Information Retrieval In Text-Based Document Using Boyer Moore Algorithm

Yudhi Setyo Purwanto*[1], M. Farid Rifai[2], Hendra Jatnika[3], Gina Afra Ardelia[4]
[1,2,3,4]Institut Teknologi PLN; Menara PLN, Jl. Lingkar Luar Barat, Duri Kosambi, Cengkareng, Jakarta Barat, DKI Jakarta, 11750, Ph. (021) 5440342
Informatics Engineering, Faculty of Telematics Energy, ITPLN, Jakarta
e-mail:[1]y.purwanto@itpln.ac.id, [2]m.farid@itpln.ac.id, [3]h.jatnika@itpln.ac.id, [4]gina1731239@itpln.ac.id*

Abstrak

Language Development Center (LDC) merupakan laboratorium pengembangan bahasa yang setiap tahunnya melayani lebih dari 1000 peserta dalam kegiatan-kegiatan seperti ujian standarisasi dan pelatihan bahasa bagi mahasiswa dan dosen. Kegiatan tersebut menghasilkan banyak data yang dalam pengelolaannya masih belum efektif dan menyebabkan beberapa masalah dalam pencarian, pengarsipan, dan pemantauan data. Oleh karena itu, penelitian ini dibuat dengan tujuan untuk membuat sebuah sistem agar pengelolaan data di LDC menjadi lebih efektif. Dalam perancangan dan pembangunan sistem ini, penulis menggunakan metode Waterfall dan juga mengimplementasikan algoritma Boyer Moore dalam proses pencarian data. Algoritma boyer moore membantu proses pencarian data lebih efektif dan dapat mencari kata dalam beberapa file. Hasil dari penelitian ini yaitu sebuah sistem informasi manajemen data berbasis web yang menjamin keamanan dan integrasi data, kecepatan dan efektifitas akses data, juga ketepatan dan kecepatan dalam pencarian data. Dari sisi pengguna, sistem ini juga memangkas alur kerja, mengurangi tenaga dan waktu yang digunakan, juga dapat menghemat biaya-biaya operasional kegiatan.

***Kata kunci**: sistem informasi, manajemen data, penarikan data, algoritma Boyer Moore*

Abstract

The Language Development Center (LDC) is a language laboratory that annually serves more than 1000 participants in activities such as standardized tests and language courses for students and lecturers. This activity generates a lot of data which is still managed ineffectively and causes several problems in data retrieval, archiving, and monitoring. Therefore, this research was made with the aim of creating a system so that data management in LDC becomes more effective. In designing and building this system, the writers uses the Waterfall method and also implements the Boyer Moore algorithm in the data search process. Boyer moore algorithm helps the data search process more effectively and can search for words in several text-based files at once. The result of this research is a web-based data management information system that ensures the data security and integration, as well as the speed and efficiency of data access processes. The algorithm provides accuracy and effectiveness in data retrieval process. From the users' perspective, this system also reduces the time and energy used and save operational costs.

Keywords: information system, data management, data retrieval, Boyer Moore algorithm

1. INTRODUCTION

The Language Development Center (LDC) is a language development laboratory located at the PLN Institute of Technology (ITPLN). LDC has many activities every year such as English test activities, courses, workshops, etc. Each of these activities requires and produces a lot of important data to be managed, such as data in the form of documents, certificates, and inventory data. However, there is no data management information system that can manage these data properly, causing several problems such as unintegrated data storage, data retrieval that takes a lot of time, and many more.

One solution in dealing with data management problems is to build a data management information system that has document and certificate search features. At ITPLN, an English exam certificate at the ToEFL level is one of the requirements for the thesis defense's program at the end of the study period. These requirements must be collected and verified by the academic supervisor. The issue of certificate authenticity is something that must be seriously considered because of the many image manipulation applications currently circulating. Previously, LDC provided time and place to accommodate this manually, but there were many obstacles and shortcomings, such as: missing data, undocumented data, the service takes a long time to process, and can only be done at certain times (office hours).

Until now, after only 4 years in service, LDC has conducted an English test for final year students of approximately 3000 participants and the number will continue to grow. With this much of information, it will be difficult for employees to find the right data and verify its authenticity. Based on these things, LDC then created a system that can make the process of checking documents, especially certificates run automatically and quickly. This system is web-based and uses the Boyer-Moore algorithm. This system is designated as part of the Smart Data Management System (SDMS) which is a one-stop learning management that contains all learning and exam features, from announcements, registration, scheduled login, learning system, exam system, to the distribution of certificates, data management, and reporting.

This research is based on previous studies, and is divided into two parts, namely those related to: 1) management information system (MIS), and 2) information retrieval system (IR system). Research on MIS was carried out by Syam [1] to assist in the formulation and implementation of policies at universities, and Bahagia [2] in making data management for victims of natural disasters so that they can convey information that is coherent, orderly, and precise. In the studies on the IR system, Bunyamin and Negara [3], also Amin and Purwaningtyas [4] used the vector space model method that uses the concept of vector space which converts documents into vectors that are used as references in determining the relevance of input to documents. While Saadah et al. [5] and Chiranjeevi and Shenoy [6] did so using the Term-Frequency-Inverse Document Frequency (TF-IDF) method which considered the occurrence of the same word order between the query and the text in the document. Danuri [7] conducted research on content-based text search using the Brute Force algorithm and Aruleba et al. [8] using a full text retrieval system in the digital library. The writers use the Boyer Moore algorithm to settle the information retrieval issue in their data management system.

In the data search process, an algorithm that can facilitate the data search process is needed. Boyer Moore algorithm is one of the string search algorithms which can be said to be the most effective algorithm in everyday applications [9]. In the string search process, the Boyer Moore algorithm reads characters in a right-to-left pattern. The Boyer Moore algorithm performs string matching using 2 techniques, namely the looking-glass technique which is a technique for matching the patterns contained in the text starting from the character in the text that is the last or rightmost then to the earliest pattern. The second technique is the character-

jump technique where if there is a mismatch between the pattern and the text, a character shift is carried out [10].

The purpose of this research is to design and build a data management information system that aims to help the data management process to be more effective and efficient and also to implement the data retrieval process in the certificate verification mechanism using the Boyer Moore algorithm in the data management information system at LDC.

2. METHODS

There are several stages in carrying out this research. These stages are starting from problem identification to report generation. In system development, to get along with the proper research flow, the writers use the waterfall method with the stages of problem identification, data collection, system requirements analysis, application design, implementation, and testing [11]. Meanwhile, in conducting research design, the authors use the IR flow process as follows:

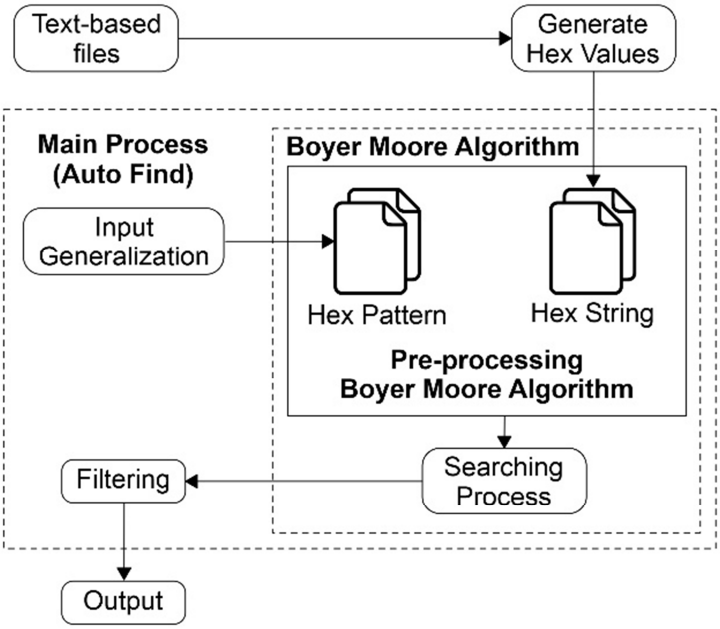

Figure 1. Information Retrieval Flow Process

Boyer Moore algorithm is an algorithm that is applied to research in the data search process. The process of the Boyer Moore algorithm is as follows:

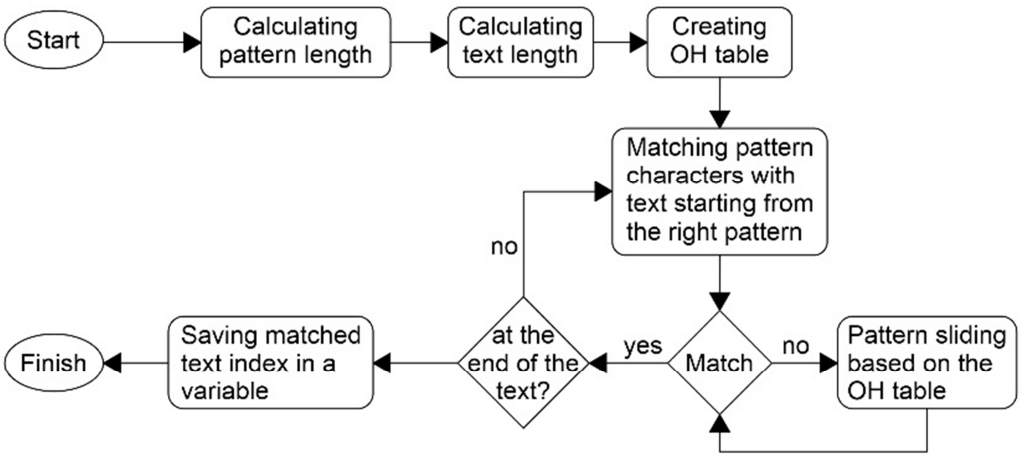

Figure 2. Boyer Moore Algorithm Flow Diagram

The data search process in this study aims to find keywords contained in the file. The explanation of the steps of the Boyer Moore algorithm according to Figure 2 above is as follows:

a. The system calculates the length of the pattern entered by the user.
b. The system calculates the length of the text contained in a file.
c. The system creates an OH table according to the pattern length that will be used to calculate the number of pattern shifts.
d. Match the pattern with the initial text starting from the rightmost index of the pattern.
e. If no match is found, then the pattern shift is carried out according to the OH table that has been made to get a match between the pattern and the text.
f. If there is a match, it will check whether the text is at the end.
g. If not, then re-matched according to step d.
h. When the text has been terminated, the index text and those that match the pattern will be stored in a variable.

In this study, the Boyer Moore algorithm is used in the data search process. The examples of calculations for the Boyer Moore algorithm are as follows:

Text (T) = GINA AFRA ARDELIA
Pattern (P) = LIA

i	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
T	G	I	N	A		A	F	R	A		A	R	D	E	L	I	A
P	L	I	A														

a. *Occurrence Heuristic* (OH) Table
The OH table is used as the shift value obtained when finding a character mismatch. OH value is obtained with max (1, number of characters – character index – 1)

Pattern	L	I	A	*
OH value	2	1	1	3

b. Pattern and Text matching

1) Matching starts from the characters on the rightmost pattern, namely 'A' and 'N' characters in the text. Because a mismatch is found, then the 'N' character will be matched with the OH table, because no similarity is found, the shift is done 3 times.

i	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
T	G	I	N	A		A	F	R	A		A	R	D	E	L	I	A
P	L	I	A														

2) The next match is 'A' character in the pattern, and 'A' character in the text. Because character A is found/matched, then the process is continued to the left side of the pattern, namely the 'I' character in the pattern and the space character in the text. Because in the second match no similarities were found, then the space character was checked again in the OH table, if it is not found, the pattern was shifted for 3 times.

i	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
T	G	I	N	A		A	F	R	A		A	R	D	E	L	I	A
P				L	I	A											

3) The next match is 'A' character in the pattern and 'A' character in the text, because a character match is found, then the matching shift is made to the left, namely 'R' character in the text and 'I' character in the pattern. Because it does not match, a shift is made according to the OH table.

i	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
T	G	I	N	A		A	F	R	A		A	R	D	E	L	I	A
P							L	I	A								

4) The next match is the 'R' character in the text and the 'A' character in the pattern. Because a mismatch was found, a shift was made according to the OH table.

i	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
T	G	I	N	A		A	F	R	A		A	R	D	E	L	I	A
P										L	I	A					

5) Next, matching the 'L' character in the text and 'A' character in the pattern, there is an inequality between the two and a check is made on the OH table, because the 'L' character is in the OH table, then the shift is carried out according to the OH value of the 'L' character, which is done in 2 shifts.

i	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
T	G	I	N	A		A	F	R	A		A	R	D	E	L	I	A
P													L	I	A		

6) Next is matching 'A' character in the text and 'A' character in the pattern. Because a match is found, then the check is continued on the character on the left and another match is found until the character in the pattern has run out. Because all the characters in the pattern match the text, the string match was found.

i	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
T	G	I	N	A		A	F	R	A		A	R	D	E	L	I	A
P															L	I	A

3. RESULT AND DISCUSSION

The result of this research is a web-based data management information system in LDC. This system is implemented using the programming language PHP, HTML, CSS, Javascript, and MySQL database. The interface of this system is as follows:

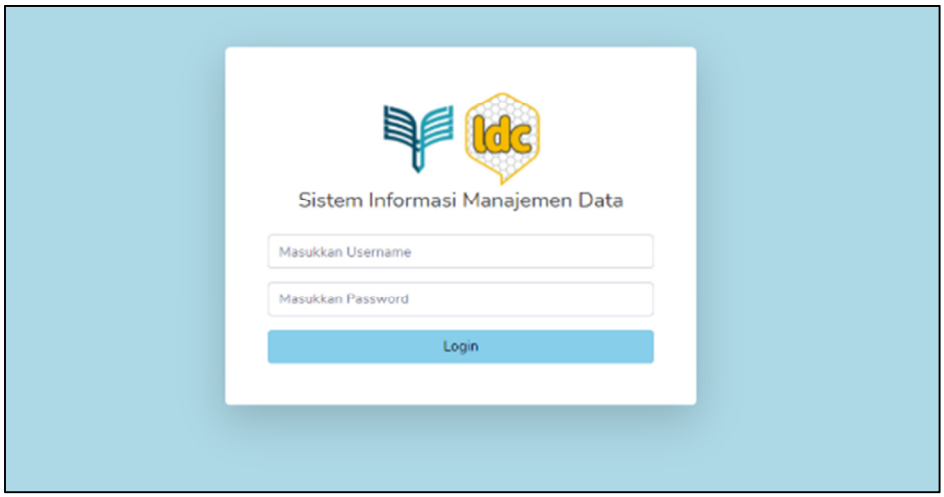

Figure 3. The Login page preview

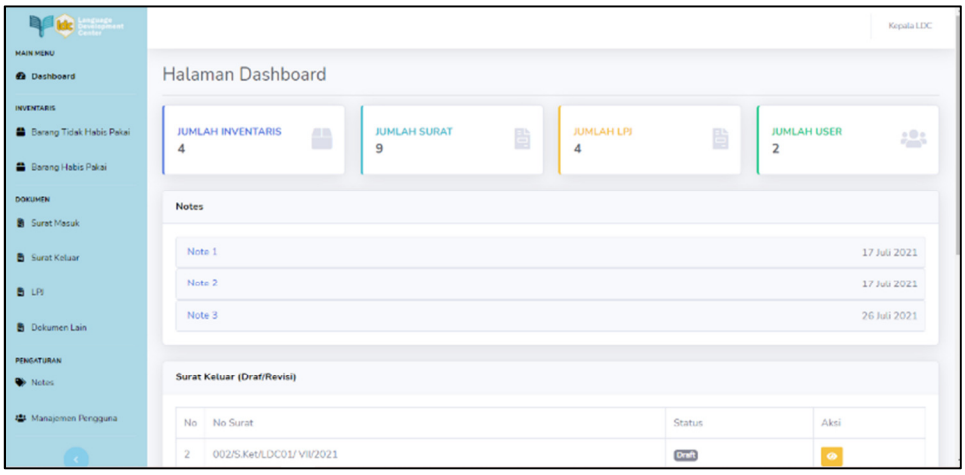

Figure 4. The Dashboard page preview

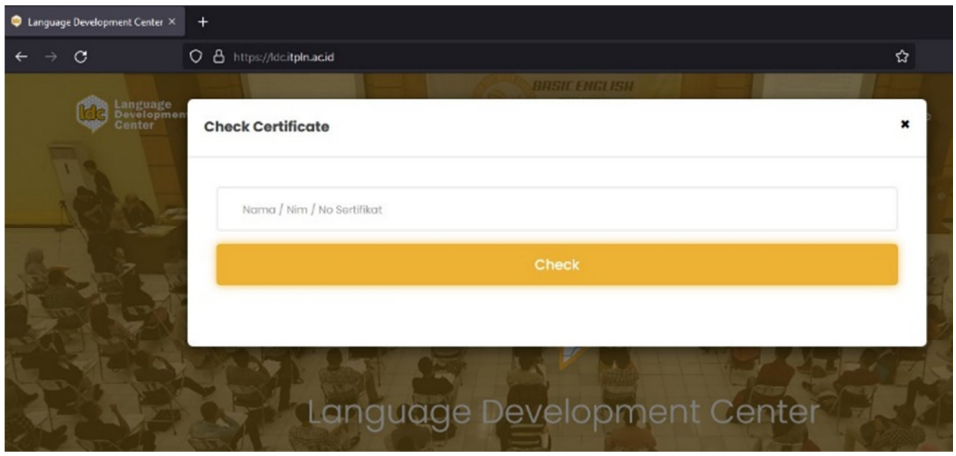

Figure 5. The Certificate Check Page Preview

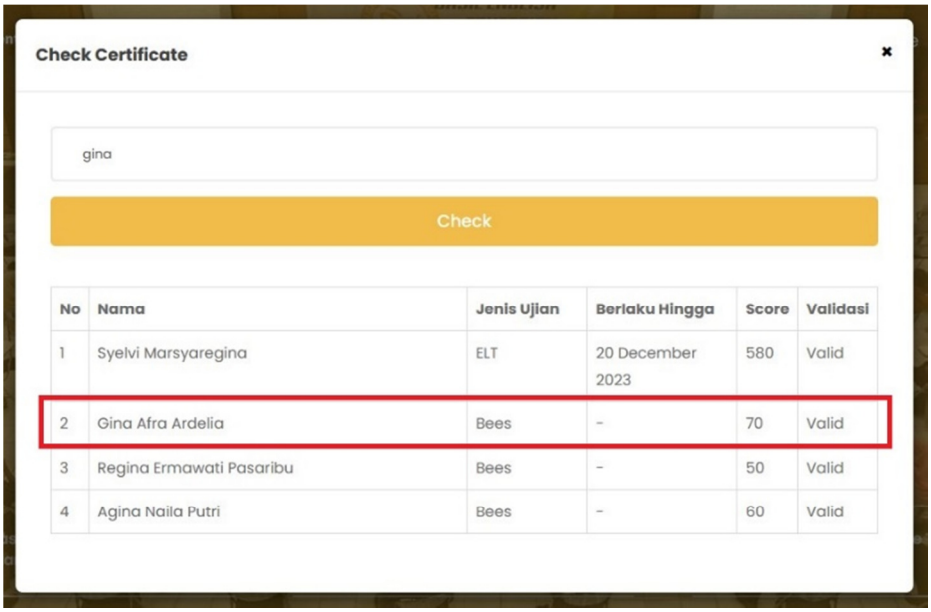

Figure 6. The Search on Certificate Check Page Preview

The results of the search for data using the Boyer Moore algorithm were tested with 8 files of EprT test results from 2018-2021 with a total data of 2400 participants as follows:

No.	*Keyword*	*Expected Result*	*Result*	Vld/Not
1.	Gina Afra	EPrT Jan 2021 participant	EPrT Jan 2021 participant	valid
2.	Galang Rizqi	EPrT Nov 2019 participant	EPrT Nov 2019 participant	valid
3.	Arief Cahya	EPrT April 2019 participant	EPrT Apr 2019 participant	valid
4.	Syahrul Rama	EPrT Jan 2020 participant	EPrT Jan2020 participant	valid
5.	Bobi Sukma	EprT Jan 2019 participant	EprT Jan 2019 participant	valid
6.	Dian	EPrT Dec 2018 participant EPrT Apr 2019 participant EPrT Apr 2019 participant EPrT Apr 2019 participant	EPrT Dec 2018 participant EPrT Apr 2019 participant EPrT Apr 2019 participant EPrT Apr 2019 participant	valid

		EPrT May 2019 participant EPrT Nov 2019 participant EPrT Jan 2020 participant EPrT Jan 2021 participant	EPrT May 2019 participant EPrT Nov 2019 participant EPrT Jan 2020 participant EPrT Jan 2021 participant	
7.	Arinil Khaerah	EPrT Jan2021 participant	EPrT Jan 2021 participant	valid
8.	Anisa Alyana	EPrT Jan2020 participant	EPrT Jan 2020 participant	valid
9.	Zulfira	EPrT Jan 2020 participant	EPrT Jan 2020 participant	valid
10.	Erlangga	EPrT Jan 2021 participant	EPrT Jan 2021 participant	valid

Out of the 10 data tested, all of them produced results in accordance with the expected results. In data that is only contained in one file, only one file is displayed, while the data contained in many files, then the file that is displayed is more than one.

4. CONCLUSION

Based on the research conducted, it can be concluded that the existence of a data management information system at LDC can help the data management process to be more effective than before, such as integrated data storage, faster data retrieval, and can be accessed directly by users. In addition, the application of the Boyer Moore algorithm can help the data search process in the LDC become more effective.

5. SUGGESTIONS

The writers would like to suggest that the next research on this issue can be focused on the development of a bigger environment with more various data, e.g., pictures, scanned documents, or even the web contents. The data retrieval can also be done by mixing the methods and algorithms to achieve a faster data retrieval process in the future.

ACKNOWLEDGEMENTS

The team expresses the greatest appreciation and gratitude to the Indonesia Directorate General for Higher Education, Research, and Technology (Ditjen DIKTI) Ministry of Education, Culture, Research, and Technology (Kemendikbud) also the Research and Community Service Bureau of the PLN Institute of Technology (LPPM ITPLN) for all assistances, both moral and material, so that this research can be completed.

REFERENCES

[1] E. Syam, *"Rancang Bangun Sistem Informasi Manajemen Data Mahasiswa dan Dosen Terintegrasi," It J. Res. Dev.*, Vol. 2, No. 2, pp. 45–51, 2018, doi: 10.25299/itjrd.2018.vol2(2).1220.

[2] Bahagia, D. Satria, and H. Ahmadian, *"Perancangan Sistem Informasi Manajemen Data Korban Bencana Berbasis Mobile Android," J. Manaj. dan Akunt.*, Vol. 3, No. 2, pp. 22–30, 2017.

[3] H. Bunyamin, C. P. Negara, F. T. Informasi, and U. K. Maranatha, *"Aplikasi Information Retrieval (IR) CATA Dengan Metode Generalized Vector Space Model," J. Inform.*, Vol. 4, No. 1, pp. 29–38, 2008.

[4] F. Amin and Purwatiningtyas, *"Rancang Bangun Information Retrieval System (IRS) Bahasa Jawa Ngoko pada Palintangan Penjebar Semangad Dengan Metode Vector Space Model (VSM)," J. Teknol. Inf. Din.*, Vol. 20, No. 1, pp. 25–35, 2015.

[5] M. N. Saadah, W. R. Atmagi, D. S. Rahayu, and A. Z. Arifin, *"Information Retrieval Of Text Document With Weighting TF-IDF and LCS," J. Comput. Sci. Inf.*, Vol. 6, No. 1, pp. 34–37, 2013.

[6] C. H S and M. K. Shenoy, *"Advanced Text Documents Information Retrieval System for Search Services," Cogent Eng.*, Vol. 7, No. 1, 2020, doi: 10.1080/23311916.2020.1856467.

[7] D. Danuri, *"Pencarian File Teks Berbasis Content Dengan Pencocokan String Menggunakan Algoritma Brute force," Sci. J. Informatics*, Vol. 3, No. 1, pp. 68–75, 2016, doi: 10.15294/sji.v3i1.6515.

[8] K. D. Aruleba, D. T. Akomolafe, and B. Afeni, *"A Full Text Retrieval System in a Digital Library Environment," Intell. Inf. Manag.*, Vol. 08, No. 01, pp. 1–8, 2016, doi: 10.4236/iim.2016.81001.

[9] G. Kowalski, *Information Retrieval Systems: Theory and Implementation*. 1999.

[10] A. Yusnita and Yunita, *"Penelusuran Katalog Perpustakaan pada Sma It Yabis Bontang Dengan Algoritma Boyer-Moore," Sebatik STMIK WICIDA*, pp. 15–21, 2018.

[11] S. M. Houston, *The Project Manager's Guide to Health Information Technology Implementation*, 2nd ed. LLC: Taylor & Francis Group, 2018.