УДК 004

РАЗРАБОТКА ИНФОРМАЦИОННОЙ СИСТЕМЫ БЫСТРОГО ПОЛНОТЕКСТОВОГО ПОИСКА В ФАЙЛАХ

© 2022 А. В. Авдеев, А. Н. Зеленина

Воронежский институт высоких технологий (Воронеж, Россия)

Актуальность темы исследования обусловлена несколькими факторами: быстрым развитием систем электронного документооборота (ЭДО), и, как следствие, часто возникающей на предприятиях необходимостью поиска информации по ключевым словам в большом количестве документов; возможностью наличия в сторонних поисковых системах недекларированных возможностей (в т. ч. «бекдоров»), либо вредоносного ПО (вирусов, «троянов» и т. д.), вследствие чего снижается безопасность использования данных систем для поиска информации в документах, содержащих коммерческую или иную тайну, либо другую конфиденциальную информацию. Полученные результаты – простая в использовании полностью работоспособная программная система информационного поиска, осуществляющая полнотекстовый поиск в файлах быстрее распространённых аналогов. Основные преимущества полученных результатов – простота использования, интуитивная понятность интерфейса, высокая скорость работы. Планируемая область применения результатов исследования: любая организация или частное предприятие, личное использование.

Ключевые слова: полнотекстовый поиск, параллельные вычисления, многопоточность, проектирование ИС.

Введение

Объектом исследования является файловая система операционной системы Windows. Предметом исследования являются подходы к быстрому поиску документов, хранящихся в файловой системе локального компьютера или сетевого носителя, в том числе с использованием регулярных выражений.

Целью работы является разработка прототипа информационной системы быстрого полнотекстового поиска в файлах. Эта цель определяет следующие частные задачи:

1) обоснование актуальности темы и разрабатываемой системы;

2) анализ принципов построения поисковых систем;

3) исследование аналогов разрабатываемой системы;

4) постановка задачи, формулирование необходимого функционала;

5) моделирование вариантов использования;

6) проектирование структуры системы;

7) разработка модулей и написание кода программы;

8) разработка руководства по использованию;

9) тестирование разработанной системы и сравнение её с аналогами.

Информационный поиск – это процесс отыскания нужных пользователю сведений в некотором множестве документов, хранящемся на локальном или сетевом носителе (рис. 1). Главной задачей любой поисковой системы является поиск информации в соответствии с информационными потребностями пользователя, формируемыми в виде запроса. Очень важно в результате проведенного поиска ничего не потерять, то есть найти в индексе все документы, относящиеся к запросу (полнота поиска), и не выдать ничего лишнего (точность поиска). На рисунке 2 представлена методология функционирования комплексной информационно-поисковой системы (ИПС).

В настоящее время существует множество поисковых систем, отличающихся друг от друга архитектурой, скоростью поиска, функционалом, удобством использования и иными характеристиками (рис. 3).

Авдеев Артем Владимирович – Воронежский институт высоких технологий, студент.

Зеленина Анна Николаевна – Воронежский институт высоких технологий, канд. техн. наук, доцент, e-mail: snakeans@gmail.com.

- документальный поиск;
- фактографический поиск;
- полнотекстовый поиск;
- поиск по изображению;
- семантический поиск;
- поиск по метаданным;

Рисунок 1. Информационный поиск

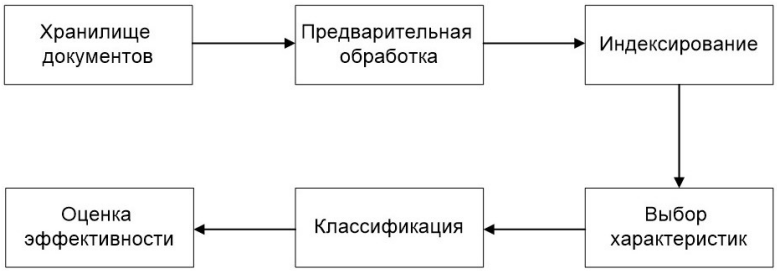

Рисунок 2. Методология функционирования комплексной информационно-поисковой системы

Рисунок 3. Классификация поисковых систем

Среди таких систем можно выделить как системы интернет-поиска (Google, Yandex и т. д.), так и отдельные поисковые программы или штатные средства приложений и операционных систем.

Как уже упоминалось выше, в случае циркулирования в информационных потоках предприятия конфиденциальной информации, использование сторонних поисковых систем может быть недостаточно безопасным. Таким образом, видится актуальной задача разработки собственной системы полнотекстового поиска в файлах, находящихся на локальном или сетевом хранилище. Решение

данной задачи позволяет получить универсальное, безопасное, открытое, быстрое и простое средство поиска в наиболее популярных на сегодня форматах текстовых документов (DOCX, XLSX, PPTX, TXT). В нашем случае поиск будет производиться по ключевым словам или регулярным выражениям.

Постановка задачи, формулирование необходимого функционала

Программная система обеспечивает выполнение следующих функций:

- выбор директории для поиска;
- ввод поискового запроса по ключевым словам или регулярным выражениям;
- вывод результатов поиска в виде таблицы с перечнем найденных файлов;
- просмотр найденных документов с помощью штатных средств операционной системы.

Также разрабатываемая ИС имеет графический интерфейс и высокую скорость работы относительно прямых аналогов. Для уменьшения времени выполнения поискового запроса использован многопоточный алгоритм с пропорциональным распределением задач поиска на все доступные ядра процессора.

При разработке использован язык программирования C++ 14 совместно с фреймворком Qt версии 5.15.2 и средой Qt Creator 5.0.0. Сборка проекта осуществляется с помощью утилиты qmake с использованием компилятора из состава Visual Studio 2019.

Разработанная ИС состоит из шести основных классов, полная схема отношений между классами представлена на рисунке 4.

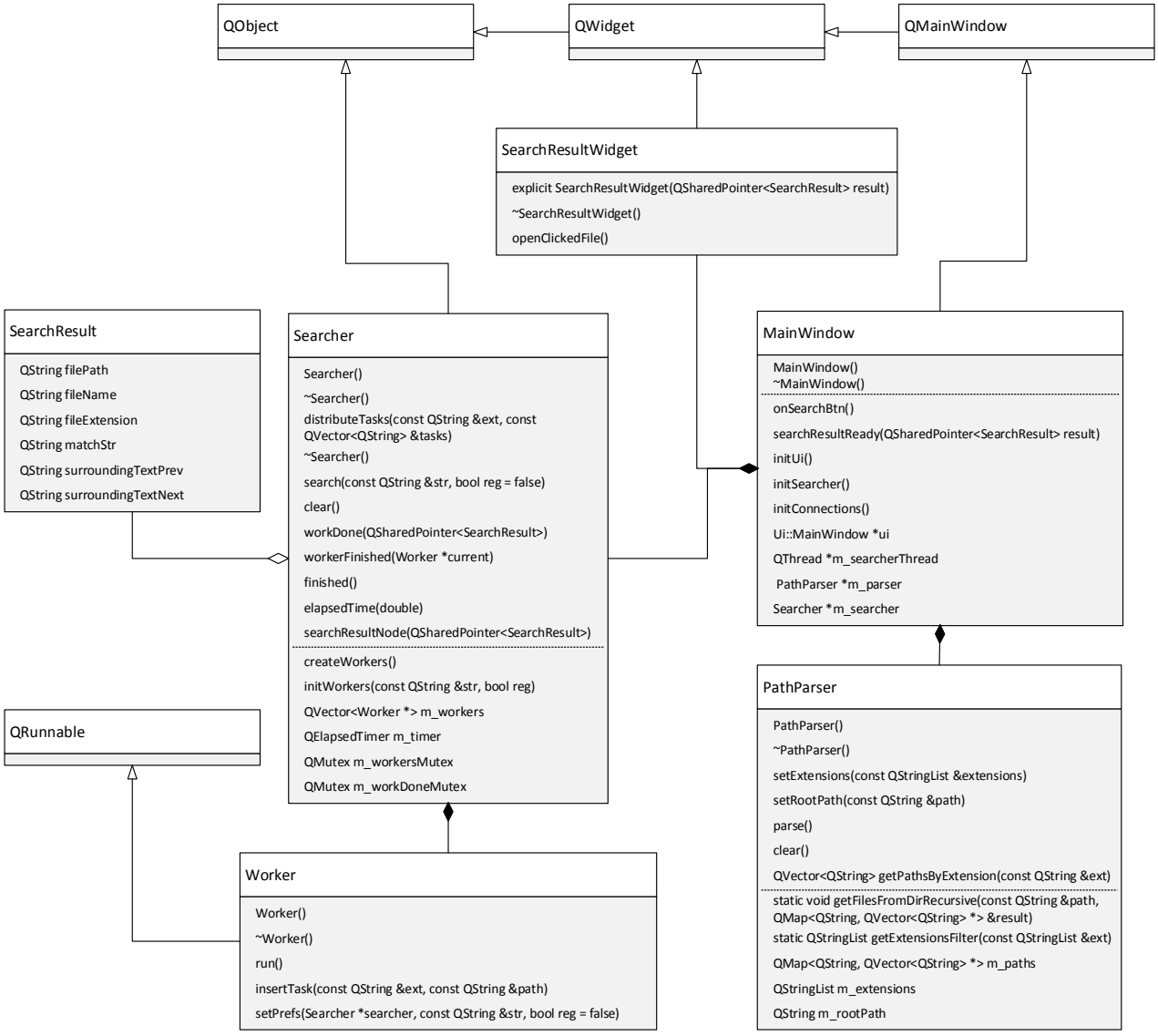

Рисунок 4. UML диаграммы основных классов ИС «Searcher»

Интерфейс разработанной ИС интуитивен и понятен, что позволяет легко использовать её на любом предприятии любым сотрудником, обладающим базовыми навыками работы с ОС Windows. Для начала работы необходимо задать директорию для поиска, выбрать режим поиска (строка или регулярное выражение) и задать соответствующий поисковый запрос (рис. 5). После осуществления поиска программа выдаёт найденную информацию в виде таблицы.

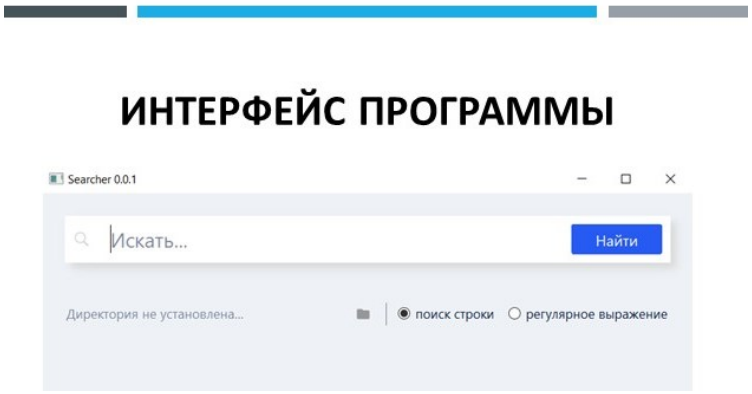

Рисунок 5. Интерфейс ИС «Searcher». Поисковой запрос

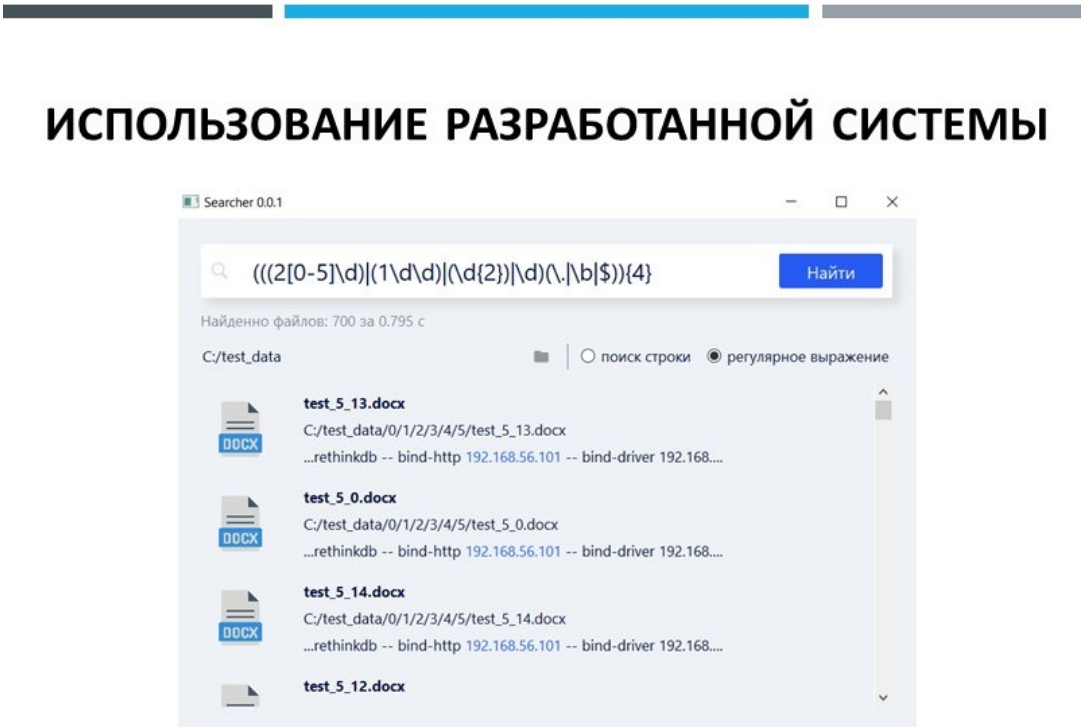

Рисунок 6. Интерфейс ИС «Searcher». Результат поиска

По сравнению со штатным средством поиска ОС Windows и коммерческим приложением Total Commander, разработанная ИС «Searcher» имеет полную поддержку поиска по регулярным выражениям, а также гораздо более высокую скорость работы (рис. 7). Согласно результатам тестирования (рис. 8), при поиске файлов по регулярному выражению, разработанная ИС осуществляет поиск за 2 секунды, ближайшие аналоги Проводник и Total Commander – за 26 и 16 секунд соответственно.

Рисунок 7. Сравнение возможностей ИС «Searcher» с аналогами

Рисунок 8. Сравнение программ по времени поиска

Заключение

Информационная система быстрого полнотекстового поиска в файлах протестирована в ООО «Центр современной педиатрии». Решение поставленной задачи позволило получить универсальное, быстрое и безопасное средство поиска, что упрощает работу и позволяет снизить вероятность случайного или злонамеренного разглашения конфиденциальной информации или заражения устройства вредоносным ПО. Сравнительный анализ с прямыми аналогами (рис. 8) показал значительное увеличение скорости работы. В дальнейшей перспективе функционал данного продукта (рис. 9) может быть расширен – при необходимости можно добавить больше форматов файлов для поиска, возможность параллельного поиска в нескольких, не вложенных друг в друга директориях, и другие функции.

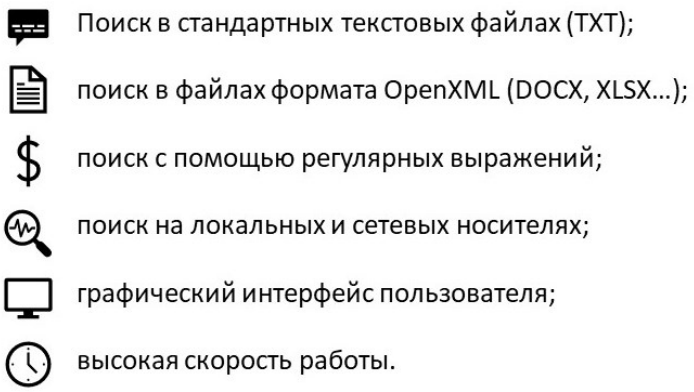

Рисунок 9. Основные положения и варианты использования ИС «Searcher»

СПИСОК ИСТОЧНИКОВ

1. Демихов М. А. Поисковые методы на основе tree-деревьев / М. А. Демихов // Вестник Воронежского института высоких технологий. – 2016. – №. 4. – С. 99-102.

2. Львович И. Я., Кравцова Н. Е., Чупринская Ю. Л. Особенности решений для обработки текстовых данных / И. Я. Львович, Н. Е. Кравцова, Ю. Л. Чупринская // Вестник Воронежского института высоких технологий. – 2019. – №. 1. – С. 89-92.

3. Преображенский Ю. П. Анализ методов нечеткого поиска / Ю. П. Преображенский, Д. Н. Мирошник // Вестник Воронежского института высоких технологий. – 2018. – №. 4. – С. 82-84.

4. Решетников А. Д. О подходах для определения меры несходства в текстовых данных А. Д. Решетников // Вестник Воронежского института высоких технологий. – 2019. – №. 3. – С. 35-39.

5. Шапаев А. В. Проблемы поиска текстовой информации в больших объемах данных / А. В. Шапаев, Д. А. Юдаков, А. А. Часовской // Вестник Воронежского института высоких технологий. – 2019. – №. 1. – С. 113-115.

6. Ширяев В. В. Извлечение текстовых данных из документов формата PDF, DOCX (DOC) с помощью сторонних библиотек / В. В. Ширяев, А. В. Турчановская // Труды семинара по геометрии и математическому моделированию. – 2019. – №. 5. – С. 158-160.

7. Шахова О. А. Статистическая обработка результатов исследований: учебное пособие / Шахова О. А. – Тюмень: Издательство «Титул», 2022. – 103 с.

8. Мельникова Т. В. Моделирование обработки больших массивов данных в распределенных информационно-телекоммуникационных системах / Т. В. Мельникова, М. В. Питолин, Ю. П. Преображенский // Моделирование, оптимизация и информационные технологии. – 2022. – Т. 10. – № 1 (36). Доступно по: https://moitvivt.ru/journal/article?id=1117 (дата обращения: 10.09.2022).

DEVELOPMENT OF INFORMATION SYSTEM FOR QUICK FULL TEXT SEARCH IN FILES

© 2022 A. V. Avdeev, A. N. Zelenina

Voronezh Institute of High Technologies (Voronezh, Russia)

The relevance of the research topic is due to several factors: the rapid development of electronic document management systems (EDM), and, as a result, the need to search for information by keywords in a large number of documents that often arises at enterprises; the possibility of presence in third-party search engines of undeclared features (including backdoors) or malware (viruses, Trojans, etc.), which reduces the security of using these systems to search for information in documents containing commercial or other secret or other confidential information). The obtained result of the research is an easy-to-use, fully functional information retrieval software system that performs full-text search in files faster than common analogues. The main advantages of the developed software system are ease of use, intuitive interface, high speed. Planned scope of the research results: Any organization or private enterprise, personal use.

Keywords: full-text search, parallel computing, multithreading, IS design.