TECHNOLOGICAL AND INSTITUTIONAL BARRIERS TO INFORMATION RETRIEVAL

Sharipova Guzal Usmanbekovna

Doctor of Philosophy (PhD) in Historical Sciences, Department of History and Islamic Source Studies – IRCICA, International Islamic Academy of Uzbekistan

g.sharipova@iiau.uz

Abstract

This article systematizes the key technological, linguistic, and institutional barriers that reduce the completeness and accuracy of data retrieval across global computer networks. It examines the fundamental reasons for the limited search coverage of the World Wide Web, whose actual volume accessible to an individual system does not exceed 30%, and describes in detail the phenomenon of "Deep Web" resources being inaccessible to automatic indexing. The article explores the specific linguistic and communicative difficulties associated with homonymy, synonymy, and the prevalence of short, nonspecific user queries that do not employ information retrieval language operators. Institutional and algorithmic distortions in search results caused by search engine optimization techniques, commercial spam, personalization algorithms, and regulatory manipulation are analyzed. Particular attention is given to the systematic rewriting of timestamps and the removal of outdated pages during index database updates, which leads to the erasure of the historical traces of web documents.

Keywords

Information retrieval, search engines, Deep Web, search query, ranking operators, search spam, erasure of historical traces.

ТЕХНОЛОГИЧЕСКИЕ И ИНСТИТУЦИОНАЛЬНЫЕ БАРЬЕРЫ ИНФОРМАЦИОННОГО ПОИСКА

Шарипова Гузал Усманбековна - доктор философии (PhD) по историческим

наукам кафедры «История и источниковедение ислама» – «IRCICA»,
Международной исламской академии Узбекистана

Аннотация. В статье систематизированы ключевые технологические, лингвистические и институциональные барьеры, снижающие полноту и точность извлечения данных в глобальных вычислительных сетях. Рассмотрены фундаментальные причины ограниченности поискового охвата Всемирной Паутины, реальный объем которого для отдельной системы не превышает 30%, и подробно описан феномен недоступности ресурсов “Deep Web” (далее “глубокого веба”, прим автора) для автоматической индексации. Раскрыта специфика лингвистических и коммуникативных трудностей, проявляющихся в омонимии, синонимии и преобладании коротких неспецифичных пользовательских предписаний без использования операторов информационно-поисковых языков. Проанализированы институционально-алгоритмические искажения поисковой выдачи под воздействием методов поисковой оптимизации, коммерческого спама, алгоритмов персонализации и регуляторных манипуляций. Отдельно освещена проблема систематической перезаписи временных меток и удаления устаревших страниц в процессе обновления индексных баз, приводящих к стиранию исторического следа сетевых документов.

Ключевые слова. Информационный поиск, поисковые системы, глубокий веб, поисковый запрос, операторы ранжирования, поисковый спам, стирание исторического следа.

Информационный поиск в цифровых сетевых средах выступает базовым механизмом удовлетворения информационных потребностей пользователей, однако реализация этого процесса сталкивается с рядом существенных системных ограничений. Поисковая система представляет собой информационно-поисковое средство, функционирующее в качестве посредника между пользователем и

хранилищами данных. Эффективность извлечения сведений снижается под воздействием комплекса технологических, лингвистических, структурных и институциональных факторов. Полнота охвата ресурсов Всемирной Паутины даже наиболее мощными отдельными поисковыми машинами не превышает 30%, вследствие чего результаты поиска практически всегда оказываются одновременно избыточными и неполными. Цель данной работы заключается в систематизации и научном анализе основных технологических и институциональных барьеров, препятствующих полноте, точности и исторической достоверности информационного поиска на основе проанализированного источника. Задачи исследования включают раскрытие факторов недоступности ресурсов “глубокого веба”, анализ лингвистических и пользовательских трудностей обработки запросов, а также определение институционально-алгоритмических искажений выдачи и проблемы стирания исторического следа сетевых документов.

Фундаментальный технологический барьер информационного поиска заключается в существовании “невидимого” или “глубокого” веба, объемы которого многократно превосходят индексируемую часть глобальной сети. Данный сегмент включает в себя базы данных, библиотечные каталоги, цифровые книги, научные журналы, патенты, отчеты и государственные архивы. Указанные ресурсы остаются недоступными для автоматического сканирования и индексации стандартными программами-пауками и краулерами.

Невозможность охвата ресурсов “глубокого веба” определяется наличием авторизованного доступа к системам, закрытием отдельных разделов сайта в файлах robots.txt, а также применением экзотических и нестандартных форматов файлов, таких как djvu, doc, swf, архивов и облачных хранилищ. Дополнительным препятствием выступает функционирование веб-ресурсов исключительно в

режиме просмотра без публикации прямых гиперссылок на документы. Полноту поиска ограничивает также внутренняя изменчивость сетевой инфраструктуры, при которой веб-страницы и целые сайты со временем меняют адреса или полностью исчезают с серверов. Кроме того, доступ к национальным и зарубежным базам данных регулярно ограничивается в результате сетевых блокировок по IP-адресам.

Качество информационного поиска снижается вследствие возникновения информационного шума и проявления многовариантности естественного языка. Явления омонимии, синонимии и морфологические особенности слов приводят к выдаче релевантных по форме, но не связанных по смыслу материалов, либо к пропуску целевых документов. Указанные лингвистические трудности усугубляются отсутствием у большинства пользователей навыков формулирования запросов на специализированных информационно-поисковых языках.

Пользовательские предписания в большинстве случаев представляют собой слишком короткие и неспецифичные наборы ключевых слов. Средняя длина запроса составляет около трех терминов, при этом запросы на естественном языке используются крайне редко. Применение специализированных поисковых операторов как, Булевой логики AND, OR, NOT, операторов близости NEAR, FOLLOWED BY, ADJ, W/n, PRE/n, фразового поиска в кавычках, а также ограничений по полям intitle:, inurl:, intext:, filetype: и mime:, составляет лишь около 10% от общего объема веб-запросов. К техническим проблемам лингвистического характера относятся также несогласованность кодировок символов и трудноизвлекаемость информации из мультимедийных файлов.

На формирование результатов поиска оказывают прямое влияние действия сторонних субъектов и алгоритмические механизмы коммерческих поисковых

машин. Владельцы веб-ресурсов активно используют методы поисковой оптимизации и поискового спама, манипулируя метатегами, заглавиями, ключевыми словами и ссылочной массой для искусственного завышения позиций своих сайтов в выдаче.

Алгоритмы коммерческого ранжирования применяют непрозрачные критерии отбора документов. Результаты поиска по идентичным текстовым предписаниям варьируются под воздействием параметров персонализации, зависящих от IP-адреса пользователя, истории его предыдущих обращений и текущего режима авторизации в системе. Поисковые сервисы также внедряют автоподстановки и специальные элементы выдачи, которые отражают коммерческие интересы, геопространственные диспропорции или становятся объектом алгоритмически и вручную регулируемых манипуляций.

В специализированных академических поисковых системах, таких как Google Scholar, также присутствуют институционально-технические ограничения. Система не проектировалась как профессиональная библиометрическая база данных и лишена контроля авторитетности заглавий и авторов. В процессе работы выявляются ошибки автоматического парсинга, отсутствует открытый легальный API для автоматической выгрузки больших массивов метаданных, действует ограничение отображения выдачи первыми 1000 результатов, а также сохраняется уязвимость к проявлениям контентного и цитатного спама.

Значимый институционально-технологический барьер возникает при попытке использования коммерческих поисковых систем в исторических исследованиях. Операционная стабильность поисковых машин достигается за счет регулярного обновления их индексных баз данных. В ходе этого процесса происходят систематический перенос временных меток веб-документов из

прошлого в настоящее или недавнее прошлое на основе даты их последнего обновления, а также бесследное удаление устаревших страниц.

Постоянная актуализация индексов приводит к стиранию исторического следа развития информационных сетей и разрушению смысловых структур семантических сетей. Коммерческие поисковые системы непрерывно конструируют “конкурирующее настоящее”, вследствие чего они оказываются непригодными инструментами для ретроспективного анализа структуры информации в конкретный исторический период. Преодоление данного барьера возможно только путем обращения к специализированным архивным веб-ресурсам, сохраняющим исходную дату публикации, или путем создания специализированных инструментов архивирования.

В заключение следует отметить что, проведенный анализ показывает, что технологические и институциональные барьеры информационного поиска обусловлены как фундаментальными свойствами Всемирной Паутины, так и спецификой функционирования коммерческих поисковых машин. Недоступность “глубокого веба”, изменчивость адресов ресурсов и сетевые блокировки создают структурные препятствия полноте охвата. Лингвистическая многовариантность и отсутствие у пользователей навыков применения языков поисковых предписаний снижают точность и полноту извлечения данных. Институциональные искажения, вызванные поисковым спамом, непрозрачностью алгоритмов ранжирования, персонализацией и систематической перезаписью временных меток документов, приводят к трансформации поисковой выдачи и стиранию исторического следа. Преодоление проанализированных барьеров требует использования специализированных библиографических систем, корректного формулирования запросов с применением поисковых операторов и привлечения профильных цифровых архивов.

СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ

1.Гримуто О. А. Эффективное использование поисковых систем Google и Google академия в учебной и научной деятельности //Физика конденсированного состояния: материалы XXXI междунар. науч.-практ. конф. аспир., магистр. и студ.(Гродно, 13-14 апр. 2023 г.)/Учреждение образования" Гродненский государственный университет имени Янки Купалы". – 2023. – С. 366.

2.Калмыков М. А., Медникова О. В. История развития поисковых систем и алгоритмы их работы //Научные известия. – 2022. – №. 28. – С. 33-35.

3. Шарипова Г. Цифровая история. Учебное пособие / под ред.— Ташкент: GRAND KONDOR PRINT, 2026.

4.Croft W. B. et al. Search engines: Information retrieval in practice. – Reading : Addison-Wesley, 2010. – Т. 520. – С. 131-141.

5.Haider J., Sundin O. Invisible search and online search engines: The ubiquity of search in everyday life. – Taylor & Francis, 2019. – С. 160.

6.Orduna-Malea E., Martín-Martín A., Delgado-Lopez-Cozar E. Google Scholar as a source for scholarly evaluation: A bibliographic review of database errors. – 2017.