УДК 004 **DOI: 10.34670/AR.2022.78.30.020**

Поисковые сайты и технологии поиска информации

Мизаев Мансур Мовсарович

Старший преподаватель,
кафедра ,изнес-информатикb,
Чеченский государственный университет им. А. А. Кадырова,
364049, Российская Федерация, Грозный, ул. Шерипова, 32;
e-mail: m.mizaev@chesu.ru;

Мусаева Мархa Сайд-Магомедовна

Ассистент кафедры информатики и вычислительной техники,
Грозненский государственный нефтяной технический
университет им. академика М. Д. Миллионщикова,
364061, Российская Федерация, Грозный, ул. Р.Х. Исаева, 100;
e-mail: markha.musaeva.98@mail.ru;

Аннотация

В статье освещается вопрос поиска информации и технологий осуществления данного процесса. На сегодняшний день существует огромное множество поисковых систем, которые посредством самых разных механизмов помогают человеку осуществлять поиск сведений и находить им нужное применение. Отмечается, что поиск информации – это широкая междисциплинарная область, которая опирается на многие другие дисциплины. Авторы рассматривают саму систему поиска информации с самого начала ее пути вместе с огромным количеством различных моделей и механизмов. Говорится о том, что автоматизированные информационно-поисковые системы первоначально использовались для управления информационным взрывом в научной литературе в последние несколько десятилетий, однако сейчас они доступны большинству и приносят пользу в самых разных сферах жизнедеятельности человека.

Для цитирования в научных исследованиях

Мизаев М.М., Мусаева М.С.-М. Поисковые сайты и технологии поиска информации // Экономика: вчера, сегодня, завтра. 2022. Том 12. № 4А. С. 203-209. DOI: 10.34670/AR.2022.78.30.020

Ключевые слова

Информация, поиск, технологии поиска информации, информационно-поисковые системы, модели теории множеств, алгебраические модели, вероятностные модели.

Введение

В современном информационном обществе данные, факты и знания имеют гораздо более высокий приоритет, чем полвека назад. Благодаря Интернету информация становится все более и более доступной. Человек очень быстро научился использовать онлайн-площадки для того, чтобы получить те знания, которые ему нужны.

Каким же образом происходит поиск и предоставление информации? Ответ кроется в самих словах «поиск информации». Сбор информации – это целое направление в области информатики, прежде всего, оно имеет большое значение для поисковых сайтов. Используя сложные информационно-поисковые системы, можно распознать намерения, стоящие за конкретными поисковыми терминами, и найти соответствующие данные в поисковых запросах.

Основная часть

Поиск информации заключается в том, чтобы сделать существующие знания доступными. Так было задолго до начала цифровой эры. Ванневар Буш, один из первых людей, которые серьезно задумались о том, как человечество может сделать свои концентрированные знания более доступными в условиях постоянно меняющегося мира, опубликовал новаторскую статью на заре информационной эры (1945) под названием «Как мы можем думать». В статье было представлено видение будущего сбора и организации информации [Байков, 2014].

Буш увидел следующую проблему в науке: эксперты все больше и больше специализируются и вместе с тем нуждаются в большем количестве информации, но из-за дифференциации, вызванной этой крайней специализацией, информацию становится все труднее найти. Конечно, это было в то время, когда библиотеки все еще были организованы с аналоговыми бумажными коробками и большими каталогами.

Поиск по ключевым словам был возможен только в том случае, если прилежный библиотекарь уже потрудился вручную проиндексировать все работы. Буш увидел свой собственный уникальный способ сделать информацию более доступной, используя технические разработки, доступные в то время, такие как микрофильмы. Его видение состояло в том, чтобы создать механизм, который представлял собой машину размером с письменный стол, которая служила бы хранилищем знаний и стала бы серьезным исследовательским оборудованием. Пусть его идея так и не была реализована, но саму технологию (пользователи переходят от одной статьи к другой) можно рассматривать как предшественницу гипертекста.

В 1950-х годах ученый-компьютерщик Ханс Петер Лун специально занимался задачей получения информации и разработал методы, которые по-прежнему актуальны сегодня: полнотекстовая обработка; автоматическая индексация; выборочная обработка информации (SDI) [там же].

Эти методы были очень важны для развития Интернета, поскольку информационно-поисковые системы нуждаются в них для осуществления навигации по океанам доступной информации в Интернете. Без них мы никогда не смогли бы найти ответы, которые ищем.

Цель поиска информации (IR) состоит в том, чтобы сделать данные, хранящиеся в машине, доступными для обнаружения: в отличие от интеллектуального анализа данных, который извлекает структуры из онлайн-записей, IR занимается фильтрацией конкретной информации из набора данных. Типичным приложением является поисковая система в Интернете.

Информационно-поисковые системы решают две основные проблемы:

1. Неопределенность: запросы пользователей часто неточны. Поисковые запросы, введенные пользователем, часто оставляют много места для интерпретации. Например, те, кто ищет термин «банк», могут искать общую банковскую информацию или могут потребовать указания направления в ближайшее финансовое учреждение. Проблема усугубляется, когда сами пользователи не уверены, какую информацию они ищут.

2. Содержание хранимой информации иногда неизвестно системе, что приводит к тому, что пользователям представляются неправильные результаты. Это происходит, например, с омонимами – словами, имеющими несколько значений. Пользователь может искать не финансовое учреждение, а информацию о географическом объекте, связанном с реками.

Кроме того, информационно-поисковая система также должна оценивать информацию, чтобы предоставить пользователям последовательность данных. Первый результат в идеале должен дать наилучший ответ на вопрос пользователей.

Технология поиска информации представлена различными моделями, которые не обязательно являются взаимоисключающими и могут быть объединены друг с другом.

Некоторые из этих моделей лишь отличаются незначительными деталями. Тем не менее, все они все еще могут быть примерно разделены на три группы:

1. Модели теории множеств. Отношения подобия определяются операциями с множествами (Булева модель).

2. Алгебраические модели: Сходство определяется парами: документы и поисковые запросы могут быть представлены в виде векторов, матриц или кортежей (модель векторного пространства).

3. Вероятностные модели: Эти модели устанавливают сходство, рассматривая наборы данных как многоступенчатые случайные эксперименты [Щербаков, 2012].

Вышеперечисленные модели очень часто используются вместе, точно так же как свое предназначение нашли и их гибриды.

Булева модель

Самые популярные поисковые системы в Интернете основаны на логическом принципе. Это логические ссылки, которые помогают пользователям либо уточнить, либо точно определить поиск. С И, ИЛИ или НЕТ (И, ИЛИ, НЕТ) или соответствующими символами ∧, ∨ запрос может быть указан, когда, например, в результате должны появиться оба термина или содержимое с определенным термином должно быть скрыто.

Эти ключи также работают в Google по тому же принципу. Недостатком этой системы является то, что она не содержит никакой системы ранжирования результатов.

Модель Векторного Пространства

При математическом подходе содержимое также может быть представлено в виде векторов. В модели векторного пространства термины отображаются в виде осей координат. Как документы, так и поисковые запросы получают определенные значения, связанные с термином, и могут быть представлены в виде точек или векторов в векторном пространстве. Впоследствии оба вектора сравниваются друг с другом.

Вектор (или содержимое), ближайший к запросу, должен отображаться первым в рейтинге результатов. Недостатком здесь является то, что без булевых операторов никакие термины не могут быть исключены.

Вероятностная модель

Вероятностная модель использует теорию вероятностей. Каждому документу присваивается значение вероятности. Затем результаты сортируются в соответствии с

вероятностью, с которой они соответствуют каждому поиску.

Насколько высока вероятность того, что определенный контент соответствует пожеланиям пользователя, определяется так называемой «обратной связью по релевантности». Например, пользователям может быть предложено оценить результаты вручную. При следующем идентичном поиске модель покажет другой (возможно, лучший) список результатов.

Недостатком этой процедуры является то, что она начинается с двух требований, ни одно из которых не гарантируется. С одной стороны, модель предполагает, что пользователи готовы участвовать в системе, предоставляя обратную связь. С другой стороны, теория также предполагает, что пользователи просматривают результаты независимо друг от друга, оценивая содержимое каждого источника так, как если бы это было первое, что они прочитали в поиске. На практике пользователи всегда ценят информацию, основанную на ранее просмотренном контенте или имеющихся знаниях.

При поиске информации используются различные методы и приемы, независимо от моделей. Цель всегда состоит в том, чтобы упростить поиск информации для пользователя и обеспечить более релевантные результаты.

Важность термина для поискового запроса рассчитывается путем объединения частоты встречаемости и обратной частоты документа. Это значение сокращенно обозначается как tf-idf [Информация и формы ее представления, www].

Частота терминов

Плотность поисковых слов указывает, как часто термин появляется в документе. Однако частота появления термина не может быть единственным показателем того, насколько релевантен текст, поскольку некоторые тексты могут содержать слово несколько раз из-за длины, а не релевантности содержания. Поэтому частота должна быть рассчитана в зависимости от размера документа.

Обратная частота документов

В IDF рассматривается весь текст, а не только один документ. Слова, которые встречаются только в нескольких документах, будут иметь более высокую актуальность, чем термины, которые встречаются почти во всех текстах [Информация и формы ее представления, www].

Объединив два текста, информационно-поисковые системы могут обеспечить лучшие результаты, чем если бы они использовались по отдельности: если бы важна была только частота терминов, то поисковый запрос «Телешоу с помощью мыши» определил бы приоритет контента, в котором появляются слова «the» и «with». Это, очевидно, было бы бесполезно. Напротив, если используется обратная частота документов, «телешоу» и «мышь» гораздо важнее для поиска и распознаются как фактические условия поиска.

Изменение запроса

Основной проблемой при сборе информации является поведение самих пользователей: крайне неточные запросы приводят к появлению неверной или неадекватной информации. Чтобы избежать этого, специалисты по информатике внедрили модификацию запросов – систему, которая автоматически изменяет введенный поисковый запрос [Поиск информации, www].

Это означает, например, что используются синонимы, которые обеспечивают лучшие результаты. Система использует тезаурусы и отзывы пользователей, чтобы найти эти синонимы.

Чтобы избежать зависимости от сотрудничества с пользователем, они могут использовать так называемую «псевдо-обратную связь». С помощью этого метода система считывает соответствующие термины из лучших результатов поиска и оценивает их как релевантные для

поиска.

Запросы могут быть расширены или улучшены с помощью следующих методов:

1. Исключение стоп-слов. Стоп-слова – это те выражения, которые лишь незначительно влияют на содержание текста. Имеет смысл не рассматривать такие слова, как «и», или такие статьи, как «the», как репрезентативные для содержания документа.

2. Идентификация групп из нескольких слов. Группы слов должны быть распознаны как таковые. Эта идентификация гарантирует, что поисковая система также считает части составных слов релевантными.

3. Сокращение до корня и формы корня. Для более эффективного поиска слова должны быть сведены к их корневым словам. В противном случае флективные формы слова не будут корректно отображаться в результатах поиска.

4. Тезаурус. В дополнение к терминам, используемым в соответствующем документе, информационно-поисковая система также должна рассматривать синонимы слова как соответствующие. Это единственный способ гарантировать, что пользователи найдут то, что они ищут.

Отзыв и точность

Эффективность информационно-поисковой системы обычно рассчитывается с использованием коэффициентов скорости и точности отзыва. Оба фактора представлены в виде частных.

Отзыв: Насколько полны результаты поиска?

Для этого количество «найденных, соответствующих» сравнивается с количеством «ненайденных, соответствующих документов». Другими словами, коэффициент указывает, насколько вероятно, что соответствующий документ будет найден:

Отзыв = ((Найденные, релевантные документы)) / ((Найденные, релевантные документы) + (ненайденные, релевантные документы))

Точность: каков именно результат поиска?

Чтобы убедиться в этом, указывается количество найденных, соответствующих количеству найденных, не относящихся к делу документов. Коэффициент указывает, насколько вероятно, что найденный документ имеет отношение к делу:

Точность = ((Найденные, релевантные документы)) / ((Найденные, релевантные документы) + (Найденные, неуместные документы))

Оба значения в основном находятся между 0 и 1, где 1 было бы идеальным значением. Кроме того, на практике исключаются идеальные результаты для обоих факторов. Те, кто увеличивает полноту результатов поиска, делают это за счет точности, и наоборот. Кроме того, выпадение (т.е. частота по умолчанию) может быть рассчитано как дополнительное значение: этот коэффициент отражает частоту ложных срабатываний; он определяется отношением найденных, не относящихся к делу документов к не относящемуся к делу содержимому, которое не было найдено. Отзыв и точность могут быть представлены в виде диаграммы осей, в которой каждое из двух значений занимает по одной оси каждое [Щербаков, 2012].

Заключение

Таким образом, поиск информации – это широкая междисциплинарная область, которая опирается на многие другие дисциплины. Поскольку она настолько широка, она обычно плохо понимается, и к ней обычно подходят только с той или иной точки зрения. Она находится на

стыке многих устоявшихся областей и опирается на когнитивную психологию, информационную архитектуру, информационный дизайн, информационное поведение человека, лингвистику, семиотику, информатику, информатику и библиотечное дело. Веб-поисковые системы, такие как Google и Lycos, являются одними из наиболее заметных приложений для поиска информации. Автоматизированные информационно-поисковые системы первоначально использовались для управления информационным взрывом в научной литературе в последние несколько десятилетий, однако сейчас они доступны большинству и приносят пользу в самых разных сферах жизнедеятельности человека.

Библиография

1. Ашманов И.С. Идеальный поиск в Интернете глазами пользователя / И.С. Ашманов. М.: Питер, 2011. 624 с.
2. Байков В. Интернет. Поиск информации и продвижение сайтов. М.: БХВ-Петербург, 2014. 288 с.
3. Галеева И.С. Интернет как инструмент библиографического поиска. М.: Профессия, 2007. 256 с.
4. Информация и формы ее представления. URL: http://256bit.rU/informat/eu_intro/il-l.htm#l.
5. Поиск информации. URL: https://psychology.fandom.com/wiki/Information_retrieval.
6. Поисковая система – это. URL: https://dic.academic.ru/dic.nsf/ruwiki/190.
7. Поисковая система: структура и функции. URL: https://uniofweb.ru/wiki/poiskovye_sistemy.
8. Технология поиска. URL: https://www.seonews.ru/masterclasses/tehnologiya-poiska-informatsii-v-internet-vidyi-poiskovyih-instrumentov-informatsionnyie-poiskovyie-sistemyi-interneta.
9. Щербаков А.Ю. Интернет-аналитика. Поиск и оценка информации в web- ресурсах. М.: Книжный мир, 2012. 173 с.

Search sites and information search technologies

Mansur M. Mizaev

Senior Lecturer,
Department of business informatics,
Chechen State University named after A.A. Kadyrov,
364049, 32 Sheripova str., Grozny, Russian Federation;
e-mail: m.mizaev@chesu.ru

Markha S.-M. Musaeva

Assistant of the Department of informatics and computer engineering,
Grozny State Oil Technical University Academician M. D. Millionshchikov,
Grozny State Oil Technical University
named after Academician M.D. Millionshchikov,
364061, 100 R.Kh. Isaeva ave., Grozny, Russian Federation;
e-mail: markha.musaeva.98@mail.ru;

Abstract

The article highlights the issue of searching for information and technologies for the implementation of this process. Today, there are a huge number of search engines that, through a variety of mechanisms, help a person to search for information and find the right application for it. It is noted that information retrieval is a broad interdisciplinary field that relies on many other

disciplines. Because it is so broad, it is usually poorly understood and is usually only approached from one point of view or another. It sits at the crossroads of many established fields and draws on cognitive psychology, information architecture, information design, human information behavior, linguistics, semiotics, computer science, computer science, and librarianship. The authors consider the information retrieval system itself from the very beginning of its journey, along with a huge number of different models and mechanisms. It is said that automated information retrieval systems were originally used to manage the information explosion in the scientific literature in the past few decades, but now they are available to the majority and bring benefits in various areas of human life.

For citation

Mizaev M.M., Musaeva M.S.-M. (2022) Poiskovye saity i tekhnologii poiska informatsii [Search sites and information search technologies]. *Ekonomika: vchera, segodnya, zavtra* [Economics: Yesterday, Today and Tomorrow], 12 (4A), pp. 203-209. DOI: 10.34670/AR.2022.78.30.020

Keywords

Information, search, information retrieval technologies, information retrieval systems, set theory models, algebraic models, probabilistic models.

References

1. Ashmanov I.S. (2011) *Ideal'nyi poisk v Internete glazami pol'zovatelya* [Ideal Internet search through the user's eyes]. Moscow: Piter Publ.
2. Baikov V. (2014) *Internet. Poisk informatsii i prodvizhenie saitov* [Internet. Information search and website promotion]. Moscow: BKhV-Peterburg Publ.
3. Galeeva I.S. (2007) *Internet kak instrument bibliograficheskogo poiska* [Internet as a tool for bibliographic search]. Moscow: Professiya Publ.
4. *Informatsiya i formy ee predstavleniya* [Information and forms of its presentation]. Available at: http://256bit.rU/informat/eu_intro/il-l.htm#l [Accessed 22/03/2022].
5. *Poisk informatsii* [Search for information]. Available at: https://psychology.fandom.com/wiki/Information_retrieval [Accessed 12/03/2022].
6. *Poiskovaya sistema – eto* [Search engine is]. [Available at: https://dic.academic.ru/dic.nsf/ruwiki/190 [Accessed 17/03/2022].
7. *Poiskovaya sistema: struktura i funktsii* [Search engine: structure and functions]. Available at: https://uniofweb.ru/wiki/poiskovye_sistemy [Accessed 12/03/2022].
8. Shcherbakov A.Yu. (2012) *Internet-analitika. Poisk i otsenka informatsii v web- resursakh* [Internet analytics. Search and evaluation of information in web resources]. Moscow Knizhnyi mir Publ.
9. *Tekhnologiya poiska* [Search technology]. Available at: https://www.seonews.ru/masterclasses/tehnologiya-poiska-informatsii-v-internet-vidyi-poiskovyih-instrumentov-informatsionnyie-poiskovyie-sistemyi-interneta Accessed 12/03/2022].