Mobile-Based Word Matching Detection using Intelligent Predictive Algorithm

https://doi.org/10.3991/ijim.v13i09.10848

Nurul Aisyiah Baharudin (✉), Hamidah Jantan
Universiti Teknologi MARA, Terengganu, Malaysia
Nurulaisyiah93@gmail.com

Abstract—Word matching is a string searching technique for information retrieval in Natural Language Processing (NLP). There are several algorithms have been used for string search and matching such as Knuth Morris Pratt, Boyer Moore, Horspool, Intelligent Predictive and many other. However, there some issues need to be considered in measuring the performance of the algorithms such as the efficiency for searching small alphabets, time taken in processing the pattern of the text and extra space to support a huge table or state machines. Intelligent Predictive (IP) algorithm capable to solve several word matching issues discovered in other string searching algorithms especially with abilities to skip the pre-processing of the pattern, uses simple rules during matching process and does not involved complex computations. Due to those reasons, IP algorithm is used in this study due to the ability of this algorithm to produce a good result in string searching process. This article aims to apply IP algorithm together with Optical Character Recognition (OCR) tool for mobile-based word matching detection. There are four phases in this study consists of data preparation, mobile based system design, algorithm implementation and result analysis. The efficiency of the proposed algorithm was evaluated based on the execution time of searching process among the selected algorithms. The result shows that the IP algorithm for string searching process is more efficient in execution time compared to well-known algorithm i.e. Boyer Moore algorithm. In future work, the performance of string searching process can be enhanced by using other suitable optimization searching techniques such as Genetic Algorithm, Particle Swarm Optimization, Ant Colony Optimization and many others.

Keywords—Intelligent Predictive, Natural Language Processing, Word matching, Mobile-based system

1 Introduction

Nowadays, mobile devices such as smartphone and tablet have become an inseparable part in our daily life activities that used for entertainment, information searching, connecting to customer services, taking photos, GPS location and many others. In advancement of mobile technologies, mobile application industry is growing exponentially whether in small as well as large scale businesses [1]. Since the mobile usage is increasing rapidly, the demand for mobile applications development

in many areas are also increases including in text processing field such as text mining, text summarization, information retrieval and many others. The variety of applications coming up in areas of data and information mining, sentiment analysis, DNA pattern matching etc., will give a new direction of mobile application development for this area.

Text Processing is one of the most common tasks in many machine learning applications such as in information retrieval, machine translation, sentiment analysis, information extraction, question answering etc. These applications deal with huge amount of text to perform classification or translation and it involves a lot of work on the back end. The process of transforming text into something that an algorithm can digest is a complicated process. Text processing for information retrieval referring to the process of searching, creating or manipulating the electronic text for string searching, writing, editing, formatting and printing tasks by using a computer program and related hardware [2]. There are techniques and tools that can be used to transform an electronic text document into the structured format and then the specified algorithm will do the required analysis. Besides that, the understanding of the string search algorithms is also important in this field especially for searching and matching the word that exist in a document until to the whole document in databases [3]. In data preparation, Optical Character Recognition (OCR) is an electronic tool can be used to prepare an electronic document for data analysis. The advantage of using this tool is the document created will be in the form of text-searchable and editable. This approach will help in fast processing and highly accurate performance to make sure the content of document is remains undamaged and it will increase the efficiency and effectiveness in text processing tasks [4].

Boyer Moore, Horspool and Knuth Morris Pratt algorithms are among the string search algorithms widely used for word matching. These algorithms have been applied in text processing, web security, data mining, search engine, medical, mobile system etc. However, these algorithms facing numerous issues especially in measuring the performance of the algorithms. The number of comparisons, comparison time and search space requirements are among the important issues to handle in measuring the performance of the algorithms [5-7]. Intelligent Predictive (IP) algorithm capable to solve several issues in word matching especially it can skip the pre-processing of the pattern, uses simple rules during matching process and it does not involve complex computations [5-6]. Due to those reasons, this study proposed Intelligent Predictive (IP) algorithm as the searching that proven can perform better performance in word matching process especially for the number of comparison [6]. This algorithm uses intelligent predictive analysis based on text features for string searching in text by finding the first occurrence of the pattern in text that consists of words that separated by blank space. In this study, the efficiency of the algorithm was evaluated based on the execution time by comparing the result obtained by another selected algorithm. This study aims to apply intelligent predictive algorithm to search a keyword for word matching detection in mobile-based environment together with the use of OCR electronic tool for data preparation.

This paper organises in the following manner: related work on word matching applications and techniques has been discussed in the second section. The third section

describes the research method. Then the fourth section is on the result analysis and discussions; and finally, followed by the conclusion and future work in the last section.

2 Related Work

2.1 Word matching application and techniques

Word matching or string matching is used to search a specific word in text document that match with the word entered by user. The aim of the applications is to reduce the time for a reader to read the whole text document. Word matching technique will search for the first appearance word and move to other words that exist in the text. A string search algorithm takes a text and a pattern, as the inputs and finds the first or all the occurrences of the pattern. The algorithm will check whether the string pattern for the searched word is appear or not in the text document that is referred [5-9]. If the word is matches, it will return to where the position of the word occurs in the text. Boyer Moore, Horspool, Knuth Morris Pratt and Intelligent Predictive algorithms are among the well-known algorithms for string search in word matching applications. This study focuses on the potential algorithms to be applied i.e. Intelligent Predictive algorithm and Boyer Moore algorithm. Table I shows several word matching application and techniques used.

Table 1. Word Matching Applications and Techniques

Application	Technique	Purpose
Text Processing	Intelligent Predictive Algorithm	Searching the string pattern using intelligent prediction searching method [5].
Web Security	Boyer Moore Algorithm	Detecting the web application vulnerabilities [7].
Data Mining	Knuth Morris Pratt and Naïve Boyer Moore Algorithm	Analysing the selected algorithms as the potential method in Text Mining [8].
Search Engine	Boyer Moore Algorithm	Developing the text editor in commands search and substitute [9].
Medical	Longest Common Substring and Sub-sequences Algorithm	Analysing the similarity of DNA sequence [10].
Mobile Visual Search	Maximally Stable External Region	Searching the paper /article by the title in online database [11].
System Application for Mobile Device	Boyer Moore and Rule Based System	Developing the mobile-based Library Book Information System [12].

2.2 Intelligent predictive algorithm

There are several word matching algorithms have been proposed in this area such as Intelligent Predictive (IP), Knuth Morris, Boyer Moore and many others. IP algo-

rithm works based on intelligent predictive behavior that can perform predictive analysis to predict what will be occurred for the next searching result. It works by searching the first occurrence of a pattern or word in that separated by blank space in a text document [5]. Figure 1 demonstrates the basic process of IP algorithm in string searching. In searching process, IP algorithm starts on the word entered by the user and aligning that word to the leftmost end. Then, it will start to compare the leftmost end of the character. If the word is matched, it will be compared with the rightmost end of the character. The other remaining characters will be comparing from right to left if the matching occurred. In case of a mismatch, this algorithm uses two rules to make a shift namely alphabet-blank mismatch and alphabet-alphabet mismatch. Alphabet-blank mismatch is applied when there is mismatch occur at the leftmost character with alphabet and blank space. In this case, the pattern will be shifted to right by one position. Alphabet that mismatch happened when the leftmost alphabet is not same with the leftmost character. In this case, the pattern is shifted by two positions towards the right. The move of two positions is occurred because the character at the next position might be either a blank space or a character that is a part of the current word to which pattern is aligned.

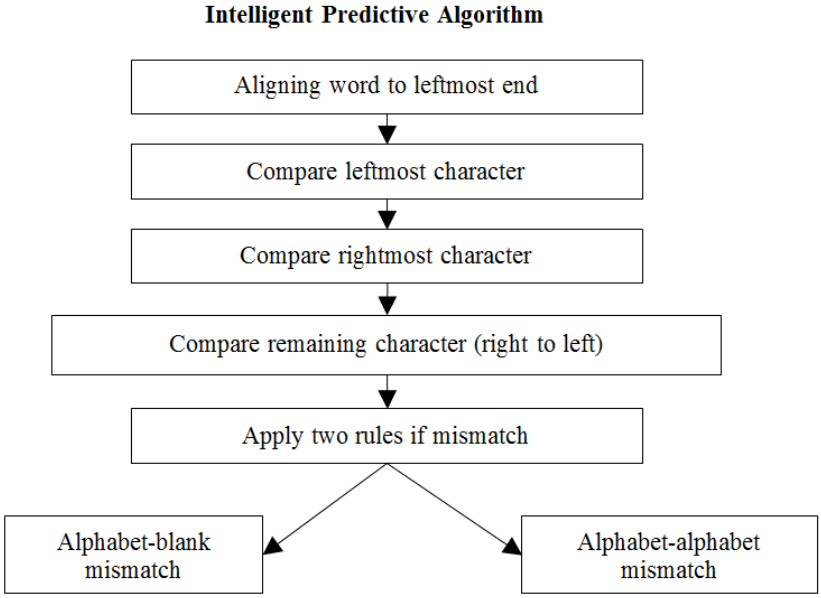

Fig. 1. Intelligent Predictive Algorithm Step-by-step Process [5]

In previous studies as shown in Table 2, intelligent predictive method has been used in text processing for word matching due to their ability in predictive analysis. Besides that, there are several studies that integrate IP algorithm with other method such as Artificial Neural Network to advance the predictive functionalities [13]. IP algorithm also used together with Greedy Dual method to integrate the Web caching and Web prefetching in web application development [14].

Table 2. Appplication using Intelligent Predictive Algorithm

Technique	Purpose	Application Area
Intelligent Predictive Algorithm	Analysis of selected string searching techniques in a text with intelligent predictive method [5-6].	Text Processing
Artificial Neural Network and Intelligent Predictive	Advancing the predictive functionalities for a portable cloudiness, solar radiation and air temperature [13].	Meteorology
Intelligent Predictive and Greedy Dual	Integrating the Web catching and prefetching to reduce the noticeable response time perceived by user [14].	Web Application

2.3 Boyer Moore Algorithm

Boyer Moore (BM) algorithm is an efficient string searching algorithm and it has been the standard benchmark for other string searching algorithm. The algorithm gathers information during the pre-processing phase which is used to skip sections of the text it is searching for [15-16]. BM algorithm does not need to check every character of the string to be searched, but rather skips over some of them. Generally, the algorithm gets faster as the word being searched for becomes longer. Figure 2 shows the basic searching process using Boyer Moore algorithm [5].

Boyer Moore algorithm starts to do searching process by aligning the pattern to the leftmost end of text. Before starting the comparison, the pattern needs to be processed first. This pre-processing step is required to determine the number of shifts in case of mismatch. After pre-processes the pattern, it will compare the pattern with the text from right to left. There are two rules uses in Boyer Moore algorithm in case of mismatch of a pattern character such as bad character heuristic and good suffix heuristic.

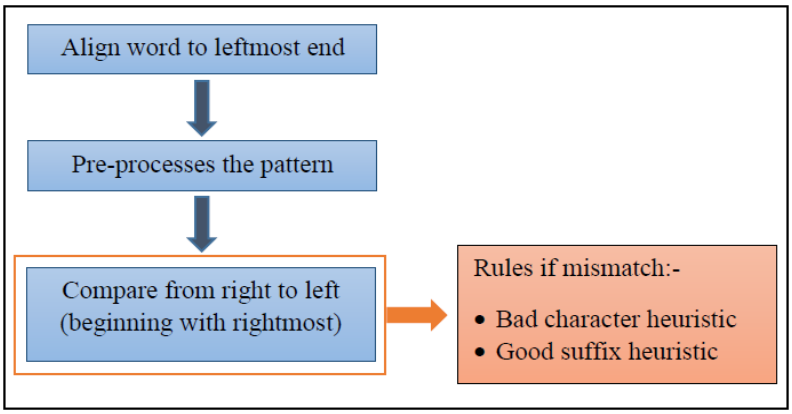

Fig. 2. Boyer Moore Algorithm Basic Process

Both rules are used to determine the pattern shift if mismatch of the character occur. Bad character heuristics applied in the Boyer Moore algorithm if the bad character which is the character in the text that causes a mismatch, occurs anywhere in the pattern. If this happens, the pattern can be shifted as to align the pattern with the char-

acter in the text. Good suffix heuristics is applied in Boyer Moore algorithm when the bad character heuristics fails. An alignment of the rightmost character occurrence of the pattern with the character in text would produce a negative shift. Instead, a shift by one would be possible. However, in this case it is better for the algorithm to derive the maximum possible shift distance from the structure of the pattern which is called good suffix heuristics.

This algorithm has been applied in many areas such as in web application for vulnerabilities detection in detecting the SQL injection, Buffer Overflow, Cross Site Scripting and Cross Site request Forgery [7]. Besides that, this algorithm has been compare with other string matching algorithms such as Knuth Morris Pratt and Naïve to determine the most suitable algorithm in Text Mining [8]. In another application in word matching is text editor development that used to search and substitute the commands [9]. Boyer Moore algorithm is known works best when the alphabet is moderately sized, and pattern is relatively long. Besides that, a hybrid of Boyer Moore and Rule Based System was introduced for library book information for mobile–based environment [12].

There are several advantages using Boyer Moore Algorithm. BM algorithm is very fast when working on a large alphabet and it very good to use when working on binary strings as well. If the pattern however is very short or has a low probability to be found then the algorithm is not so optimal to use [7, 9]. However, for the disadvantages, this algorithm suffers from the phenomenon that it tends to work inefficiently on small alphabets like DNA [10]. The skip distance tends to stop growing with the pattern length because substrings re-occur frequently. Besides that, the pre-processing for the good suffix heuristic is difficult to understand and implement.

3 Research Method

This section presents the step-by-step processes involved in this study that consist of four stages i.e. data preparation; mobile user interface design; algorithm implementation; and the evaluation and analysis. Figure 3 shows the research framework for word matching detection using IP algorithm. As comparative study, Boyer Moore algorithm was selected for word matching task using same research design to analyses the IP algorithm performance compared to this algorithm. The first stage is data preparation that involving the process of capturing the text character for a document, analyses the captured image and transform it into character code using OCR tool. This tool can be used to produce printed text in image form and convert it into text file [17]. In this stage, the captured or scanned text will be transformed into image file such as in PNG, JPEG or BMP files.

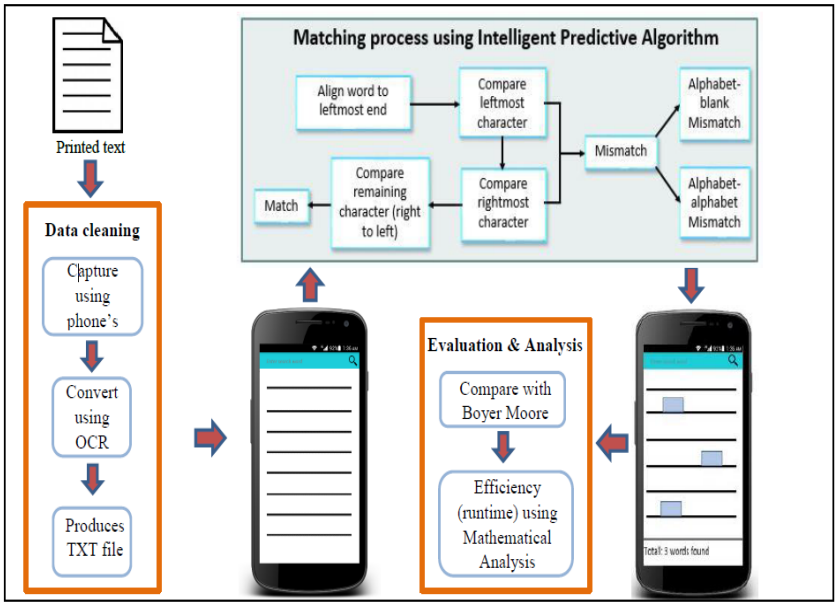

Fig. 3. Mobile-Based Word Matching Process

Then, the OCR tool is used to extract the word from the captured image and convert it into the readable text which is .txt file. Figure 4 shows the sample of image data that captured using phone.

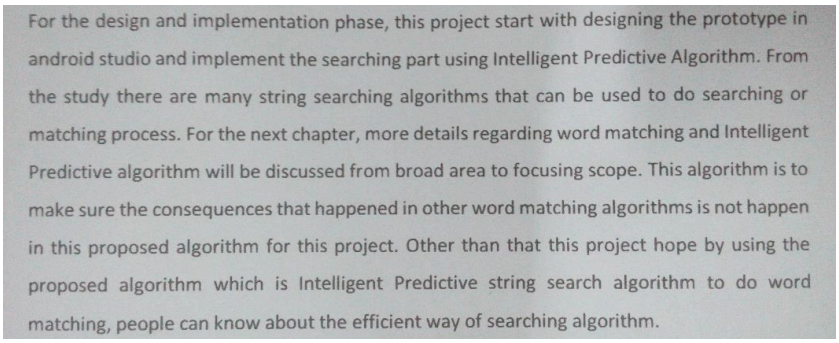
For the design and implementation phase, this project start with designing the prototype in android studio and implement the searching part using Intelligent Predictive Algorithm. From the study there are many string searching algorithms that can be used to do searching or matching process. For the next chapter, more details regarding word matching and Intelligent Predictive algorithm will be discussed from broad area to focusing scope. This algorithm is to make sure the consequences that happened in other word matching algorithms is not happen in this proposed algorithm for this project. Other than that this project hope by using the proposed algorithm which is Intelligent Predictive string search algorithm to do word matching, people can know about the efficient way of searching algorithm.

Fig. 4. Sample Printed Text

The second stage is mobile user interface was designed that based on Human Computer Interaction (HCI) design. The user interface designed was very simple and it follow the concepts of HCI design which is easy to use and user friendly. The first user interface of this application was the home view. In this first view design, user must click on the Click to Start Button to start using the application. Once the button is clicked, this application moved to the camera view which is the second interface design for this application. For the camera view, this application needs to access the

mobile device camera to capture the image. At this point, the user will capture the printed text to search for the word. The printed text can be either in English or Malay text. Figure 5 shows the design of home view and camera view.

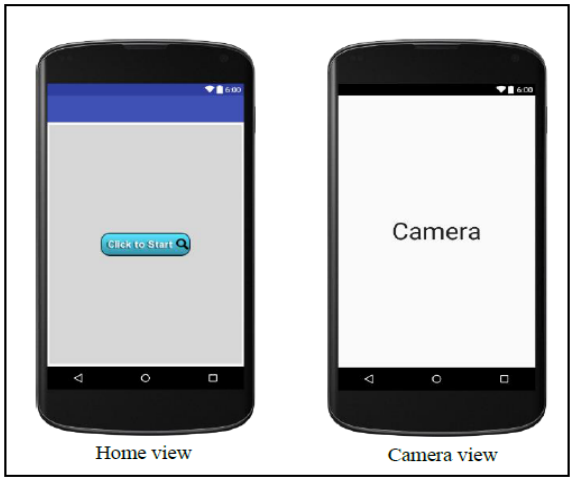

Fig. 5. User Interface for Home and Camera Views

The third and fourth interface design are text and result views for displaying the text as input and result after word matching process being implemented. Figure 6 shows the design of text view and result view.

Fig. 6. User Interface for Text and Result Views

The image of text will be capture using camera button, and then the application will be moved to the third view which is the text view. At this view, it will show the

result of captured image that has been converted into the .txt file performed by the OCR tool. This text view design also includes the search box design. The use of this search box is to allow the user to key in any searching word. Once user key in the searching word, this application will show the result view. At this view, the result is shows by highlighting the word if the word searched by user was found. At the bottom of this view, it shows the result of how many words was found in the text. The result will show 0 if the word entered by user is not found in the text. The third stage is focus on algorithm implementation for string searching process using IP algorithm. The algorithm has been executed together with mobile application development in searching process. In the fourth stage, Boyer Moore algorithm are applied in searching process as comparative study for performance analysis. The performance of the algorithms was evaluated based on the execution time required in their searching process.

4 Result & Discussions

This study proposes the potential string search algorithm for mobile–based word detection by analyzing their execution time as performance analysis in searching process. The performance of Intelligent Predictive (IP) algorithm and Boyer Moore algorithm as a benchmark of string-matching algorithms was conducted. The evaluation method used is execution time analysis. This method measures the efficiency or running time of the algorithm takes to search the string or pattern. Table 3 shows the result of execution time analysis for Intelligent Predictive algorithm and Boyer Moore algorithm based on the number of words.

Table 3. Execution Time Analysis

Number of Word	Execution Time (ms) of Intelligent Predictive Algorithm	Execution Time (ms) of Boyer Moore Algorithm
100	0.2648	1.1696
200	0.5887	1.3931
300	1.1712	1.5653
400	1.3162	1.6306
500	1.4167	1.7213

The number of words used in testing process for both algorithms can be categorized into five categories number of words used which is from 100 data to 500 words. The execution time was measured in nanosecond and recorded in millisecond. The results of the execution time for both algorithms increased as the number of words increased. The result of the time efficiency for both algorithms was represented in Figure 7

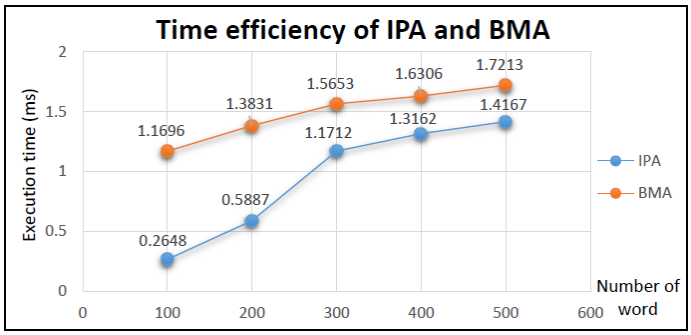

Fig. 7. Time Efficiency of Intelligent Predictive Algorithm and Boyer Moore Algorithm

The graph demonstrates that increasing number of words will cause increasing in execution time for both algorithms. Intelligent Predictive algorithm shows more efficient in execution time as compared with Boyer Moore algorithm. Intelligent Predictive algorithm takes less time to do the searching process. It shows that, in this study the searching process of the Intelligent Predictive algorithm has proven more effective and faster. Boyer Moore algorithm works well in finding a long pattern in text but work inefficiently for the small pattern in text. Intelligent Predictive algorithm is working well in finding of small pattern and suitable for application that required small set of text.

5 Conclusion

In the proposed mobile based application, user can find a word from printed text esaily and quickly. The OCR tool is used in this study to process the string input in image form and tranform it into text file. Intelligent Predictive algorithm was applied due to the ability of this algorithm to improve the time taken for the searching process. In future work, this study can be enhanced by implementing other optimization algorithm to improve the searching process in word matching such as bio-inspired algorithms, Genetic Algorithm, Particle Swarm Optimization, Ant Colony etc.

6 Acknowledgement

This research has been supported by Ministry of Education (MOE) to Universiti Tekonologi MARA (UiTM) for Fundamental Research Grant Scheme (FRGS) (600-RMI/FRGS 5/3 (3/2016).

7 References

[1] Mantra (2017), Future of Mobile Application Development", available online: https://www.valuecoders.com/blog/technology-and-apps/future-of-mobile-application-development/ last visit:28.01.2017

[2] Forgac R, Krakovsky R and Mokri I (2015), A Contribution to Modification of PART Clustering Algorithm for Text Processing, IEEE 19th International Conference on Intelligent Engineering Systems (INES), Bratislava, Slovakia, pp:421-425, https://doi.org/10.1109/ines.2015.7329747

[3] Miner G, Delen D, Elder J, Fast A, Hill T and Nisbet R (2012), The Seven Practice Areas of Text Analytics,in Practical Text Mining and Statistical Analysis for Non-Structured Text Data Applications, ed: Elsevier, pp.29-41. https://doi.org/10.1016/b978-0-12-386979-1.03001-2

[4] Scan P (2017), The Advantages of OCR (Optical Character Recognition), available online: https://www.pearl-scan.co.uk/blog/advantages-of-ocr-optical-character-recognition, last visit:24.01.2017 https://doi.org/10.3403/30309311

[5] Gurung D, Chakraborty U K and Sharma P (2016), Intelligent Predictive String Search Algorithm, Procedia Computer Science, Vol. 79, pp.161-69, https://doi.org/10.1016/j.procs.2016.03.116

[6] Gurung D, Chakraborty U K and Sharma P (2017), An analysis of the Intelligent Predictive String Search Algorithm: A Probabilistic Approach, Information Technology & Computer Science, Vol. 2, pp.66-75, https://doi.org/10.5815/ijitcs.2017.02.08

[7] Saleha, Ain Zubaidah Mohd, AmizahRozalia, Nur, Buja, Alya Geogiana, Jalil, Kamarularifin Abd., Ali, Fakariah Hani Mohd and AbdulRahman, Teh Faradilla (2015), A Method for Web Application Vulnerabilities Detection by Using Boyer-Moore String Matching Algorithm, Procedia Computer Science, Vol. 72, pp. 112-121, https://doi.org/10.1016/j.procs.2015.12.111

[8] Sheshasayee A and Thailambal G (2015), A comparitive analysis of single pattern matching algorithms in text mining, International Conference on Green Computing and Internet of Things (ICGCIoT), Noida, India, pp. 720-725, https://doi.org/10.1109/icgciot.2015.7380557

[9] Kurniawan D H and Munir R (2015), A New String Matching Algorithm Based on Logical Indexing, The 5th International Conference on Electrical Engineering and Informatics, Denpasar, Indonesia, pp. 394-399, https://doi.org/10.1109/iceei.2015.7352533

[10] Alsmadi I and Nuser M (2012), String Matching Evaluation Methods for DNA Comparison, International Journal of Advanced Science and Technology, Vol. 47, pp. 13-32

[11] Tsai S S, Chen H, Chen D, Vedantham R, Grzeszczuk R, and Girod B (2011), Mobile Visual Search on Printed Documents using Text and Low Bit-rate Features, 18th IEEE International Conference on Image Processing (ICIP), Brussels, Belgium, pp. 2601-2604, https://doi.org/10.1109/icip.2011.6116198

[12] Basari, Abd Samad Hasan, Yusop, Noorrezam, Hussin, Burairah and Shibghatullah, Abdul Samad (2014), Hybrid of Boyer Moore and Rule based System for Mobile Library Book Information, International Journal of Computer Applications, Vol. 90, pp. 21-28. https://doi.org/10.5120/15570-4144

[13] Ferreira P, Gomes J, Martins I, and Ruano A (2012), A Neural Network Based Intelligent Predictive Sensor for Cloudiness, Solar Radiation and Air Temperature, Sensors, Vol. 12, p. 15750-15777, https://doi.org/10.3390/s121115750

[14] Patil J B and Pawar B V (2011), Integrating Intelligent Predictive Caching and Static Prefetching in Web Proxy Servers, International Journal on Computer Science and Engineering (IJCSE), Vol. 3, No. 2, pp. 697-704.
[15] Mukku B S, Rachita B and Preeti N (2018), String Matching Algortihms, International Journal of Engineering& Computer Science (IJECS), Vol. 7, No. 3, pp. 23769-23772.
[16] Charras C and Lecroq T (2004), Handbook of Exact String Matching Algorithms: College Publications
[17] Wankhede P A and Mohod S W (2017), A Different Image Content-based Retrievals using OCR Techniques, International conference of Electronics, Communication and Aerospace Technology (ICECA), Coimbatore, India, Vol. 2, pp. 155-161, https://doi.org/10.1109/iceca.2017.8212785

8 Authors

Nurul Aisyiah Baharudin holds a bachelor's degree in computer science that was obtained from Universiti Teknologi MARA, Terengganu, Malaysia in 2017.

Hamidah Jantan received the first degree in Computer Science in 1989 from Universiti Teknologi Malaysia (UTM), Kuala Lumpur, Malaysia. She obtained her master's degree in information technology (Science and System Management) from Universiti Kebangsaan Malaysia (UKM) in 2002, and Ph.D. degree in the Faculty of Information Science and Technology (FTSM), Universiti Kebangsaan Malaysia (UKM), Malaysia in 2011. Currently, she is an associate Professor at Faculty of Computer & Mathematical Sciences, Universiti Teknologi MARA (UiTM) Terengganu, Malaysia. Her research interests include Intelligent System, Decision Support System and Data Mining technology.

Article submitted 2019-05-12. Resubmitted 2019-07-10. Final acceptance 2019-07-10. Final version published as submitted by the authors.