СИСТЕМЫ ПОИСКА ИНФОРМАЦИИ НА ОСНОВЕ АНАЛИЗА СЛАБОСТРУКТУРИРОВАННОГО ТЕКСТА

Митина Ольга Алексеевна
кандидат пед. наук.,
МИРЭА – Российский технологический университет,
г. Москва

INFORMATION SEARCH SYSTEMS BASED ON THE ANALYSIS OF WEAKLY STRUCTURED TEXT

Mitina Olga Alekseevna
Candidate of Science
MIREA – Russian Technological University
Moscow

АННОТАЦИЯ

С середины XX века проблема хранения информации и доступа к ней привлекает все больше внимания. Все крупные новостные издания уже давно создали собственные веб-страницы, где размещают свои новости. Социальные сети и форумы позволяют миллионам людей обмениваться миллиардами сообщений ежедневно. Многие корпоративные и медицинские документы хранятся на выделенных серверах. Каждое сообщение или документ представляет собой блок неструктурированного или слабоструктурированного текста. Возникает необходимость создания систем поиска информации на основе анализа слабоструктурированного текста.

ANNOTATION

Since the middle of the XX century, the problem of information storage and access to it has attracted more and more attention. All major news outlets have long created their own web pages where they post their news. Social networks and forums allow millions of people to share billions of messages daily. Many corporate and medical documents are stored on dedicated servers. Each message or document is a block of unstructured or weakly structured text. There is a need to create information search systems based on the analysis of weakly structured text.

Ключевые слова: информационный поиск; слабоструктурированные данные; системы поиска информации.

Keywords: information search; weakly structured data; information search systems.

В последнее десятилетие наблюдается увеличение объема информации в цифровом формате. Каждое сообщение или документ представляет собой блок неструктурированного или слабоструктурированного текста [1][2]. В связи с этим остро встает вопрос обеспечения быстрого и эффективного доступа к этой информации. Таким образом, для решения этого вопроса начали использоваться компьютеры, поскольку скорость их вычисления превосходит человеческую.

В процессе поиска решения в сфере обработки информации была создана отдельная дисциплина, называемая информационным поиском (information retrieval).

Задачами данной дисциплины являются анализ и структурирование данных, а также быстрый и эффективный поиск среди этих данных.

На сегодняшний день информационный поиск представляет собой междисциплинарную область науки, которая стоит на пересечении библиотечного дела, информатики, лингвистики, семиотики, когнитивной психологии и информационного дизайна [3].

Определение информационного поиска наиболее полно отражено в государственном стандарте. Информационный поиск (IR) – «все методы и процессы, используемые для того, чтобы выбрать из документной коллекции или сети информационных ресурсов документы, релевантные информационным потребностям», в котором присутствует ключевое слово – «релевантный». Релевантность – «соответствие полученной информации запросу» [4].

Информационный поиск уже стал основной формой доступа к информации, вытеснив собой традиционный поиск по идентификатору.

При хранении больших объемов информации проблема точного и быстрого доступа становится только сложнее.

Одним из последствий этой проблемы является игнорирование релевантной информации, к которой не удается получить доступ, что приводит к повторению уже проделанной работы. С развитием компьютеров особое внимание уделяется использованию интеллектуальных систем доступа к информации [5]. Многие монотонные задачи доверены компьютерам. Однако проблема эффективного доступа к информации остается нерешенной.

Принцип работы данных систем прост. Человек формулирует запрос, в ответ на который из набора документов выбираются те, что содержат удовлетворяющие требованию информацию. В идеальном варианте, система выбирает только релевантные документы, отбрасывая все

остальные. Использование человеческих ресурсов непрактично в силу затрат по времени.

Когда компьютеры смогли обрабатывать не только числа, возникла возможность «прочитать» всю коллекцию документов и извлечь нужные документы.

Вскоре стало ясно, что использование естественного языка в документах вызывает проблемы не только с хранением и вводом информации, но также требует решения проблема интеллектуальной характеризации содержимого документа. Если в скором времени развитие аппаратной части разрешит проблему с хранением и вводом данных, то автоматическая характеризация, представляющая собой попытку машиной повторить человеческий процесс чтения, является, несомненно, неприятной проблемой.

В частности, «чтение» включает в себя попытку изъять из текста и синтаксическую, и семантическую информацию и, пользуясь полученной информацией, определить релевантность документа относительно конкретного запроса. Относительно медленный прогресс в семантической части современной лингвистики и недавние неудачи попыток машинного перевода показывают, что эти проблемы далеки от решения [6].

В центре процесса изъятия информации лежит идея релевантности. Цель автоматического изъятия – получить всю релевантную информацию, оставив как можно меньше нерелевантной. Делается это с помощью присвоения документам индексов, представляющих содержимое этих документов.

Человек может легко определить относится ли документ к запросу. Компьютеру для этого требуется модель, численно определяющая релевантность.

Если в 90-х годах исследования показывали, что люди все еще предпочитали получать информацию от других людей, то уже в 2004 году результаты опроса продемонстрировали, что 92% пользователей интернета представляют его как удобное место для получения ежедневной информации.

Сфера информационного поиска эволюционировала в ответ на различные вызовы для предоставления доступа к информации, развивая все новые подходы для поиска различных видов контента. Первоначально информационный поиск использовался только в сферах научных публикаций и библиотечной документации. Однако вскоре стал необходим для профессионалов – журналистов, юристов и докторов.

Многие исследования информационного поиска были сделаны в этих сферах. Большинство текущей практики работы с информационным поиском включает предоставление доступа к неструктурированной информации в различных корпоративных и государственных областях.

В последнее десятилетие, главным толчком в инновации стала всемирная паутина, давшая доступ для публикации десятком миллионов создателей контента. Такой взрыв информации был бы странным, если информацию невозможно было найти, охарактеризовать и проанализировать, чтобы каждый пользователь мог быстро найти информацию, отвечающую его нуждам.

На данный момент самыми популярными и широко используемыми IR-системами являются поисковые сервисы Google, Яндекс и т.д. Неструктурированные данные – данные, которые не имеют структуры, с которой компьютеру легко работать. Такие данные представляют полную противоположность данным, содержащимся в обычных базах данных, создающихся для работы с данными персонала и инвентаризации.

На самом деле почти не существует полностью неструктурированных данных. Любые текстовые данные имеют структуру, определенную языком, на котором написан текст. Также текст может иметь дополнительную структуру – заголовки, параграфы и сноски [7]. Эти элементы в документах обычно представлены с помощью открытых маркировок. Поэтому информационный поиск часто используется для работы со слабоструктурированными данными.

Системы информационного поиска могут быть разделены по объемам информации, с которыми они работают. В веб-поиске система должна обеспечивать поиск среди миллиардов документов на миллионах компьютеров. Необходимость сбора документов для индексации, создание систем для работы с такими объемами данных, обработка гипертекста и защита от обмана сайтами с содержимым, использующимся для повышения ранга в поисковой системе, являются отличительными проблемами веб- поиска.

Другим примером является персональный поиск информации. Потребительские операционные системы включают в себя информационный поиск. Программы электронной почты не только дают возможности поиска, но также классифицируют текст сообщения, в простейших случаях отсеивая спам, а также предоставляют ручную или автоматическую классификацию, помещая сообщения в определенные папки.

Проблемы данной области информационного поиска включают в себя работу с широким спектром видов документов на персональном компьютере, создание поисковой системы, являющейся свободной и легко запускаемой, обработку и использование дискового пространства, которое может использовать машина без беспокойства для пользователя.

Между этими крайностями находится корпоративный и институциональный поиск и поиск в определенных сферах. Данный тип поиска используется в коллекциях типа внутренних документов корпораций, собраний патентов или поиска статей в определенной научной сфере. Документы хранятся в централизованных файловых системах, поиск по которым обеспечивается одной или несколькими машинами [8].

Целью информационного поиска является вывод как можно большего числа документов, подходящих под запрос пользователя, и как можно меньшего числа неподходящих .

Запрос представляет выражение на естественном языке, которое пользователь вводит в поле запроса.

Одним из эффективных методов информационного поиска является моделирование тем (topic modeling). Модель присваивает каждому документу одну или более тем, которые выявляются из текста в процессе обучения модели. Как правило, используется обучение без учителя. Запросу пользователя также присваиваются темы из выявленных в процессе обучения, затем выводятся документы в порядке, определяемом количеством совпадающих с запросом тем [9].

Поисковые системы типа Google и Яндекс используют полнотекстовый поиск (full-text search). Данный метод оценивает все слова в документе на совпадение с поисковым запросом. Для ускорения полнотекстового поиска используется процесс индексации: система сканирует все документы в доступной ей коллекции и составляет список поисковых терминов – индексов. На стадии поиска система ищет совпадения слов в запросе с индексами, присвоенными документам [10]. Одним из популярных движков полнотекстового поиска является Elasticsearch.

Таким образом, значимость многоязычных и многосторонних подходов к информационному поиску трудно переоценить. Информационный поиск начал представлять не только академический интерес, но и стал базисом для удовлетворения информационных потребностей большинства людей.

Проблема информационного поиска может быть преуменьшена неосведомленным пользователем, который, впрочем, меняет точку зрения при работе с цифровым поиском.

Список литературы

1. Николаев А.А. Разнообразие структур данных в современной информации // Молодой ученый. 2019. №23 (261). С. 21-23.

2. Укуев Б.Т. Особенности обработки неструктурированных данных в информационной базе научных исследований вуза // Естественные и технические науки. 2018. № 3. С. 75-76.

3. Смирнов Ю.В. Информационный поиск для облачных библиотечных систем: особенности лингвистического обеспечения: дисс. к.т.н. – Москва, 2019. – 228 с.

4. ГОСТ Р 7.0.91_2015. СИБИД. Тезаурусы для информационного поиска. – Введ. 2016-07-01. – Москва: Стандартинформ, 2016. – С. 4

5. Магомедов Р.М. О развитии интеллектуальных систем // Территория науки. 2015. №6. С 39-44

6. Батура Т.В. Семантический анализ и способы представления смысла текста в компьютерной лингвистике // Программные продукты и системы. 2016. №4. С. 45-57

7. Цитульский А.М., Иванников А.В., Рогов И.С NLP – Обработка естественных языков // Научно-образовательный журнал для студентов и преподавателей «StudNet». 2020. №6. С.467-475.

8. Панфилова О.А., Крюкова Д.Ю., Давыдова Е.Н. Информационные ресурсы. Системы поиска. Вологда. 2019. С. 81-88.

9. Леонов Е.А., Синицин И.В., Шептунов С.А. Применение методов тематического моделирования для анализа успеваемости студентов в рамках мониторинга образовательного процесса // Качество. Инновации. Образование. Москва. 2018. С. 15-19

10. Миронычев Д.А., Тихонов В.Д. Обзор существующих решений для организации поиска в корпоративных системах информации // Передовые инновационные разработки. Перспективы и опыт использования, проблемы внедрения в производство. Москва. 2019. С. 39-41.

УДК 624.139.26

ОПРЕДЕЛЕНИЕ ОПТИМАЛЬНОЙ ДЛИНЫ БУРОНАБИВНОЙ СВАИ С УШИРЕНИЕМ

Серватинский В.В.
канд. техн. наук, доц.
Преснов О.М.
канд. техн. наук, доц.
Холодов С.П.
канд. техн. наук, доц.
Холодов В.С.
студент ИСИ СФУ
Сибирский федеральный университет,
660041, Россия, Красноярск, проспект Свободный, 79.

Состояние вопроса.

В работе [1] показано, что при работе буронабивных свай с уширением на вертикальную нагрузку, удельная несущая способность сваи (несущая способность отнесенная к объему сваи, кН/м^3) зависит от размеров уширения. В ней приведена методика определения оптимальных его размеров.

В работе также показано, что на удельную несущую способность сваи влияют и другие