ДИАЛОГИ СО СПЕЦИАЛИСТАМИ

Научная статья
УДК 513.2; DOI: 10.61260/2304-0130-2024-4-28-31

ПОИСКОВЫЕ СИСТЕМЫ СЕТИ ИНТЕРНЕТ

✉**Лабинский Александр Юрьевич.**
Санкт-Петербургский университет ГПС МЧС России, Санкт-Петербург, Россия
✉***labinskyi.a@igps.ru***

Аннотация. Рассмотрены особенности поисковых систем, используемых для поиска информации по ключевым словам в различных хранилищах информации, в том числе в сети Интернет. Представлены различные поисковые системы, включая самую популярную поисковую систему в мире Google и популярную в России систему Яндекс. Показаны особенности работы поисковых систем, включая основные компоненты и этапы поиска информации, связанные с компонентами поисковой системы. Подробно рассмотрены основные типы поисковых систем. Для каждого типа поисковой системы приведены примеры практической реализации. В статье основное внимание уделено примеру поисковой системы, использующей нечеткий поиск. Приведены преимущества нечеткого поиска по сравнению с четким поиском информации, в частности, учет таких особенностей текста запроса, как языковые и смысловые взаимосвязи, грамматические формы слов запроса, а также возможные ошибки и опечатки. Для нечеткой поисковой системы, реализованной в виде программы для ЭВМ, приведены блок-схема алгоритма сравнения строки с образцом запроса и блок-схема событий интерфейса программы поиска информации.

Ключевые слова: поисковая система, нечеткая поисковая система, ключевые слова, типы поисковых систем, компоненты поисковых систем, алгоритм сравнения строк, программа для ЭВМ

Для цитирования: Лабинский А.Ю. Поисковые системы сети Интернет // Надзорная деятельность и судебная экспертиза в системе безопасности. 2024. № 4. С. 28–31. DOI: 10.61260/2304-0130-2024-4-28-31.

Введение

Поисковой системой (Search engine) называется совокупность компьютерных программ, предназначенная для поиска документов по ключевым словам в различных хранилищах информации, в том числе в сети Интернет. Результатом поиска могут быть веб-страницы, изображения, аудиофайлы и т.п. Качество поиска оценивается по количеству документов, релевантных (в наибольшей степени отвечающих) запросу пользователя.

В настоящее время самой популярной поисковой системой в мире, в том числе и в России, является система Google, алгоритм которой был разработан в начале 2000-х гг. Сергеем Брином и Ларри Пейджем, основателями компании Google. Приблизительное число серверов Google в середине 2000-х гг. составляло 1 млн, суммарная мощность центров обработки информации Google оценивалась в 250 МВт, а расходы на эти центры обработки информации составляли около 2,5 млрд долл. в год.

Сформулируем постановку задачи. Нужно рассмотреть особенности систем поиска информации, включая особенности работы поисковых машин, используемых в интернете. Тема статьи актуальна, так как объем информации, размещаемой в интернете, постоянно увеличивается. Поэтому поисковым системам посвящено значительное число работ [1–11].

Особенности работы поисковых машин

Основными компонентами поисковой машины являются поисковый робот, индексатор и поисковик. Каждый компонент поисковой машины реализует определенный этап поиска информации.

На первом этапе поисковый робот загружает данные, соответствующие запросу пользователя.

Поисковый индекс позволяет в максимальной степени ускорить поиск информации. В результате создается индексная база данных.

В поисковой машине Яндекс используется поисковый механизм (программа) «Метапоиск».

Некоторые поисковые системы позволяют производить нечеткий (приближенный) поиск, учитывающий расстояние до ключевых слов.

Пример поисковой системы, использующей нечеткий поиск [10]

Нечеткий поиск информации в процессе поиска учитывает расстояние редактирования, определяющее количество операций редактирования, которые нужно применить к строке поискового запроса, чтобы учесть все возможные искажения, ошибки и опечатки.

Расстояние редактирования характеризует степень «похожести» строк.

В 1965 г. советский математик В.И. Левенштейн [6] решил задачу похожести строк и предложил широко используемое сейчас расстояние редактирования, для вычисления которого используется следующий алгоритм: первый символ текста сравнивается с образцом запроса, если символы совпадают, выполняется переход к следующему символу и т.д.

Процесс сравнения останавливается, если все символы совпадают (поиск останавливается) или встречаются несовпадающие символы.

Блок-схема алгоритма сравнения строки с образцом запроса представлена на рис. 1.

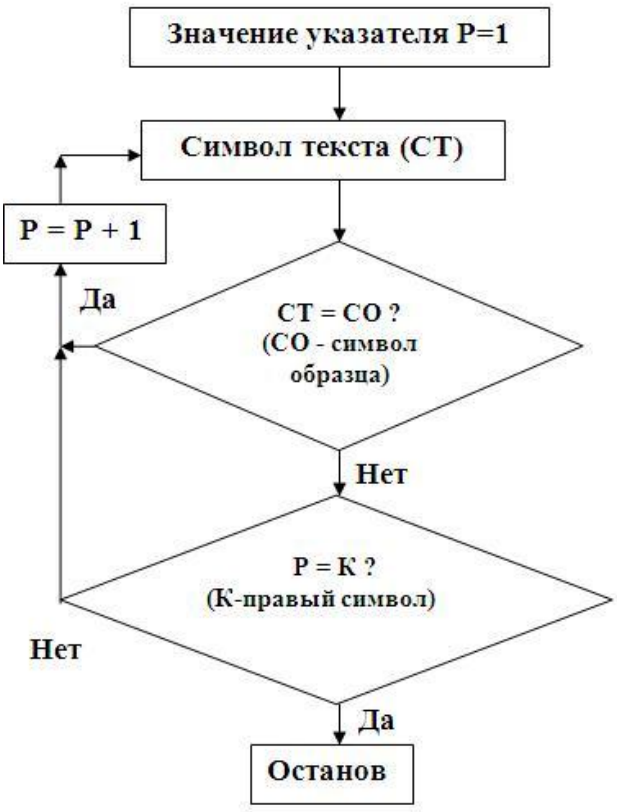

Рис. 1. **Блок-схема алгоритма сравнения строки с образцом запроса**

Компьютерная модель нечеткого поиска

В качестве примера системы нечеткого поиска представлена компьютерная модель, заимствованная в работе [11] и реализованная в виде программы для ЭВМ (рис. 2). Программа позволяет производить четкий и нечеткий поиск подстроки теста в файлах интернета (типы файлов *.htm, *.xml, *.txt). Поиск подстроки текста осуществляется в файлах, расположенных в указанном каталоге.

В строку ввода искомого текста «Найти подстроку:» вводится искомая подстрока и путем нажатия кнопки «Поиск» запускается процесс поиска файлов, содержащих указанную подстроку. Имена найденных файлов помещаются в список «Файл для просмотра».

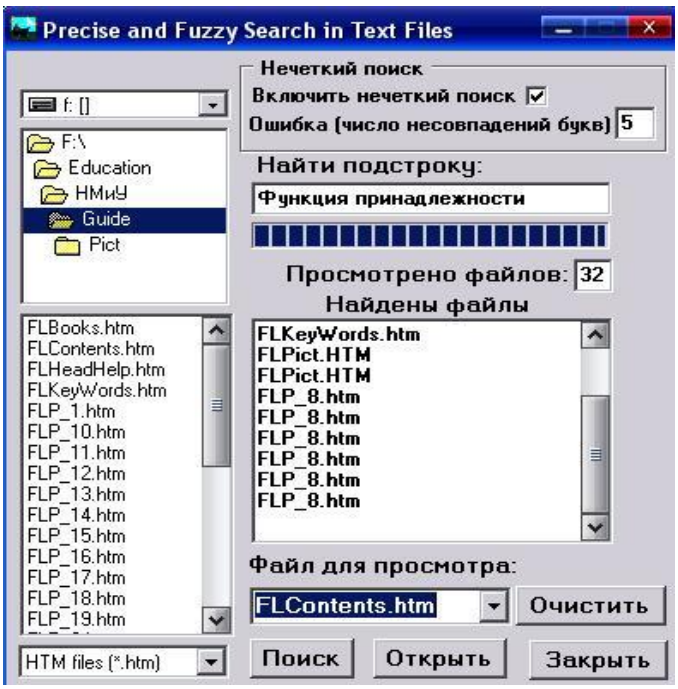

Рис. 2. **Программа четкого и нечеткого поиска информации**

Блок-схема событий интерфейса программы поиска информации представлена на рис. 3.

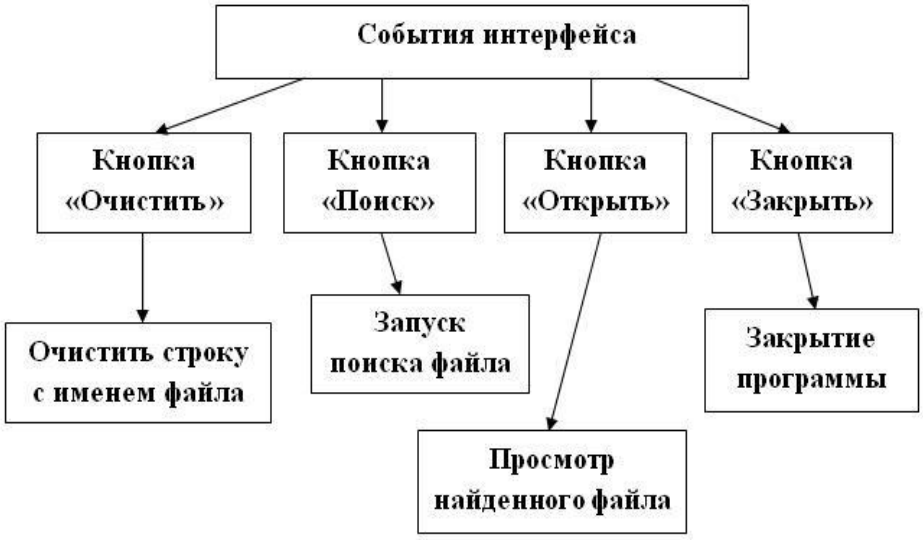

Рис. 3. **Блок-схема событий интерфейса программы поиска информации**

Вывод

Рассмотрены особенности поисковых систем, используемых для поиска информации по ключевым словам в сети Интернет. Нечеткие поисковые системы имеют преимущества перед традиционными четкими системами, так как позволяют производить нечеткий (приближенный) поиск, учитывающий расстояние до ключевых слов.

Список источников

1. Ашманов И.С., Иванов А.А. Продвижение сайтов в поисковых системах. М.: Вильямс, 2007.

2. Кириллов А.В. Поисковые системы: компоненты, логика, методы ранжирования // Бизнес-информатика. 2009. № 4.

3. Колисниченко Д.Н. Поисковые системы и продвижение сайтов в Интернете. М.: Диалектика, 2007.

4. Ландэ Д.В. Поиск знаний в Internet. М.: Диалектика, 2005.

5. Ландэ Д.В., Снарский А.А., Безсуднов И.В. Интернетика: навигация в сложных сетях: модели и алгоритмы. М.: Либроком, 2009.

6. Левенштейн В.И. Двоичные коды с исправлением выпадений и вставок символа // Проблемы передачи информации. 1965. № 1.

7. Chu H., Rosenthal M. Search engines for the World Wide Web // American Society for Information. 2006. Vol. 33.

8. Risvik K.M., Michelsen R. Search engines and web dynamics // Computer Networks. 2002. Vol. 39.

9. Tarakeswar M.K., Kavitha M.D. Search Engines // Journal of Computer Applications. 2011. Vol. 4.

10. Сонькин М.А., Лещик Ю.В. Применение алгоритмов нечеткого поиска в системах мониторинга лесопожарной обстановки // Труды ТГТУ. 2012. № 2.

11. Лабинский А.Ю. Нечеткий поиск текстовой информации // Природные и техногенные риски (физико-математические и прикладные аспекты). 2019. № 2. С. 42–45.

Информация о статье: статья поступила в редакцию: 26.10.2024; принята к публикации: 19.11.2024

Информация об авторах:

Лабинский Александр Юрьевич, доцент кафедры прикладной математики и информационных технологий Санкт-Петербургского университета ГПС МЧС России (196105, Санкт-Петербург, Московский пр., д. 149), кандидат технических наук, доцент, e-mail: labinskyi.a@igps.ru, https://orcid.org/0000-0001-2735-4189, SPIN-код: 8338-4230