивного пользователя. Необходимо зарегистрировать только сайт, дальше все сделает поисковый робот. Он просканирует ресурс страницу за страницей, пытаясь проанализировать содержание каждой из них. В такой схеме работы поисковым системам необходимо изменять алгоритмы и правила индексирования ресурсов и построения поискового индекса.

Конечно, большинство пользователей пользовались, пользуются, и будут пользоваться классическими поисковиками. Это просто, удобно и распространено. Это, как привычка, пользоваться поисковиками.

Поэтому выбранная тема курсовой работы является актуальной.

Объектом исследования в курсовой работе является глобальная сеть Интернет.

Предмет исследования – поисковые системы, используемые для нахождения необходимой информации в Интернете.

Целью исследования в курсовой работе является углубленное изучение существующих поисковых систем в сети Интернет.

Для достижения поставленной цели сформулированы следующие задачи:

1)изучить сущность поисковых систем в сети Интернет;

2)охарактеризовать процесс поиска информации в сети Интернет при использовании различных поисковых систем;

3)дать сравнительную характеристику поисковых систем.

Методической основой курсовой работы являются учебная и методическая литература, статьи в периодической печати и Интернет-ресурсы.

Глава 1. Теоретические основы работы поисковых систем

1.1 Понятие поисковых систем, их цели и задачи

Постоянное увеличение накапливаемой текстовой информации ведет к необходимости реализации эффективного поиска конкретной необходимой информации. Для поиска информации на естественном языке применяются методы полнотекстового поиска. Задача реализации полнотекстового поиска относится к классу слабо формализованных.

Полнотекстовый поиск – автоматизированный документальный поиск, при котором в качестве поискового образца документа используется его полный текст или

существенные части текста[2].

В некоторых информационно-поисковых системах полнотекстовый поиск рассматривается в качестве механизма управления данными большого объема на нижнем уровне их интеграции. При этом общий поисковой сервер, индексирующий информационные ресурсы, к которым имеется доступ, выступает в роли универсального интерфейса для работы с ними. Отмечается, что в некоторых случаях такого механизма управления данными вполне достаточно, а затраты на его внедрение являются минимальными. В случае если достаточный уровень управления данными нельзя обеспечить за счет полнотекстового поиска ввиду ограничений языка запросов на использование логической структуры документов, то переходят к другим подходам управления данными на основе графовых или древовидных моделей их представления, что зачастую связано со значительной модификацией данных и существенными затратами.

Толчком к использованию подхода к управлению данными на основе полнотекстового поиска послужили следующие объективные предпосылки[3]:

- простота и доступность html-стандарта представления данных и манипулирования ими с помощью веб-браузеров, которые в значительной степени основаны на инженерно-эвристических подходах;
- высокие темпы внедрения данного подхода во все сферы деятельности;
- отсутствие, по крайней мере, на первых этапах развития технологий полнотекстового поиска, необходимости применения строгих математических моделей, связанных с ними сложных алгоритмов и реализующего их программного обеспечения.

Несмотря на постоянное развитие технологии полнотекстового поиска всемирно известными IT-компаниями некоторые проблемы остаются неразрешенными:

- низкая релевантность поиска;
- относительно невысокие скорости поиска и анализа данных (особенно при решении сложных поисковых задач в плохо разработанных областях).

Причиной низкой релевантности может служить распределение в сетях различного масштаба разнородной тематической информации по многочисленным источникам, что приводит к необходимости выделения из нее данных близких по тематикес целью вторичной обработки и анализа. Разработки в этой области реализуют иностранные[4] и российские[5] компании. Еще одной причиной низкой релевантности является наличие в естественном языке: омонимов (разные по

значению, но одинаковые по звучанию и написанию слова); синонимов (слова, принадлежащие, как правило, к одной и той же части речи, различные по звучанию и написанию, но имеющие похожее смысловое значение) и т.д. В настоящее время проводятся исследования по поиску решения указанных проблем.

1.2 Критерии качества работы поисковых систем в сети Интернет

На сегодняшний день механизмы полнотекстового поиска реализованы в поисковых машинах Яндекс, Google,Yahoo, Rambleri многих других программных продуктах на основе программно-технических комплексов[6].

Качество выполненного поиска зависит от того, насколько найденный документ релевантен поисковому запросу пользователя. Такая оценка производится, в том числе, на основе методов ранжирования документов.

Выделяют такие внестраничные критерии релевантности документов, как, например:

1. ссылочное ранжирование: PageRank– это числовая величина, характеризующая «важность» веб-страницы. Чем больше ссылок на страницу, тем она «важнее». Кроме того, «вес» страницы A определяется весом ссылки, передаваемой страницей B. Таким образом, PageRank– это метод вычисления веса страницы путем подсчета важности ссылок на нее.
2. тип запроса:

- навигационный, информационный, общий, геозависимый и др.
- индекс цитирования;
- описание сайтов в каталогах;
- релевантность запросу сайта в целом; и т.д.

Все они имеют высокую значимость для релевантности по значительной доле запросов в поиске по Интернету. Релевантность текста страницы для таких запросов также имеет значение, однако при этом бывает достаточно ее грубой оценки, тонкие различия практически не влияют на релевантность результатов по подобным запросам. В то же время, не менее значительна и доля запросов, для которых внестраничная информация практически отсутствует и решающим оказывается страничное ранжирование. Таким образом, можно говорить, что хотя

релевантность результатов в поиске по Интернету определяется не только качеством алгоритмов страничного ранжирования, их влияние на качество поиска достаточно велико. Конечно, здесь надо иметь в виду, что релевантность результатов поиска в Интернете зависит не только от качества ранжирования, но и от других факторов. Объем и частота обновления базы, отслеживание нечетких дубликатов, фильтрация спама - все это также оказывает значительное влияние на качество поиска.

Выделяют следующие критерии выбора поискового механизма:

- скорость индексирования и переиндексации,
- поддерживаемые API(ApplicationProgrammingInterface, интерфейс программирования, интерфейс создания приложений),
- поддерживаемые протоколы,
- размер базы и скорость поиска,
- поддерживаемые типы документов,
- работа с разными языками и стемминг,
- поддержка дополнительных типов полей в документах,
- платформа и язык,
- возможность расширения встроенных механизмов ранжирования и сортировки.

Основные принципы определения релевантности:

1. Количество ключевых слов запроса в тексте документа.
2. Тэги, в которых эти слова располагаются.
3. Местоположение искомых слов в документе.
4. Удельный вес слов, относительно которых определяется релевантность, в общем количестве слов документа.
5. Время – как долго страница находится в базе поискового сервера.

6. Индекс цитируемости – как много ссылок на данную страницу ведет с других страниц, зарегистрированных в базе поисковика.

Критерием результата поиска является получение пользователем списка документов, одного документа или их частей, максимально удовлетворяющего его потребностям, сформулированным в поисковом запросе. Различают критерии смыслового и формального соответствия между поисковым предписанием и выдаваемым документом.

Полнота и точность поиска являются взаимосвязанными показателями. Увеличение одного из них ведет к снижению другого. Следует учитывать ситуацию, при которой список выданных поисковой системой ссылок содержит несколько, а порой и десятки разных адресов с одним и тем же текстом. Подобные ссылки характеризуются как дубликаты. Из них, при подсчете коэффициентов учитывается только один документ.

Значимой мерой релевантности в реальных поисковых системах является степень удовлетворенности пользователя полученными результатами. Естественно, этот критерий не поддается точному формальному определению, в отличие от критериев, используемых в экспериментах по информационному поиску. Вопрос о степени применимости традиционных формальных критериев к реальному поиску в Интернете остается малоисследованным. Например, такие значимые в экспе-риментальных исследованиях критерии как Precision, Recall, Average Precision ориентированы на ситуацию, когда пользователя интересуют все релевантные документы, и он просматривает всю поисковую выдачу. В реальном же поиске по Интернету подобная модель поведения пользователя является всего лишь одной из многих и встречается не столь уж часто. Возможно, в будущем будут разработаны системы оценки релевантности, учитывающие вероятную модель поведения пользователя для оцениваемого запроса и выбирающие адекватный критерий ранжирования, хотя и это будет лишь частичным решением проблемы.

С другой стороны, бесспорно наличие корреляции между формальными критериями и качеством поиска с точки зрения пользователя.

В настоящее время различают несколько общих моделей информационного поиска:

1. Булева модель, когда документы при поиске делятся на две группы – либо соответствующие, либо несоответствующие запросу, при этом никакие их оценки не вычисляются. В первоначальном варианте модели этого типа не поддерживали ранжирование документа (отсутствовал метод определения степени соответствия документа запросу – оценок релевантности документа запросу), выдавалось все множество документов, соответствующих запросу, без какого-либо ранжирования.
2. Модель векторного пространства – документ и запрос представляется в качестве вектора и ищется скалярное произведение векторов, которое позволяет оценить близость документа и термина.
3. Вероятностная модель, где вычисляется вероятность того, что документ релевантен, т.е. соответствует запросу с использованием полного

вероятностного подхода. Существует множество методов вычисление вероятности.
4. Модель обратной связи по релевантности и расширения запроса - позволяет при поиске учитывать ответы пользователя. Классический вариант подразумевает несколько итераций поиска, при каждом следующем шаге алгоритм улучшает результаты поиска.
5. Языковые модели информационного поиска - рассматривают задачу поиска со стороны документа. Если данный документ может породить запрос, то этот документ релевантен.

1.3 Особенности реализации поисковых технологий

Система StackSearch осуществляет поиск с учетом:

- морфологии нескольких естественных языков;
- атрибутовполнотекстовых документов (при необходимости с логическим объединением);
- эвристического алгоритма определения жизненного цикла документа, для мониторинга изменений в индексируемых документах с целью исключения при сборе информации документов, которые не были изменены;
- взаимодействия с другими поисковыми системами.

Stack Search состоит из различных модулей:

1. Краулер - модуль сбора документов для индексирования из различных источников;
2. Индексатор - модуль формирования поискового индекса по сформированной ранее коллекции документов;
3. Поисковой сервер - сложная программа (программный комплекс), осуществляющая реализацию поисковых запросов с применением поискового индекса;
4. Клиентские средства - программные библиотеки и утилиты, реализованные на различных языках программирования для взаимодействия с сервером поиска.

Поиск в Google. Google состоит из следующих модулей:

1. Модуль загрузки - обрабатывает URL-адреса из собственной базы данных URL, очищает соответствующий документ от нетекстовой информации и помещает его в базу данных html-документов;
2. Модуль обработки документа - обнаруживает в имеющихся html-документах ссылки и добавляет их в соответствующее хранилище, а составляющие документы слова помещает в хранилище слов, обработанные модулем документы далее размещаются в индексе;
3. Модуль обработки ссылок - при получении ссылки на не проиндексированный документ добавляет URL в соответствующее хранилище.
4. Модуль вычисления веса документа относительно запроса пользователя.

Таким образом, можно выделить следующие хранилища информации:

а) URL – содержит адреса страниц для индексирования;

б) HTML – хранит тексты документов, из которых удалены скрипты, картинки и пр.;

в) слов – хранит номера слов и сами слова для последующего обращения по номеру;

г) индексное – различные индексы, которые указывают, в каком документе находится данное слово, и наоборот, из каких слов состоит документ;

д) ссылок – хранит ссылки из обработанного документа;

е) ссылок на сайт – хранит данные о перекрестных ссылках с сайтов.

Поисковая система Яндекс. Реализует распределенную поисковую технологию, на всех уровнях поисковой системы производится распараллеливание нагрузки.

При обращении пользователя к системе его запрос перенаправляется на поисковой веб-сервер, который в настоящий момент менее загружен.

Далее производится обработка на уровне поисковой системы, на котором располагаются базы параллельного поиска (реализуется деление большой базы документов).

Современные реализации предполагают создание полнотекстового индекса, содержащего все слова с указанием мест их встречаемости. Таким образом, поиск заданных слова осуществляется в этом индексе, после чего доступен список документов, в которых он встречается. Кроме того документы индексируются

после исключения их дубликатов (либо по всем терминам, либо по основным, определенным некоторым специфическим для различных систем способом, ключевым словам). Большинство существующих программных реализаций информационно-поисковых систем позволяют ограничивать поиск по дате публикации, источнику информации, автору, учитывать морфологическую изменчивость ключевых слов и область поиска, если имеется возможность указать таковую. Область поиска также ограничивается посредством тематического рубрикатора. Для уточнения запросов в программах применяют словари синонимов, а также предлагаются слова, часто встречающиеся в сочетании с ключевыми словами.

В настоящее время большое распространение получили метапоисковые системы, которые в результате поиска выдают данные с десятка поисковых систем, при этом объем информации может быть весьма значительным. Чтобы пользователь не потерял в полученном массиве необходимую ему информацию, результирующие данные представляются в виде общего списка, где в первых элементах расположены данные, наиболее релевантные запросу. Альтернативным решением явились тематические поисковые системы на веб-сайтах – узконаправленные порталы. Кроме того некоторые системы позволяют экспортировать результирующий список для использования в других программных продуктах. Как правило, такой список содержит ссылки на документы, удовлетворяющие запросу, а также похожие документы. Сортировка в списке может осуществляться по релевантности, дате и т.п. При просмотре полнотекстового документа в нем осуществляется указание на найденные ключевые слова, например, путем подсветки. Существуют поисковые системы, в которых реализована и возможность сохранения, модификации самих пользовательских запросов, а результаты полнотекстового поиска, полученные в различных информационно-поисковых системах, могут быть индивидуализированы путем отнесения к определенному пользователю, который и осуществил запрос. Такая персонификация позволяет экспортировать запрос и, соответственно, результаты, проводить дальнейший мониторинг с оповещением пользователя об изменении результатов запросов.

Однако ряд проблем в поисковых системах, реализующих полнотекстовый поиск, остаются открытыми[7]. Продолжаются работы по совершенствованию алгоритмов реализации полнотекстового поиска, по разрешению проблемы индексирования текстов, которая состоит в том, что от ключевых слов (индексов) требуется соблюдение двух взаимоисключающих принципов: ключевые слова должны как можно точнее идентифицировать текст; ключевые слова должны как можно более

точно отражать смысл текста.

Предлагаются различные варианты моделей полнотекстового поиска, сравнительная характеристика которых будет рассмотрена в следующей главе курсовой работы.

Глава 2. Сравнительный анализ поисковых систем

2.1 Достоинства и недостатки поисковых систем

Для облегчения поиска и ориентирования в постоянно растущем объёме информации создаются различные поисковые средства. В распоряжении пользователей интернета достаточно много поисковых систем, которые по охвату индексируемых сайтов можно разделить на две группы:

- глобальные, осуществляющие поиск по всем сайтам сети (например Google, Bing, Yandex и т.д.);
- локальные, встроенные в один или несколько родственных сайтов, которые ведут поиск только по ним.

Стоит отметить, что почти в се глобальные поисковые системы могут использоваться и в качестве локальных, однако относить их к этой группе неправомерно, поскольку поиск по отдельному сайту для них является уточнением запроса.

Все эти системы обладают определёнными достоинствами, в числе которых простота и удобство использования, что позволяет неподготовленному пользователю сразу приступить к поиску информации; ранжирование или сортировка результатов поиска от наиболее релевантных к менее релевантным; отображение заголовка страницы и небольшого экстракта (обычно 2–3 строки) рядом со ссылкой на сайт, что позволяет составить первое впечатление о релевантности сайта или выданного результата.

Вместе с тем все эти системы обладают общими недостатками:

- коммерциализованность: большинство этих систем коммерческие, основная их цель – приносить прибыль, поэтому они часто и не всегда к месту размещают рекламу, а также «продвигают» сайт, т.е. искусственно повышают его

релевантность;

- уязвимость: поскольку механизмы индексации поисковых систем автоматические, это позволяет создателям страниц вводить для повышения релевантности ключевые слова, которые не имеют отношения к содержанию страницы, но при этом видны только поисковым системам (т.е. при загрузке страницы они не отображаются);
- сортировка только по релевантности: не учитывается дата создания страницы, поэтому очень часто на первых страницах результатов поиска идут ссылки на релевантные, но устаревшие материалы;
- избыток релевантных ссылок, число которых иногда доходит до нескольких миллионов;
- отсутствие уточнения запроса по интересующим областям;
- иногда отсутствуют релевантные ссылки.

Каждая поисковая система старается улучшить результаты поиска и избавиться от перечисленных выше недостатков или хотя бы минимизировать их. Одни системы пытаются совершенствовать алгоритмы поиска, другие – предлагают пользователю уточнить поисковый запрос.

Многие поисковые системы реализовали функцию «подсказок», которая при наборе текста в поисковом поле выдаёт небольшой список наиболее часто встречающихся запросов. Большинство глобальных поисковиков предлагают уточнить запрос по типу информации, например: Yandex – выбрать из небольшого списка (Поиск, Картинки, Видео, Карты, Маркет, Новости, Музыка, Диск, Перевод, Почта, Словари, Всё), что именно ищет пользователь.

Некоторые поисковики обеспечили пользователям возможность задать временные рамки запроса. Например, Google предлагает либо выбрать из списка период создания страниц, либо задать собственный временной интервал. Также некоторые поисковые системы для уточнения поиска предлагают воспользоваться специальными операторами и пунктуацией.

В качестве примера приведём некоторые операторы и знаки пунктуации, предлагаемые Google[8]:

«*» (звёздочка) служит для замены любого слова в запросе;

«-» (дефис) – для исключения слова из запроса;

«”текст”« (текст в кавычках) – для поиска полной фразы, заключённой в кавычки;

«OR» (оператор «ИЛИ») – для поиска одного из слов , разделённых этим оператором, и т.д.

Система уточнения запросов, несомненно, полезна, однако поиск научных статей в глобальных поисковиках по-прежнему затруднён, поскольку производится по всей сети. Для исключения сайтов, не содержащих научной информации, компания Google предлагает воспользоваться поисковой системой – Академией Google (Google Scholar)[9] , которая ведёт поиск научных публикаций по статьям как со свободным, так и с ограниченным или платным доступом. Результаты поиска представляют собой ссылки либо на полный текст статьи, либо на страницу с кратким описанием.

Эта поисковая система имеет также небольшую систему уточнения запросов: уточнение времени публикации; выбор сортировки результатов поиска (по релевантности или по дате); возможность включить в результаты поиска либо исключить из них патенты, показывать либо скрывать цитаты.

Однако Академия Google обладает такими серьёзными недостатками, как недостаточность данных об охвате базы данных; неизвестная частота обновления; отсутствие опубликованного списка научных журналов, представленных в базе данных.

Таких недостатков нет у коммерческих поисковиков (например, Web of Science[10]), являющихся локальными.

Одна из лучших систем уточнения запросов создана для поисковой системы сайта Web of Science, представляющего собой реферативную базу данных публикаций в научных журналах (компания Thomson Reuters). В боковой панели слева расположены все доступные типы уточнения поиска, например: базы данных; направления исследования; авторы; годы публикаций; языки; страны/территории и т.д. В каждом из этих типов есть небольшой список наиболее часто встречающихся вариантов во всех документах основного запроса. Например, на запрос «folksonomy» и для типа уточнения «Годы публикаций» предлагается следующий список: 2016, 2017, 2015, 2014, 2018.

Это означает, что наибольшее количество публикаций по основному запросу «folksonomy» пришлось на 2016 г.

Система также предлагает воспользоваться операторами поиска (например: «AND» для поиска записей, содержащих все условия) и символами усечения (например:

«?» (знак вопроса) для замены одного символа).

При всех достоинствах эта поисковая система обладает одним существенным недостатком, особенно для русскоязычных пользователей. Несмотря на то, что сайт русифицирован, запрос в основной базе данных – Web of Science Core Collection – вводится только латинскими символами, а значит, возникают сложности с транслитерацией. Зачастую автор транслитерирует свою фамилию и имя по-разному в разных публикациях, поэтому сотрудники компании Thomson Reuters предлагают пользоваться символами усечения, однако это не уменьшает количество результатов, а наоборот увеличивает.

Необходимо обратить внимание и на то, что основной вид поиска всех упомянутых нами поисковых систем – вербальный, т.е. базируется на естественном языке. Поэтому релевантная информация, опубликованная на одном языке, при поиске по ключевым словам на другом языке не попадает в результаты поиска. Это один из главных недостатков вербального поиска.

Компания Thomson Reuters попыталась обойти это ограничение и приняла решение вести основную БД Web of Science Core Collection на английском языке. Несомненно, английский является языком международного общения, однако не все люди хорошо владеют им, поэтому предпочитают искать информацию на родном языке.

Рейтинг популярных систем мира по данным исследовательской компании NetMarketShare представлен на рисунке 1.

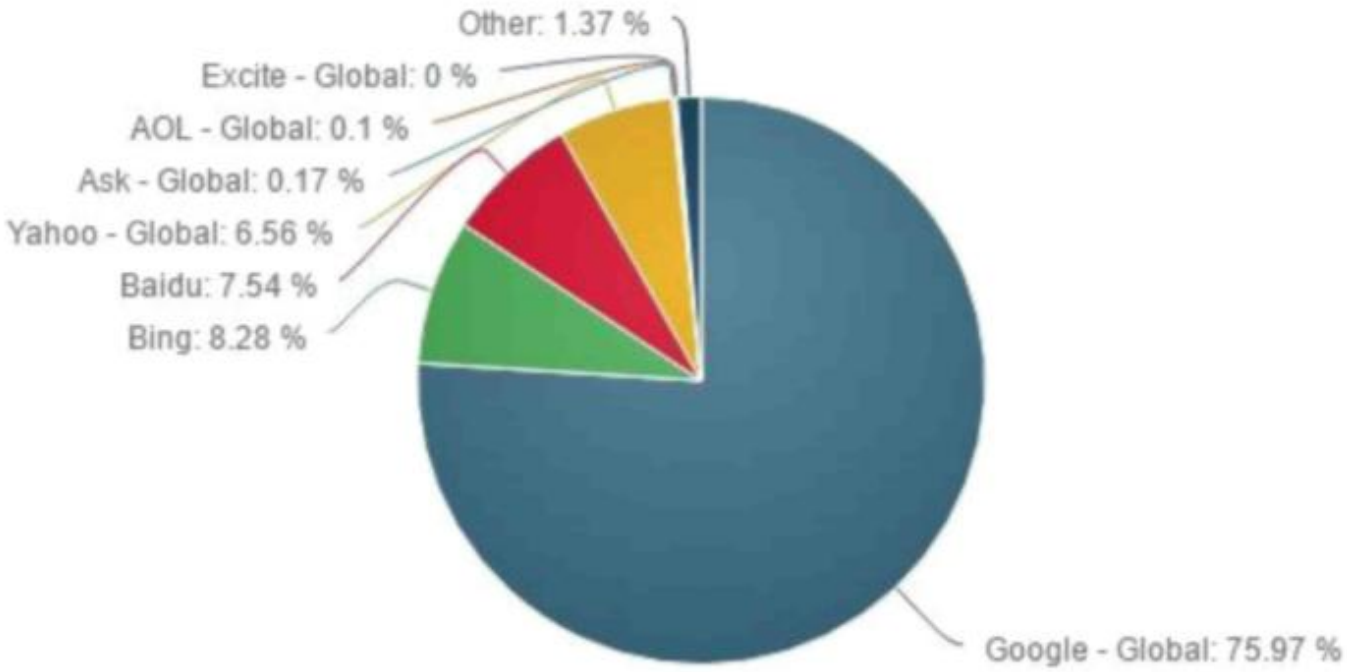

Рисунок 1. Рейтинг поисковых систем мира по популярности[11]

Популярными поисковыми системами в России по данным российского онлайн – сервиса Liveinternet на ноябрь 2018 года являются следующие[12]:

Яндекс – 50,7%;

Google – 45%;

Mail – 3,9%;

Rambler – 0,2%;

Bing – 0,1%.

Опираясь на данные статистики, в рамках курсовой работы рассматриваемыми системами будут Google, Yandex, Bing, Mail.ru и Rambler.

Google - На сегодняшний день, система Google является общепризнанным лидером среди поисковых систем мира. Появление системы произошло в 1996 году, а корпорации Google - двумя годами позже. Google - это не только поиск, но и еще более 50 сервисов, включая самый популярный браузер Google Chrome. По мнению многих специалистов, Google Chrome самый быстрый браузер в мире, на сегодняшний день. Что касается оценки пользователей, то претензий к скорости

работы не было выявлено, браузер открывает страницы практически мгновенно.

Yandex - Крупнейшая поисковая система. Появление системы произошло 23 сентября 1997 года. В последние годы Яндекс активно выходит на международный уровень. Сейчас он имеет версии сервиса в Беларуси, Украине, Казахстане и Турции. В последнее время Yandex активно продвигает свой собственный браузер.

Bing - Поисковик компании Microsoft, который быстро набирает популярность. Появление Bing произошло 1 июня 2009 года. На 2016-й год ее можно назвать быстроразвивающейся поисковой системой с достаточной долей рынка, и это позволяет назвать её конкурентом Google.

Mail.ru - Поисковая система, появление которой произошло 16 октября 2006 года. Сейчас ей принадлежат такие сервисы, как «Одноклассники» - социальная сеть для нахождения новых и старых знакомых, виртуального общения, обмена информацией между пользователями, которые смогут разделить общие интересы и увлечения, «Мой мир» - сеть, для поиска новых знакомых, друзей, одноклассников, обмена сообщениями, размещения фото и видео, поиска групп по интересам и Афиша, Агент, «Вопросы и ответы», Майл Деньги —около 40 крупнейших сервисов в Рунете, среди которых и сам поиск. Mail.ru занимает третью строчку после Google и Яндекс среди популярных поисковиков в России.

Rambler – Поисковая система, существовавшая с 1996 по 2011 года. На сегодняшний день это крупнейший российский интернет-портал. Поиск по Rambler осуществляется силами движка Яндекса, объективных причин падения его популярности нет.

2.2 Сравнительный анализ обработки запросов поисковыми системами

Рассмотрим пятерку поисковых систем по двум главным характеристикам: по полноте и точности поиска. Качество поиска в информационно-поисковых системах можно определить двумя критериями –точностью и полнотой. Точность определяется соотношением между найденными релевантными и нерелевантными документами, а полнота поиска - общим количеством найденных документов. Релевантным будем считать документ, который удовлетворяет запросу пользователя. Нерелевантные документы, сравниваемые с релевантными, иногда могут называться шумом, по аналогии с теорией передачи информации.

Релевантные документы в таком случае называют сигналом, а эффективность поиска оценивают по соотношению «сигнал – шум»[13].

Назначим весовые коэффициент - параметры, которые отражают в сравнении с другими критериями относительную важность, значимость, «вес» данных критериев. Сумма всех весов должна быть равной 1, поэтому для точности поиска весовому коэффициенту даем значение, равное 0.8, для полноты поиска – 0.2. Оформим результаты в виде таблицы 1.

Таблица 1

Весовые коэффициенты

Критерий	**Весовой коэффициент**
Точность поиска	0,8
Полнота поиска	0,2

Сформулируем тринадцать запросов на разные темы и выполним каждый запрос в каждой из пяти исследуемых поисковых системах. Из полученных списков результатов выберем следующую информацию:

Общее количество найденных документов (Д).

Количество релевантных документов различной ценности (РД)

Количество релевантных документов оценивается при просмотре текста первых 10 найденных документов. Также определяется ценность найденной информации (степень удовлетворения найденном документом информационных потребностей). Ценность информации оценивается по 3-х бальной шкале: 2 балла - информация имеет ценность, 1 балл – информация имеет частичную ценность, 0 баллов – информация не имеет ценности[14]. Результаты выполнения запросов сведем в таблицу 2, а затем выполним первичную обработку результатов (таблица 3, строка 1).

Для нахождения лучшей поисковой системы для начала вычислим средние арифметические значения показателей для каждой поисковой системы Д, РД(0),

РД(1) и РД(2).

Далее необходимо определить место каждой поисковой системы по критерию «Полнота поиска». Для этого будем использовать среднее количество найденных документов Д. Наилучшей считается та система, которая нашла больше документов. Ей присваивается первое место, самой худшей – место N (где N – это количество всех исследуемых систем). Коэффициент точности поиска Р для каждой поисковой системы определим по формуле:

P = a / (a+b) ,

где a – число релевантных документов, которые выдала поисковая система в ответ на запрос, a = 0.5 * РД(1)+ РД(2) ;

b - число документов, которые полностью не имеют ценность, b=РД(0).

Далее необходимо определить место каждой поисковой системы по критерию «Точность поиска». Лучшей будет считаться система, которая имеет большее значение коэффициента точности поиска Р. Ей присваивается первое место, самой худшей - место N (где N - это количество исследуемых систем).

Таблица 2

Результаты выполнения запросов

	Bing				**Google**				**Mail.ru**				**Rambler**				**Yandex**				
№ темы	**Д**	**РД**			**Д**	**РД**			**Д**	**РД**			**Д**	**РД**			**Д**	**РД**			**РД**
		2	**1**	**0**		**2**	**1**	**0**		**2**	**1**	**0**		**2**	**1**	**0**		**2**	**1**	**0**	**2**
1	1 810 000	9	1	0	3 180 000	10	0	0	7 000 000	5	2	3	41 000 000	9	1	0	40 000 000	9			
2	116 000	8	2	0	711 000	10	0	0	2 000 000	8	1	1	943 000	10	0	0	942 000	9			
3	420 000	7	2	1	2 330 000	8	2	0	1 000 000	9	1	0	3 000 000	8	0	2	2 000 000	9			

4	62 000	8	2	0	964 000	10	0	0	1 000 000	8	1	1	3 000 000	9	1	0	2 000 000	9
5	2 340 000	8	1	1	2 380 000	8	0	2	7 000 000	7	1	2	2 000 000	9	0	1	17 000 000	9
6	103 000	7	0	3	1 020 000	10	0	0	1 000 000	9	0	1	3 000 000	10	0	0	2 000 000	10
7	746 000	9	0	1	5 430 000	9	0	1	13 000 000	10	0	0	3 000 000	9	0	1	1 000 000	9
8	19 900	7	2	1	214 000	10	0	0	18 000	9	1	0	9 000	8	0	2	9 000	10
9	42 400	9	1	0	140 000	9	0	1	766 000	7	1	2	431 000	9	0	1	430 000	10
10	999 000	10	0	0	4 190 000	9	0	1	9 000 000	9	1	0	6 000 000	10	0	0	4 000 000	10
11	1 940 000	10	0	0	683 000 000	9	1	0	9 000 000	10	0	0	7 000 000	10	0	0	6 000 000	9
12	73 700	9	1	0	812 000	8	1	1	863 000	9	0	1	834 000	9	1	0	869 000	10
13	115 000	10	0	0	301 000	10	0	0	3 000 000	9	0	1	2 000 000	10	0	0	2 000 000	10

Таблица 3

Результаты сравнительного анализа поисковых систем

Критерий	**Bing**	**Google**	**Mail.ru**	**Rambler**	**Yandex**
Полнота поиска (Д)	675923,1	1719615,4	4203615,4	6939769,2	6019230,8
Место(полнота поиска)	5	4	3	1	2

Среднее количество пертинентных документов (РД2)	8,5	9,23	8,38	9,23	9,46
Среднее количество частично пертинентных документов (РД1)	0,92	0,31	0,69	0,23	0,15
Среднее количество непертинентных документов (РД0)	0,53	0,46	0,92	0,53	0,38
Коэффициент точности поиска (Р)	0,94	0,953	0,904	0,945	0,961
Место (точность поиска)	4	2	5	3	1
Коэффициент поискового шума (S)	0,056	0,046	0,095	0,054	0,038

Следующим шагом будет вычисление коэффициента поискового шума S по формуле: S=1 - P

В заключении необходимо вычислить по следующей формуле рейтинг каждой исследуемой системы R:

где i - номер критерия оценки поисковой системы,

m - это количество критериев оценки,

w_i - весовой коэффициент для критерия оценки i,

q_i - это место поисковой системы по критерию оценки i.

N - общее количество исследуемых систем.

Результат расчета рейтинга приведен в табл.4

Таблица 4

Результаты сравнительного анализа поисковых систем

Критерий	**Bing**	**Google**	**Mail.ru**	**Rambler**	**Yandex**
Рейтинг (R)	2,8	4,6	2,4	4,4	**5,8**

По результатам расчетов лучшей поисковой системой Интернет из исследуемых - является Yandex.

Заключение

В целом поисковая система представляет собой совокупность информационно-поискового языка, программных средств и правил перевода текстов на этот язык (индексирования), а также обеспечения поиска по заданным критериям. Главная цель поисковых систем – предоставить пользователям доступ к сайтам, информация на которых наиболее релевантна их запросам. Принцип работы поисковой системы следующий: поисковый робот обходит страницы сайта, сканирует их содержимое (код, стили, контент, изображения, ссылки и др.), далее страница отправляется на индексирование, где на основе алгоритмов начинается анализ всех собранных роботом материалов. Таким образом, поисковая система создаёт базу данных, где хранятся все обработанные алгоритмом документы. У каждой поисковой системы разработаны собственные алгоритмы оценивания качества сайтов.

В процессе выполнения данной работы был проведен анализ популярных среди пользователей поисковых систем. Была проанализирована пятерка систем, а именно поисковые системы Yandex, Google, Mail.ru Bing, и Rambler, произведено их сравнение и, была выбрана лучшая система. Опираясь на расчеты, можно с уверенностью сказать, что таковой является Яндекс. Поставленные задачи были

полностью выполнены. Результат работы поможет пользователям сети выбрать быструю и надежную поисковую систему, выполняющую запросы с наибольшей точностью и за максимально короткие промежутки времени. Также не стоит забывать, что от содержания самого запроса зависит и скорость его воспроизведение, поэтому рекомендациями составления запроса могут выступить: учет морфологии слов, четкость и составление запроса из нескольких слов, адекватно передающих содержание необходимой информации.

Список использованных источников

1. ГОСТ 7.73-96. Поиск и распространение информации. Термины и определения.
2. Адаманский А. Обзор методов и алгоритмов полнотекстового поиска [Электронный ресурс] – Режим доступа: www.dialog-21.rudialog2016materialshtmlFedorovsky.htm.
3. Борисова Н.В., Кочуева З.А. Индексирование полнотекстовых документов для задачи интеллектуального поиска информации по ключевым словам//Восточно-Европейский журнал передовых технологий. – 2014. – № 1/2(67). – С. 4–8. URL: http://journals.uran.ua/eejet/article/view/20332.
4. Васенин В.А. Управление тематическими данными в больших и сверхбольших хранилищах: механизмы, модели, программное обеспечение (состояние, задачи, решения) // Проблемы информатики, № 1, 2018. – С. 71–84.
5. Вишняков Ю.М., Вишняков Р.Ю. Вычислительная семантическая интерпретация текстов научно-технического стиля // Современные наукоемкие технологии, 2016, № 12-2. – С. 53–30.
6. Королева О.Н., Мажукин А.В. Поисковые системы сети INTERNET. – М.: Изд-во Моск. гуманитар. ун-та, 2017. – 35 с.
7. Мировые информационные ресурсы [Текст]:Учебное пособие/ В.К.Иванов; под. ред .В. К.Иванова.-Тверь:Изд-во ин-та ТвГТУ, 2017. - 37с.
8. Симакина Н.И. Разработка подсистемы полнотекстовой индексации и полнотекстового поиска для платформы облачного контент-репозитория / Симакина Н. И., Шипулина К. В., Костарев А. А., Окунев А. Ф. // Вестник Пермского университета. Серия: Математика. Механика. Информатика 2014, №4. – С. 92–96.
9. Рейтинг топ 5 самых лучших отечественных и мировых поисковых систем: [Электронный ресурс]. М. – URL: bestseoblog.ru
10. Статистика сайта. Переходы из поисковых систем: [Электронный ресурс]. М. – URL: liveinternet.ru

11. Beall J. TheWeaknesses of Full-Text Searching // Journal of Academic Librarianship, Sep. 2018, Vol.34, No.5, pp. 438–444.
12. Dong-Jin Kim, Sang-Chul Lee, Ho-Yong Son, Sang-Wook Kim, Jae Bum Lee. C-Rank and its variants: A contribution-based ranking approach exploiting links and content // Journal of Information Science 1(18), September 2014, pp. 761–778.
13. Franklin M., Halevy A., Maier D. From Databases to Dataspaces: A New Abstraction for Information Management // SIGMOD Record. 2015, Vol. 34, No. 4; URL: http://www.sigmod.org/ sigmod/record/issues/0512/p27-article-franklin.pdf .
14. Kai A. Olsen, Kenneth M. Sochats, & James G. Williams. Full Text Searching and Information Overload // International Information & Library Review, 30 (June, 1998), pp. 105–122.
15. Yury M. Vishnyakov, RenatYu. Vishnyakov/ The Linguistic Proximity in Information Retrieval and Document Classification. 14th IEEE International Symposium on Computational Intelligence and Informatics to be held on November 19-21, 2013 in Budapest, Hungary. – P. 131–134.
16. Yury Vishnyakov, Renat Vishnyakov // Representation of semantically cohesivesenten cefragments inscientificandte chnicaltexts // 2014 IEEE 12th InternationalSymposiumonAppliedMachineIntelligenceandInformatics, pp. 295–298, DOI: 10.1109/SAMI.2014. 6822425
17. IDC – Режим доступа: http://www.idc.com/home.jsp
18. Операторы в поисковых запросах. – Режим доступа: https://support.google.com /websearch/answer/2466433?hl=ru
19. Академия Google. – Режим доступа: https://scholar.google.ru/
20. Web of Science. – Режим доступа: http://apps.webofknowledge.com/

1. IDC – Режим доступа: http://www.idc.com/home.jsp ↑

2. Вишняков Ю.М., Вишняков Р.Ю. Вычислительная семантическая интерпретация текстов научно-технического стиля // Современные наукоемкие технологии, 2016, № 12-2. – С. 53–30. ↑

3. Васенин В.А. Управление тематическими данными в больших и сверхбольших хранилищах: механизмы, модели, программное обеспечение (состояние, задачи, решения) // Проблемы информатики, № 1, 2018. – С. 71–84. ↑

4. Franklin M., Halevy A., Maier D. From Databases to Dataspaces: A New Abstraction for Information Management // SIGMOD Record. 2015, Vol. 34, No. 4; URL: http://www.sigmod.org/ sigmod/record/issues/0512/p27-article-franklin.pdf . ↑

5. Королева О.Н., Мажукин А.В. Поисковые системы сети INTERNET. – М.: Изд-во Моск. гуманитар. ун-та, 2017. – 35 с. ↑

6. Адаманский А. Обзор методов и алгоритмов полнотекстового поиска [Электронный ресурс] – Режим доступа: www.dialog-21.rudialog2016materialshtmlFedorovsky.htm. ↑

7. Beall J. TheWeaknesses of Full-Text Searching // Journal of Academic Librarianship, Sep. 2018, Vol.34, No.5, pp. 438–444. ↑

8. Операторы в поисковых запросах. – Режим доступа: https://support.google.com /websearch/answer/2466433?hl=ru ↑

9. Академия Google. – Режим доступа: https://scholar.google.ru/ ↑

10. Web of Science. – Режим доступа: http://apps.webofknowledge.com/ ↑

11. Рейтинг топ 5 самых лучших отечественных и мировых поисковых систем: [Электронный ресурс]. М. – URL: bestseoblog.ru ↑

12. Статистика сайта. Переходы из поисковых систем: [Электронный ресурс]. М. – URL: liveinternet.ru ↑

13. Мировые информационные ресурсы[Текст]:Учебное пособие/ В.К.Иванов; под. ред .В. К.Иванова.-Тверь:Изд-во ин-та ТвГТУ, 2017. - 37с. ↑

14. Мировые информационные ресурсы [Текст]:Учебное пособие/ В.К.Иванов; под. ред .В. К.Иванова.-Тверь:Изд-во ин-та ТвГТУ, 2017. - 37с. ↑