Automated social media account identification using Simplified Brute Force

G. Harianto¹, E. B. Setiawan² and Y. R. Murti³

School of Computing, Telkom University, Telekomunikasi street 01, Terusan Buah Batu, Bandung 40257, Indonesia.

E-mail: ¹godvestroia@gmail.com, ²erwinbudisetiawan@telkomuniversity.ac.id, ³yuszaa@telkomuniversity.ac.id

Abstract. Social media is a media where millions of people can connect and obtain information through the Internet network. The features provided by social media has made it easier to exchange data on the internet. Social media has covered so many people who have more than one social media account. Day by day the number of social media users has increased and has reached a huge number plus everyone can have more than one account. This huge amount makes social media into a data warehouse that can be used to obtain information. By crawling data on multiple social media at once, the search process will be faster so that we can identify accounts on more than one social media at once. In this research, we will crawl to some social media at once using DOM Parser. After the data collected, we performed the matching using the simplified Brute Force algorithm with a time efficiency of 78.16% where Facebook received the highest 90% accuracy on unverified account identification and Instagram received the highest 80% accuracy for verified account identification. We apply crawling to obtain account information from several social media as well as directly. The crawling process uses the DOM structure to parse the content on the site and SQL to insert data into the database.

1. Introduction
Social media is a warehouse of information where everyone can get a lot of information. The availability of search engines on any social media also makes it easier for users to search someone's account. Crawling data is one of the data collection methods that download information from a website page using a software called web spider or web robot. Then the web spider will crawl on the intended web page and enter the data that the web permits into the database [1].

Some studies usually use the API to do the crawling of social media. For example, the research done to crawling Facebook, Instagram and Twitter [2,3,4]. The researchers used the API to download data from social media into the database. But in this study crawling on social media will not use the help of API to obtain data but using DOM Parser. The advantages of using DOM Parser is the ease of data retrieval where users simply search for tags, class or id of the elements of the website whose data will be entered into the database while the API user must have a token first and find out the keyword index of data elements to be taken. After the data are collected then the system will perform matching on the crawling results of some social media to estimate an account of each social media identified as the user being searched for the account. To match the result, we using the simplified Brute Force where this algorithm works based on the Brute-Force algorithm in general [9], but we do not match keywords to all data in the database. Matching keywords only performed on the name attribute to shorten the time needed during the matching process. This is done based on our analysis of databases where the keywords tend to have no compatibility with other attributes in the database other than name attributes.

2. Related work
Crawling on social media is the process of collecting data and information provided by social media to be processed as done on Facebook [2], Instagram [3], Twitter [4] and Google+ [5]. We can be crawling
to collect various data types such as user id, link profile, profile name, profile picture, biography, account type, status / post and much more depending on the data provided by the social media. The collected data can be used to perform analysis such as grouping of user types and categories [3], detecting cloned account [4], bot detection [6], detecting spam [7], detecting fake news [8] as well as other analyzes using data from social media. However, we have not found the use of social media crawling data to identify a person's account from multiple social media at once.

Therefore in this paper, we create social media account identifiers using the Brute Force approach where the process of crawling on search results in social media will run in real-time. Generally, any social media search engine will only display search results related to social media itself, but the crawling process in this study will directly crawling on three social media at once that are Facebook, Twitter and Instagram using PHP, SQL and DOM Parser help for take the data and put it into the database. Then the data will be divided into several tables according to the type of social media before being matched using Brute Force algorithm and displayed to the screen.

Brute Force algorithm is one of the algorithms applied to the search engine. The way it works is quite simple where the system will check the keywords on all the data to find its match [9]. However, the use of brute force algorithms in this paper will minimize the scope of matching in which the keyword is matched only with the name attribute to reduce time the matching process. The keyword used to match is the username entered on the search form page. The information that we need to enter on the search form page is the full name and account type (verified or unverified) and another detail like city of origin, city of residence and workplace used to do filtering on Facebook search results.

3. Proposed algorithm and implementation

Based on related work to be done step-step in doing social media calculations. The first step is to enter the user information to search then the information will be used to do the crawling (3.1) and after the data collected will be done using Brute Force algorithm (3.2) then will be displayed to the screen.

3.1. Crawling

One technique of collecting data from social media is crawling and assisted by API [2]. To use this API, we need a token and we need to know the keyword index of the attributes to be retrieved. Another case with crawling using DOM Parser, to crawling data only by finding out tags, class, or id of data to be entered into the database. In this paper crawling using DOM Parser will be applied to collect social media account data related to user name, user id, profile link and some user details that may be listed on his account. After the user enters the information in the search form then the information will be converted into a search URL where at this stage there is a process called encoding. Encoding serves to change the characters we enter into characters that can be read by the system so that search format can run correctly [10]. We do the encoding on the space character as shown in table 1.

| No | Social Media | Symbol | Encoding |
|----|--------------|--------|----------|
| 1  | Facebook     | Space  | %20      |
| 2  | Instagram    | Space  | +        |
| 3  | Twitter      | Space  | %20      |
After we get the data output of social media users we can minimize the scope of social media search by applying filtering that adds keywords about the details of the person to be searched based on the information entered on the search form. In figure 1 it is explained that when starting the filtering system will initialize the new keyword to be entered into the search engine then start the new crawling. After that the data will be entered into the database grouped according to the type of social media.

![Diagram](image)

**Figure 1.** Social media account identification system.

### 3.2. Matching with Brute Force Algorithm

The collected data will then be matched to the keywords. The matching process in this paper uses the Brute Force algorithm. But before starting the matching process, data from the database and also keywords will be converted to lowercase using case folding to avoid sensitive case error at matching. The matching process in this algorithm is simplified by minimizing the scope of matching in where the attributes that will be matched only the name attribute.

---

3

---
Brute Force algorithm in this paper is classified into four processes done on each social media account data table. In figure 2 we can see an overview of the matching process and the order employed by the Brute Force algorithm. The explanation of each process are described as follows,

- **Keyword initialization**: put the username entered into the search form that has been converted with case folding as a keyword.
- **Data calling**: the system will call data from the database and convert it using case folding repeatedly until found matches with keywords.
- **Matching**: checking the match between keywords and data called from the database until one of the Brute Force conditions is found.
- **Print identification result**: displays matching results from each social media to the screen.

### 4. Result and analysis

In this paper, we use 100 account data consisting of 50 verified accounts (identifying Facebook, Twitter and Instagram) and 50 unverified accounts (identifying Facebook, Twitter, Instagram and Google+). Because Google+ does not provide badges or anything specific for verified accounts, Google+ does not include in the identification of verified accounts. What deceives a verified account is an original account that reflects well-known figures or brands for boring forgery. Before starting the search using the program, we do a manual search about related user account profile that will be included. The process of calculating accuracy is done by matching the identification results obtained by the system with the results obtained manually obtained. If the results obtained by the system are the same as the manual search results will be assigned a value of "1" (true) and if the results obtained by the system does not match the search results manually it will be given a value of "0" (false).

When the identification process has not implemented preprocessing data, Facebook has the highest 90% accuracy on unverified account identification as shown in figure 3. Then in verified account identification, Instagram obtains the highest 80% accuracy as shown in figure 4.
Based on these results, we try to find the problems in account identification. After analyzing we found some causes of error are as follows,

- Incorrect matching due to case sensitive.
  The matching error referred in this case is when the username is placing an irregular capital letter that is not accordance with the name that should be written. So, we do case folding to overcome it and get a higher percentage of compatibility than before using case folding.
- Account protected.
Account protection is a feature provided by several social media to protect user data so that it cannot be seen by public people. It can also cause the protected account to not appear on search engine result.

- The name listed does not match the real name.
  Some people sometimes use nicknames that are not same or do not match with their real name. So, the matching process will state that the user does not match with the name specified in the keyword.
- Unverified account (for case of identification of verified account type).
  What is meant by a verified account is an account owned by people or brands that are known to many people. While unverified accounts are ordinary accounts that people use in general. Sometimes there are public figures, political figures or other famous names who do not register their account as a verified account. So, the matching process on verified account will state that the account is not match because the account status has not been verified.

From some of the above problems one that can be solved in this paper is the error caused by the sensitive case. This problem can be solved using one of the data preprocessing techniques called case folding. In figure 3 and figure 4 it can be seen that the accuracy level after case folding on verification of verified accounts and unverified accounts, but does not change the highest level of accuracy in the identification of verified and unverified accounts. For the identification of unverified accounts, Facebook gets a fairly high level of accuracy because the search engine of Facebook has a feature that is able to find the name of the person with the details to minimize the scope of the search.

| Table 2. Comparison of the average execution time between original and simplified Brute Force. |
|-----------------------------------------------------|
| Execution Time(s) | Efficiency(%) |
|-------------------|---------------|
| Original Brute Force | 0.010422945 |
| Simplified Brute Force | 0.002275944 |
|                    | 78.16         |

The matching process in this paper uses the simplified Brute Force algorithm where the matching process only focuses on the username attribute on the database. The result of the identification obtained remains the same if doing the overall matching using the original Brute Force, but the time obtained by the simplified Brute Force is faster as shown in table 2.

5. Conclusion

Brute Force algorithm is a method that can be implemented to identify social media accounts with data obtained using crawling based on user name and type. The identification process can be performed in real time because the process of collecting data using this crawler will directly access social media pages to collect data and update the database automatically. The importance of identifying this account is to make it easier for others who want to find someone's account without having to access social media pages one by one. Based on the results obtained we know that crawling using DOM Parser can be used to collect data from social media and by using the simplified Brute Force algorithm we get the highest accuracy of 90% on Facebook for unverified account identification then Instagram get the highest accuracy of 80% for verified account identification. The simplification of our brute-force algorithm succeeded in making a time efficiency of 78.16% of the time required by the original brute force with the same accuracy.
References

[1] Ahuja M and Bal J 2014 *International Journal of Computer Trends and Technology (IJCTT)* **13**

[2] Berger P, Hennig P, Klingbeil T, Kohnen M, Pade S and Meinel C 2013 *Proceedings of the International Conference on Information and Knowledge Engineering (IKE)* (Citeseer) p 1

[3] Hu Y, Manikonda L, Kambhampati S et al. 2014 *Icwsm*

[4] Hudson B, Gurajala S and Hu W 2013 *Flock: Crawling the Twitter Social Graph*

[5] Russel M A 2013 *Mining the Social Web: Data Mining Facebook, Twitter, LinkedIn, Google+, GitHub, and More* (O'Reilly Media, Inc.)

[6] Dickerson J P, Kagan V and Subrahmanian V 2014 *Proceedings of the 2014 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining* (IEEE Press) pp 620–627

[7] Stringhini G, Kruegel C and Vigna G 2010 *Proceedings of the 26th annual computer security applications conference* (ACM) pp 1–9

[8] Allcott H and Gentzkow M 2017 *Journal of Economic Perspectives* **31** 211–36

[9] Abdeen R A 2011 *An Algorithm for String Searching Based on Brute-Force Algorithm*

[10] Agarwal P, Thakur R N and Gavini A K 2014 *Policy driven fine grain url encoding mechanism for ssl vpn clientless access* (Google Patents)