Journal of Innovative Technology Convergence
Vol. 1, No. 2 December 2019, pp. 25-32

File Comparison Prototype: A Statistical Tool for Comparing Two Text Files

Rizalyn H. Declines [1], Nikki Jane C. Ducado [2], Kenneth E. Hiponia [3], April Joy C. Orquia [4], Gwyn A. Paduganao [5], Kathy Mae M. Toledo [6], Kenrick Agustin S. Secugal [7]*

Abstract: File comparison methods have been widely used by popular applications for file searching and checking plagiarized contents. This study aims to match two text files and get their average rate of accuracy. A file comparison prototype has been developed to perform the match, and its accuracy was determined with regards to the number of files and their sizes. The prototype's acceptability will be determined with regards to usability, performance, and design. The arithmetic mean will be used to determine the accuracy of the matching text file, and the prototype will display the matched and unmatched results. The results show that the acceptability level for matching two text files is accurate, and the design of the GUI is easy to learn and use.

Keywords: File Comparison, Prototype, Matching text file, File matching

1. Introduction

Comparing text files can be found today in many applications, such as File Finder [1], Data Compare, and Google Search Engine [2]. These file comparison applications can help avoid the copying of file contents from already existing or published contents [3]. Nowadays, there's a lot of applications for comparing files [4-6], and to understand the algorithm for matching the files, this study will be matching text files. An example application in file comparison is Ashisoft's Duplicate File Finder [1], which is a

[1] College of Computer Studies, University of Antique, Sibalom, Antique, Philippines
Email: rizalyn.declines@antiquespride.edu.ph
[2] College of Computer Studies, University of Antique, Sibalom, Antique, Philippines
Email: nikkijane.ducado@antiquespride.edu.ph
[3] College of Computer Studies, University of Antique, Sibalom, Antique, Philippines
Email: kenneth.hiponia@antiquespride.edu.ph
[4] College of Computer Studies, University of Antique, Sibalom, Antique, Philippines
Email: apriljoy.orqiua@antiquespride.edu.ph
[5] College of Computer Studies, University of Antique, Sibalom, Antique, Philippines
Email: gwyn.paduganao@antiquespride.edu.ph
[6] College of Computer Studies, University of Antique, Sibalom, Antique, Philippines
Email: kathymae.toledo@antiquespride.edu.ph
[7]* College of Computer Studies, University of Antique, Sibalom, Antique, Philippines
Email: kenrickagustin.secugal@antiquespride.edu.ph (Corresponding Author)

Received [April 21, 2019]; Revised [June 11, 2019]; Accepted [August 22, 2019]

Published by InnoCon Publishing
ISSN 2704-4440

free application to find and remove duplicate files. It supports an unlimited number of files, folders, and drives. Duplicate Finder finds duplicate photos, songs, documents, spreadsheets, MP3 files, and more. Duplicate files can be found in the storage of your computer. Your computer is not fully optimized until you have removed all unnecessary duplicate files. Thus, Duplicate Finder locates and eliminates these duplicates (*i.e.*, files with the same contents that may include duplicate songs by title, artist, and album). The Duplicate Finder has a built-in picture viewer and music player, as shown in Figure 1. It protects the system files and folders. It is also convenient and easy to use, and it exports the list of duplicates to HTML, CSV, and TXT files.

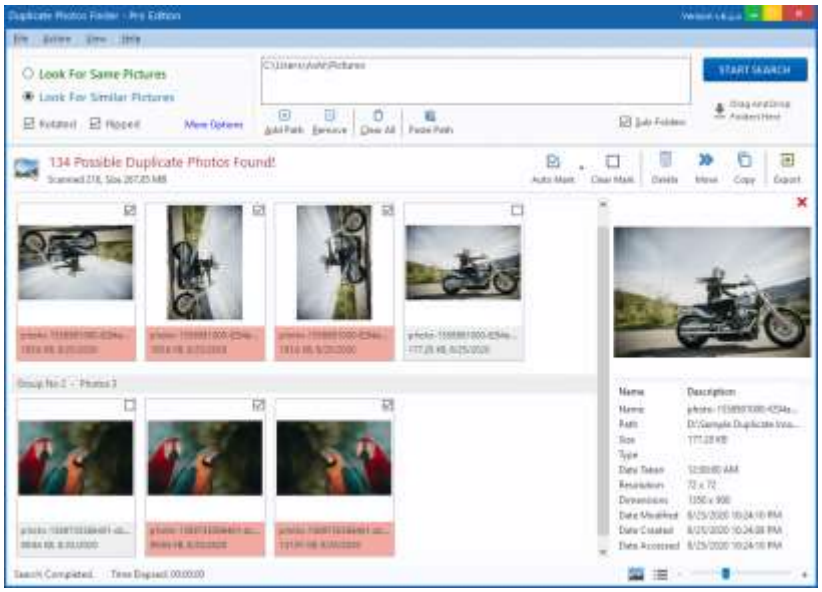

Figure 1. Ashisoft's Duplicate File Finder [1]

Another application for file comparison is ExamDiff [7], which is a freeware Windows tool for visual file comparison, as shown in Figure 2. It is quick and very simple to use, and it has a number of convenient features that many users have been asking for a long time from a file comparison tool.

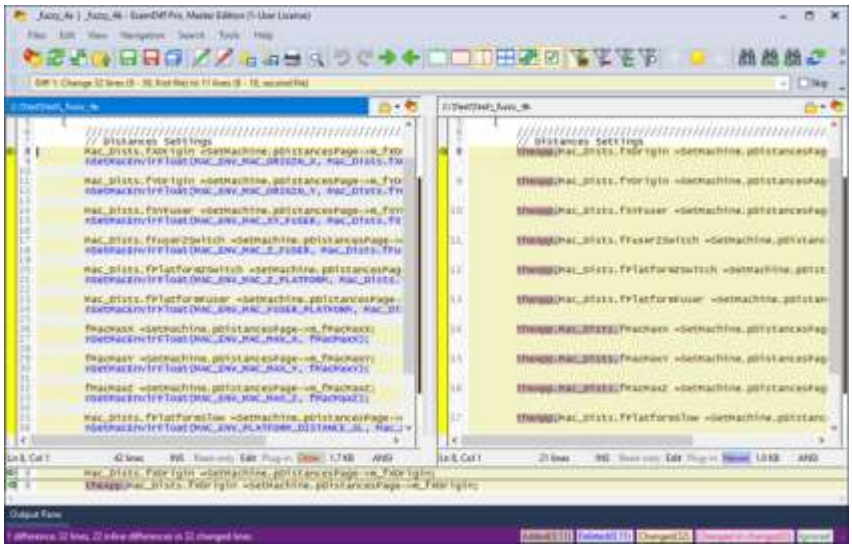

Figure 2. ExamDiff Visual File Comparison Tool [7]

This study aims to develop a prototype that will match two text files. The prototype determines the average rate of accuracy in comparing two text files with regards to their file sizes. In addition, this study aims to determine the average acceptability level of the prototype when evaluated according to its usability, performance, and graphical user interface (GUI) design. The two text files will be matched at a time, and the average rate of accuracy in matching the two text files will be shown. The speed of the prototype in performing the match between the two text files will also be determined by the varying file sizes.

The remainder of this paper is organized as follows: Section 2 outlines the development of the proposed file comparison prototype; Section 3 presents the results and discussion; and Section 4 concludes the study.

2. The Proposed File Comparison Prototype

In this study, a file comparison prototype that can show the average rate of accuracy of the two text files that were matched successfully was developed. This study can be a guide for developing new technological ideas and discoveries about file comparisons.

Figure 3 shows the conceptual design of the operational process of the proposed file comparison prototype. This conceptual design shows the flow of operations for matching two text files using the file comparison prototype. The matching starts by opening the two text files (*i.e.*, File 1 and File 2) that will be matched by the prototype. The two text files are declared matched or not matched and show the average rate of matching accuracy.

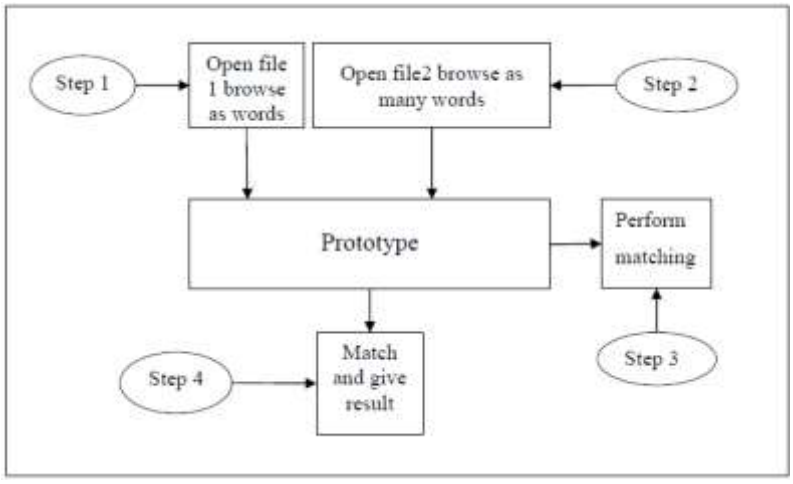

Figure 3. Conceptual Diagram of File Comparison Prototype

The prototype displays the results of the matched contents of the two text files and a decision is given if the two text files, are matched or not, as shown in Figure 4.

Figure 5 shows the main menu of the file comparison prototype GUI for comparing two text files and showing the average rate of matching accuracy of their contents. The main menu has a simple design, only providing two browse buttons to open the two text files to be matched. When the two text files have already been opened, the user may click on the Match button to begin the matching process. Under the match button, a box showing the matching percentage and the time elapsed in seconds is displayed. The match button begins the matching process. The matching percentage shows how much of the contents of the two text files are matched with each other.

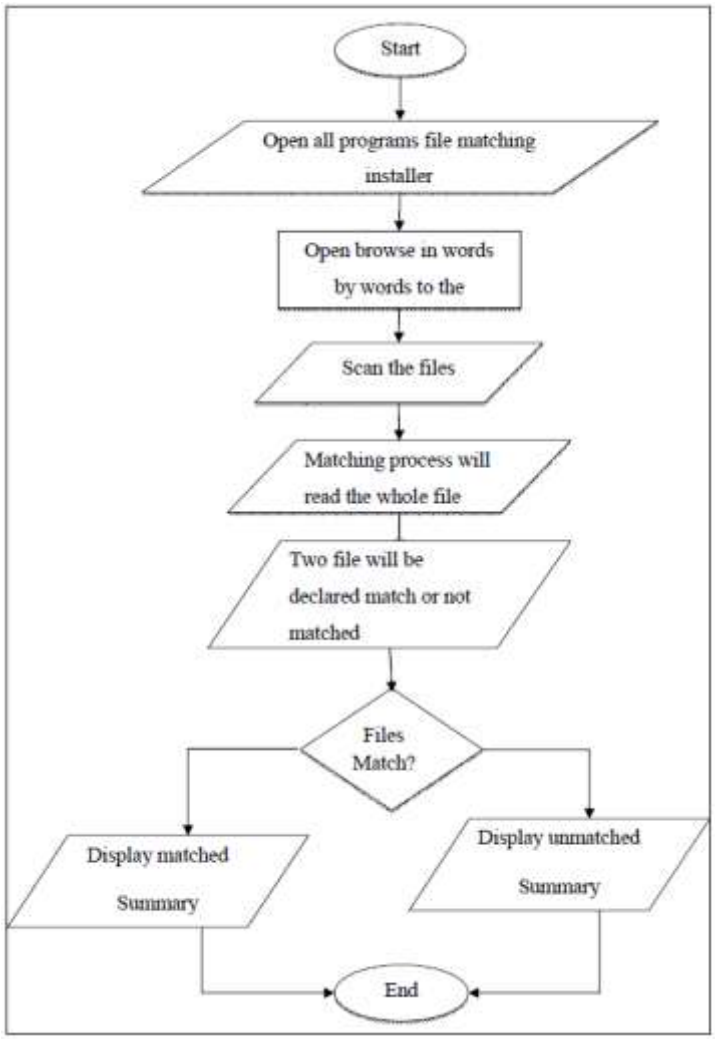

Figure 4. File Comparison Flowchart

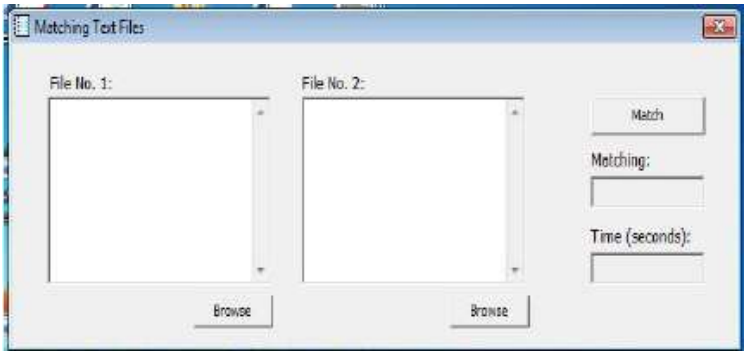

Figure 5. File Comparison Prototype Main Menu

Figure 6 shows the file browser to locate the files to be matched. In the main menu, as shown in Figure 5, files to match can be browsed to locate and open the files to match. The contents of the files to match will then be shown, as depicted in Figure 7.

Figure 6. File Browser

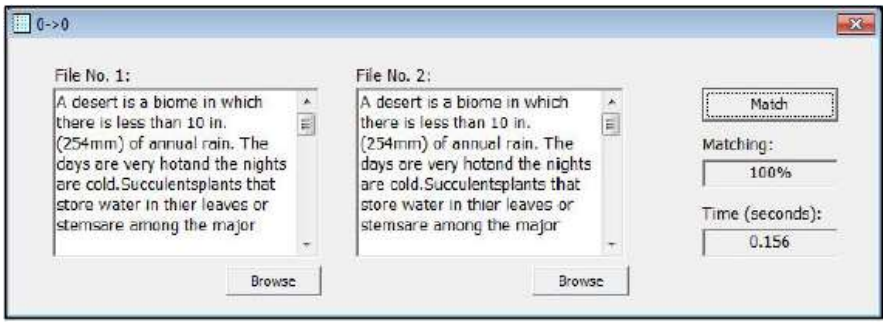

Figure 7. Matching Results

Figure 5 shows the GUI of the matching results of comparing two text files, which indicates the average rate of accuracy and its matching speed. File 1 is matched to File 2 (*i.e.*, the original file) word by word and shows the matched percentage and the time elapsed in matching.

3. Results and Discussion

The following steps were performed in order to evaluate the developed file comparison prototype: the first is to match the two text files using a prototype; the next step is to show the average rate of matching accuracy of the two files; and the final step is entropy coding using the regular expression. A survey is also conducted to measure the acceptability of the prototype. The performance and GUI design of the prototype were evaluated by students and instructors. The percentage of matching a file from the original file was determined using regular expressions and by computing the weighted mean.

Table 1 shows the speed of comparing the two text files in accordance with their file size and contents, as tested by three respondents to show the matching speed of the file comparison prototype. The first respondent tested the file comparison prototype using two files of the same size but with different contents.

The sizes of File 1 and File 2 were the same, but the contents were different in the first test by Respondent 1. In the second test, the two files were of different file sizes and different contents, which shows that the matching of files is the slowest with 0.297 seconds. The third respondent tested the file comparison prototype with two files of different file sizes and different contents as well, which yielded a faster matching speed but a lower matching percentage.

Table 1. Results of Prototype Testing by Three Respondents

Respondents	File 1 size	File 2 size	Match Speed	Matched Percentage
1	1Kb	1Kb	0.172s	100%
2	1Kb	2Kb	0.297s	55%
3	1Kb	2Kb	0.188s	45%

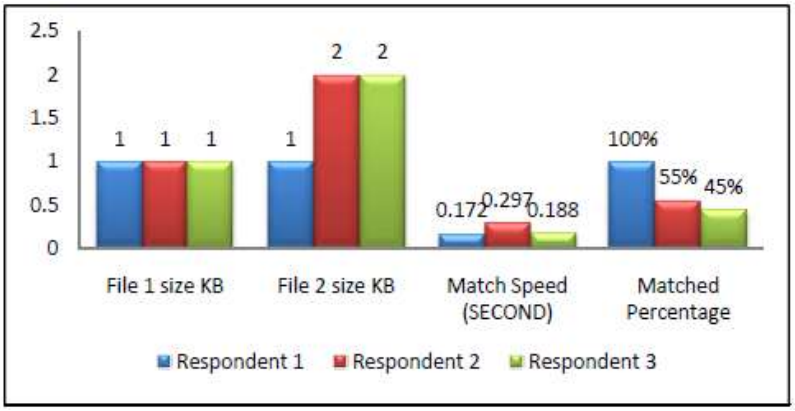

Figure 8. The Results of Prototype Testing by Three Respondents

Figure 8 shows the results of the matching speed of the file comparison prototype as tested by three respondents. In the first testing, the matching speed was determined for comparing two files with the same file size and different contents. The second and third tests are performed with two files of different file sizes and different contents.

Table 2. Results of Evaluation based on the Prototype's Usability, Performance, and Design

Respondents	**Usability**	**Performance**	**Design**	**Total**
Students	3.33	3.67	4.67	3.59
Instructors	3.06	3.67	4.06	3.89

The evaluation was performed by ten respondents, who are comprised of five students and five instructors. Table 2 shows the results of the evaluation with regards to the file comparison's usability, performance, and design of its graphical user interface (GUI). The prototype achieved average usability from five students with a rating of 3.33 and from five instructors with a rating of 3.06, both were interpreted as "Good". In terms of performance, a rating of 3.67 was achieved from both students and instructors. For the GUI design, students' rating was 4.67, and instructors' rating was 4.06. Thus, the average acceptability level from students was 3.59 and from instructors was 3.89.

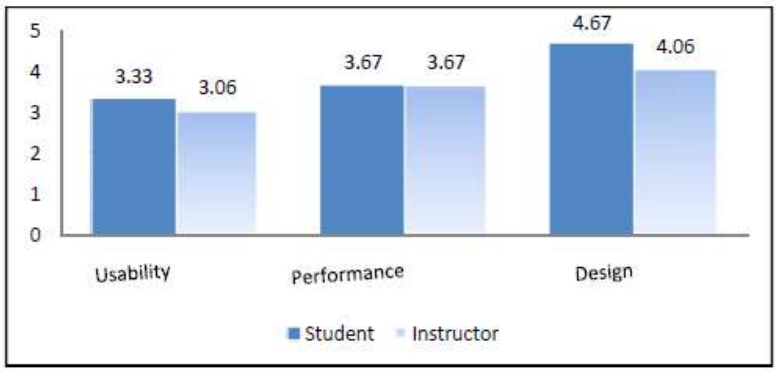

Figure 9. Total percentage Students and Instructor

Figure 9 depicts the results of the evaluation with regards to usability, performance, and design, which were evaluated by five students and five instructors.

The results of the acceptability evaluation of the file comparison prototype are depicted in Table 3. When used, the file comparison prototype exhibits an acceptability weighted mean score of 3.59 from instructors and a 3.89 weighted mean score from students. Both scores are interpreted as "Good".

Table 3. Acceptability Evaluation from Students and Instructors

Respondents	Mean (x)	Interpretation
Students	3.59	Good
Instructors	3.89	Good

4. Conclusion

This paper has proposed a file comparison prototype that measures the matching accuracy of the two text files, as well as the speed of the matching process. It is concluded that the combination of regular expressions is an alternative approach for matching two text files. This file comparison prototype is applicable only to matching two small text files. It matches word by word at a time and performs very fast. The file comparison prototype can perform document matching on different file sizes as the acceptability rate is high. The prototype also managed to achieve a good mark for its usability, performance, and GUI design. The prototype is accurate in matching two text files, even with different file sizes.

References

[1] Ashisoft, "*Duplicate File Finder*", www.ashisoft.com/ (January 14, 2019).

[2] Google Search Central, "*In-depth guide to how Google Search works*", www.developers.google.com/search/docs/fundamentals/how-search-works (January 14, 2019).

[3] Wikipedia, "*Comparison of File Comparison Tools*", www.en.wikipedia.org/wiki/Comparison_of_file_comparison_tools (January 14, 2019).

[4] Diffchecker, "*Diffchecker Desktop*", www.diffchecker.com/ (January 14, 2019).

[5] SQL Data Compare, "*Compare and Deploy SQL Server Database Contents*", Redgate, www.red-gate.com/products/sql-data-compare/ (January 14, 2019).

[6] Textxompare, "*Text Compare!*", www.text-compare.com/ (January 14, 2019).

[7] PrestoSoft, "*ExamDiff*", www.prestosoft.com/edp_examdiff.asp (January 14, 2019).