

An Effective Fast Searching Algorithm for Internet Crawling Usage

Chia Zhen Hon, Nor Azhar Ahmad

Fakulti Sistem Komputer & Kejuruteraan Perisian, Universiti Malaysia Pahang, Lebuhraya Tun Razak, 26300 Gambang, Pahang, MALAYSIA
codychia96@gmail.com, nazhar@ump.edu.my

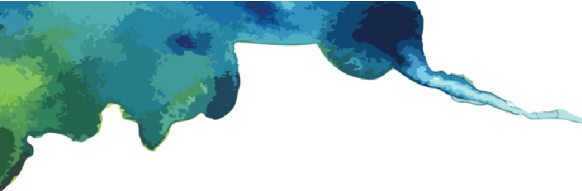
Highlights: The search algorithm is a crucial part in any internet applications. One of the crucial usages is internet crawling to find the best information using keywords. The searching algorithm has been one of the famous algorithm which is used throughout most of the existing software, such as Google, Facebook, Twitter. Therefore, an effective search algorithm is very crucial as most of the users are depending on it. This project will benefit to all users. The prototype has been developed using PHP, JavaScript enveloped in a Bootstrap web creator. It has been tested in eBook searching for UMP Library and produced success story. It could search faster in comparison to the other approach when crawling more than 10k eBook titles and content. Further development could be a combination of a faster method inside the proposed searching engine.

Key words: Search, Algorithm, Effective search

Introduction

The algorithm has been existed and been used in our daily life ever since the early civilization. Search algorithm itself is being commonly used and it is known as the universal problem –solving mechanism. Actually search algorithm is just an algorithm for searching an item (pattern) within a set of collection of data (text). Usually, it is a list of data that matches the predefined matching criteria which is the outcome of the search





form. This type of algorithm is used to look for more specific information from the available database. For instance, a search algorithm is applied when to search for a phrase in a document or search for a specific date or time. Some of the common text searching algorithm is Brute-Force algorithm (BF), Rabin-Karp algorithm (RK), Boyer-Moore algorithm (BM) and Boyer Moore Horspool algorithm (BMH).

This research is concerned about the pattern matching problems. It is applied to check on the matches between pattern matching and the text from the database itself. In the field such as computer science, bio- chemical engineering, literature and more, text matching processing is essential and important. To check whether the specific pattern is a substring of the test, an algorithm is used to compare a short text string with a longer string text.

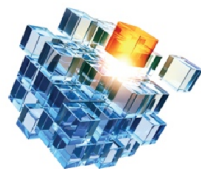
E-book Search Algorithm Development Framework

Prototype model exploration applied in the system. Each stage has been implemented to make sure the effective prototype development. **Figure 1** shows the project flow.



Figure 1: Development Framework

The model starts with Requirement Planning. In the RAD model, planning is a combination of planning and



analysis. Without the brief planning, the project cannot be arranged well. Hence, this stage is needed in the system in order to identify the algorithm selected which is carried out along this stage.

In this phase, the suitable title, the problem statement, objectives and scopes of the project are determined. Hence, the title for this project is Implementation of Boyer Moore Horspool Algorithm for Library e-book Searching. This problem is proposed due to amount of memory space needed by the computations. The second problem statement is the means of standard measures of a running time in order to search in the fastest way.

The objectives are to develop a searching function using Boyer Moore Horspool algorithms based on a keyword in e-book database, to provide an automatic search routine in the e-book software and evaluate performance of the proposed algorithm through character match percentage.

For the design phase, it is needed to collect data provided from the e-book database. This data need in order to be applied to the database in the Boyer Moore Horspool algorithm. Besides, this phase will involve in design for system prototype for BMH algorithm. In project interface design, it will include only one interface which is designed for searching algorithm interface where user need to input their search and display the related output.



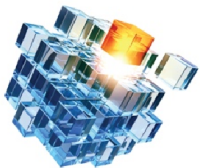


Result and Discussion

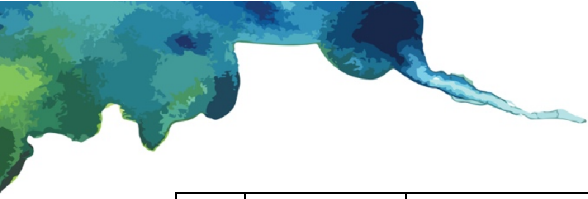
The character matching percentage is being used to evaluate performance of the proposed searching algorithm. As we can see, after experimental data was handled, we can conclude that the total average of the Boyer Moore Horspool algorithm is a strong match with a 78 % accuracy result. In order to evaluate the performance of purpose algorithm, a few experiments on data is conducted. In the experiment, 1000 data from different size is taken as a sample. Therefore, based on 1000 data, the data are divided into five datasets which will each handle 200 data. Twenty different keywords is selected and search in order to test the accuracy for the data. Based on the calculation, we will know that the algorithm is suitable use for searching or not. Below are the one of the dataset results gains from the testing.

Table 1: Testing Set

No	Keyword	Actual result	Expected result	Accuracy (%)
1	Blood	Blooded, warm-blooded, cold-blooded, bloodstream, blood	Blood	62
2	Weak	Weak, Weakness, weaken	Weak	50
3	Loss	Loss, Losses, Glossaries, glossary	Loss	40



4	Today	Today, today's	Day	75
5	Morning	Morning, morning's	Morning	78
6	God	Godmother, god, godforsaken	God	43
7	Knight	Knight, knightly, knighted	Knight	60
8	bottle	Bottle, bottles	bottle	86
9	car	Carousing, carriage, career, car	car	55
10	computer	Computer- based, computer, computers	computer	73
11	Apple	Apple, apples	apple	83
12	research	Research, researcher	search	72
13	human	Human, humans	human	83
14	mouse	Mouse, mousse, mouser	mouse	83
15	database	Database, databases	database	89
16	hot	Hot, hotel	hot	60
17	baby	Baby, babyish	baby	57
18	honey	Honey, honeys, honeymoon	honey	60
19	sword	Sword, swordfish	sword	55



20	airplane	Airplanes, airplane	airplane	89
21	shaking	Shaking	Shaking	100

TOTAL AVERAGE	(62 + 50 + 40 + 75 + 78 + 43 + 86 + 55 + 73 + 73 + 83 + 67 + 83 + 89 + 60 + 57 + 60 + 55 + 89 + 100) / 21 = 72 %
--------------------------	---

Advantages

1. Able to provide a real-time search routine in the system.
2. Ability to calculate the accuracy of the proposed searching algorithm through character match percentage.
3. Increase of knowledge among Malaysian towards search algorithm.

Acknowledgement

The researcher would like to thank UMP for giving the student's chance to carry out their research.

References

Reviewer-Kintali, S. (2015). Algorithms Unplugged by B. Vöcking, H. Alt, M. Dietzfelbinger, R. Reischuk, C. Scheideler, H. Vollmer, and D. Wagner: The Power of Algorithms by Giorgio Ausiello and Rossella Petreschi. ACM SIGACT News, 46(4), 14-16.

Eng, C. H. B. (2015). Efficient compression of large repetitive strings (Doctoral dissertation, RMIT University, Melbourne).

Cisłak, A. (2015). Full-text and Keyword Indexes for String Searching. arXiv preprint arXiv:1508.06610.

