МИНОБРНАУКИ РОССИИ
Федеральное государственное бюджетное образовательное учреждение
высшего образования
**«САРАТОВСКИЙ НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ ИМЕНИ Н.Г. ЧЕРНЫШЕВСКОГО»**

Кафедра теоретических основ компьютерной безопасности и криптографии

# Расширенный поиск файлов

АВТОРЕФЕРАТ
дипломной работы

студента 6 курса 631 группы
специальности 10.05.01 Компьютерная безопасность
факультета компьютерных наук и информационных технологий
Смирнова Данилы Антоновича

Научный руководитель
к.ю.н., доцент _____ 23.01.2021 г. А. В. Гортинский

Заведующий кафедрой
д. ф.-м. н., доцент _____ 23.01.2021 г. М. Б. Абросимов

Саратов 2021

# ВВЕДЕНИЕ

В результате работы создана программа поиска файлов некоторых типов файлов.

Схожий функционал предоставляет продукция организации FileHold. Данная организация реализовала программу управления документами, а также поиска документов по специфическим характеристикам.

Так как программа от организации FileHold не является свободным программным обеспечением, то плюсом работы является свобода составленной программы в правовом плане и, следственно, ее бесплатность.

К целям работы можно отнести:

- построение простой в управлении программы, способной к поиску файлов;
- возможность поиска документов по содержимому;
- возможность поиска документов по пользовательским критериям;
- поддержка сложных критериев как для метаданных, так и для атрибутов файловых систем;
- возможность поиска документов по содержащихся в них изображений;
- возможность нахождения документов, содержащих незначительно измененное искомое изображение;
- независимость от кодировки какого-либо поля метаданных или атрибута файловой системы;
- поддержка поиска электронных документов наиболее распространенных форматов (doc, docx, odt, pdf);
- поддержка поиска электронных таблиц (xls, xlsx, ods);
- поддержка поиска изображений (jpg, png).

В частности, были рассмотрены открытые спецификации, требующиеся для поиска указанных выше типов файлов.

Приведены характеристики основных используемых полей метаданных, атрибутов файловых систем. Описана общая структура файлов, с которыми проводилась работа.

Дипломная работа состоит из введения, 5 разделов, заключения, списка использованных источников и 3 приложений. Общий объем работы – 72 страниц, из них 28 страниц – основное содержание, включая 6 рисунков и 2 таблицы, список использованных источников из 14 наименований.

## КРАТКОЕ СОДЕРЖАНИЕ

**1 Открытые форматы файлов**

Дается общее описание извлекаемых метаданных и приводится список типов файлов, с которыми работает программа.

**1.1 Общее описание OLE-хранилищ**

Приводятся сведения о составных документах производства корпорации Microsoft в соответствии с открытой спецификацией*.

**1.2 Спецификация файла SummaryInformation**

Приводится характеристика обозначенного в названии подраздела файла и его содержимого**.

Дается описание полей и структуры данного файла.

Приводится общее описание таких типов структур***.

**1.3 Спецификация файла DocumentSummaryInformation**

Аналогично предыдущему подразделу приводится характеристика обозначенного в названии подраздела файла и дается описание полей метаданных данного файла. Так как по структуре рассматриваемые файлы из этого и предыдущего подразделов одинаковы, внутренняя структура DocumentSummaryInformation не указывается.

---

* Daniel Rentz, OpenOffice.org's documentation of the Microsoft Compound Document file format [Электронный ресурс] : (на 15 октября 2020 года, версия 1.5) / Daniel Rentz // OpenOffice.org - URL: www.openoffice.org/sc/compdocfileformat.pdf (дата обращения 10.10.2020). -Загл. с экрана. - Яз. англ.

** SummaryInformation Property Set [Электронный ресурс]: [сайт]. URL: docs.microsoft.com/en-us/openspecs/windows_protocols/ms-oleps/3f9119dc-faa2-4bb9-af95-5cf128fa5fbd?redirectedfrom=MSDN (дата обращения: 7.12.2020). Загл. с экрана. Яз. англ.

*** Property Set [Электронный ресурс]: [сайт]. URL: https://docs.microsoft.com/en-us/openspecs/windows_protocols/ms-oleps/aefcbddf-f299-4f5e-a9da-65ce4ca55075 (дата обращения: 13.10.2020). Загл. с экрана. Яз. англ.

### 1.4 Структура JPEG

Приводится описание JPEG-файлов[*], удовлетворяющих спецификации EXIF[***].

Дается список полей метаданных, которые используются в работе.

Приводится общее описание секций и маркеров[**], необходимое для работы со структурой таких файлов.

### 1.5 Структура ZIP

Описывается внутренняя структура ZIP-контейнеров[****], их заголовки и служебные поля, необходимые для извлечения файлов из упомянутых контейнеров.

Приводятся некоторые особенности таких контейнеров, а также некоторые сходства с OLE-хранилищами.

### 1.6 Спецификация Office Open XML

Приводится характеристика документов, соответствующих структуре, указанной в названии подраздела.

Описываются необходимые файлы и поля метаданных, содержащие необходимую при поиске информацию.

---

[*] Exploring JPEG [Электронный ресурс]: [сайт]. URL: https://www.imperialviolet.org/binary/jpeg/ (дата обращения: 20.11.2020). Загл. с экрана. Яз. англ.

[**] Description of Exif file format [Электронный ресурс]: [сайт]. URL: https://www.media.mit.edu/pia/Research/deepview/exif.html (дата обращения: 6.10.2020). Загл. с экрана. Яз. англ.

[***] Exchangeable image file format for digital still cameras: Exif Version 2.2 [Электронный ресурс] : (на 10 октября 2020 года, версия 2.2) // Standard of Japan Electronics and Information Technology Industries Association. - URL: www.exif.org/Exif2-2.PDF (дата обращения 10.10.2020). -Загл. с экрана. - Яз. англ.

[****] ZIP File Format Specification [Электронный ресурс]: [сайт]. URL: pkware.cachefly.net/webdocs/casestudies/APPNOTE.TXT (дата обращения: 7.11.2020). Загл. с экрана. Яз. англ.

## 2 Описание данных

В данном разделе приводится описание типов данных и групп метаданных, с которыми работает программа.

Рассматриваются основные типы данных, используемых в работе. Приводится список групп метаданных. Описываются способы определения типов полей метаданных и типов файлов.

### 2.1 Тип FileTime

Приводится описание специфичного для систем Windows типа данных, указанного в названии подраздела.

### 2.2 Группы метаданных

В данном разделе описываются группы в соответствии со стандартом MIME.

Кроме того, приводятся примеры тех групп метаданных, с которыми работает основная программа.

### 2.3 Определение типов полей

В данном подразделе описывается то, как программа поддерживает соответствие каждого названия поля метаданных и его типа.

Это необходимо, так как позволяет сильно упростить извлечение метаданных из файлов.

### 2.4 Определение типов файлов

Данный подраздел содержит методы, при помощи которых возможно с достаточной вероятностью узнать тип файла.

Приводится описание сигнатурного метода.

## 3 Реализация анализатора запросов

Данный модуль содержит функции, которые готовят требования, вводимые пользователем, к преобразованию в логическое выражение в виде обратной польской нотации, при этом разбивая на отдельные наборы строк.

### 3.1 Функция определения знака сравнения

В этом подразделе описывается функция, определяющая, является ли символ знаком сравнения.

### 3.2 Функция сравнения логических операторов

В этом подразделе описывается функция сравнения логических операторов по порядку вычисления результата логического выражения.

### 3.3 Функция определения логического оператора

В данном подразделе содержится описание функции, проверяющей символ, который может быть логическим оператором.

### 3.4 Функция разбиения логического выражения

В этом подразделе дается характеристика функции, разбивающей набор символов на логические блоки, необходимые для поиска файлов.

### 3.5 Функция разбиения сравнения

В данном подразделе описываются этапы выделения из текстового блока частей, которые относятся к сравнению. В случае успешного завершения функции будет получено ровно 3 текстовых блока, соответствующих правильно сформулированному сравнению.

## 4 Поисковые алгоритмы

В данном разделе приведены необходимые определения и функции, используемые при поиске файлов по пользовательским критериям.

Описываются функции, применяемые для:

- проверки введенного пользователем запроса на корректность;
- преобразования в более удобную в работе нотацию;
- проверки рассматриваемых файлов в соответствии с требованиями, введенными пользователем.

Описываются действия, совершаемые программой при некорректном вводе данных.

### 4.1 Обратная польская нотация

В данном подразделе приводится описание структуры, к которой приводятся все логические выражения в данной работе.

Все логические выражения приводятся к описываемой структуре, так как она проста в реализации.

Кроме того, процесс приведения к данной нотации не требует больших временных затрат.

### 4.2 Функция преобразования выражения в обратную польскую нотацию

В этом подразделе описывается алгоритм приведения к вышеупомянутой нотации пользовательского текстового выражения.

Наглядно показывается то, как выглядит запрос пользователя после преобразования в рассматриваемую нотацию.

### 4.3 Функция проверки выражения

Данный подраздел содержит описание функции, показывающей, подходит ли рассматриваемый программой файл под пользовательские критерии.

### 4.4 Функция перебора сравнений

В этом подразделе описывается функция, которая перебирает все текстовые блоки и для каждого проверяет корректность.

Если же встречен символ оператора, то проверка не проводится.

### 4.5 Проверка выражения на корректность

В подразделе описывается функция проверки одного выражения на корректность. Это необходимо, так как в случае опечатки результат поиска файлов может оказаться не верным.

### 4.6 Функция проверки файлов

В данном подразделе описываются действия, выполняемые при переборе программой файлов в целевой файловой системе.

В результате программа составляет вывод о необходимости включения рассматриваемого файла в список результатов поиска.

### 4.7 Сравнение изображений

В этом подразделе описываются действия, необходимые при сравнении изображений при поиске документов по изображению, которое может содержаться в каком-либо из рассматриваемых документов.

Рассматривается простейший количественный способ приблизительного сравнения изображений, удовлетворяющих модели RGB*.

---

* Википедия [Электронный ресурс] : свободная энциклопедия / текстдоступен по лицензии Creative Commons Attribution-ShareAlike ; WikimediaFoundation, Inc, некоммерческой организации. - Электрон. дан. - Wikipedia®,2021- URL: ru.wikipedia.org/wiki/ (дата обращения: 7.10.2020). - Загл. с экрана. -Яз. рус.

## 5 Дополнительные функции

В данном разделе приведены вспомогательные функции, используемые программой.

### 5.1 Функция извлечения атрибутов файловой системы

В подразделе приводится описание функции, извлекающей атрибуты файла, которые ему приписаны файловой системой.

Кроме того, приводится список возможных файловых атрибутов.

### 5.2 Функция автоматического извлечения метаданных

В этом подразделе описываются действия, осуществляемые при рассмотрении файлов со сложными структурами, которые не были рассмотрены в данной работе полностью.

### 5.3 Утилитарные функции

В этом подразделе описываются вспомогательные функции, используемые в процессе функционирования основной программы.

# ЗАКЛЮЧЕНИЕ

Была реализована программа поиска файлов по их метаданным и другим характеристикам.

Задачи реализованы не в полной мере: для некоторых типов файлов со сложными внутренними структурами используются сторонние подключаемые средства.

Эффективность основной программы может быть оценена как недостаточная, так как в случае со сложными структурами время работы увеличивается. Это также связано со средствами, использовавшимися при составлении программы.

Составленная программа может использоваться как альтернатива существующим программам расширенного поиска файлов.