Focus Web Crawler on Drug Herbs Interaction Patterns

Fatini Nadhirah Mohd Nain[1], Nurul Hashimah Ahamed Hassain Malim*[1], J. Joshua Thomas[2] and Mei Lan Tan[3]
Email: fatininadhirahnain@student.usm.my, nurulhashimah@usm.my, jjoshua@kdupg.edu.my, tanml@usm.my
* Corresponding author
[1]School of Computer Sciences, Universiti Sains Malaysia, Gelugor 11800, Pulau Pinang, Malaysia.
[2]Department of Computing, UOW Malaysia KDU Penang University College, Georgetown 10400, Pulau Pinang, Malaysia.
[3]School of Pharmaceutical Sciences, Universiti Sains Malaysia, Gelugor 11800, Pulau Pinang, Malaysia.

Keywords: Drug-herb interactions, Focus Web crawler, Breadth-First Search (BFS), PageRank.

Received: March 14, 2021

The types of pharmaceutical products include cosmetics and drugs. Some of the pharmaceutical products comprise a mix of drugs and herbs without considering their interaction effects. Drug-herb interactions (DHIs) refer to the interactions between conventional drugs and herb medicines. However, the available information on DHIs is scattered because it has heterogeneous databases and website resources, apart from some of the paid or subscribed databases. Easy access to information on DHIs would allow researchers to explore more. Therefore, this study proposes improvements in the focus web crawler to collect DHIs information from the heterogeneous resources on the Internet, present priority levels of a resource link through anchor text and URLs, and traversing the link with the aid of depth. The improved focused crawler was tested on two algorithms namely the Breadth-First Search (BFS) and PageRank. Information of DHIs crawled 4,744 herbals from the focus web crawler. The accuracy values for Chinese Med Digital Projects and MedlinePlus were 98% for PageRank and 71% for BFS. Additionally, a focused web crawler may gather more relevant web pages in the same amount of time as a wide crawler. Hence, the proposed crawler may successfully gather DHIs on the web in response to the user queries.

Povzetek: Razvit je nov algoritem za preiskovanje spleta za iskanje vzorcev medsebojne odvisnosti zdravil.

1 Introduction

Despite the advancements in modern medicine, most people still use herbals to cure their illnesses. In the 17th century, many countries practiced herbal medicine based on their traditional knowledge of a plant that was used by the local communities and was passed down from one generation to another. Now, many of the products have mixed conventional drugs with herbals. Therefore, it will lead to a large gap in increasing the number of chemicals consisting of primary and secondary metabolites of the active substance using single pharmacology that contributes to the effects of either moderate, resisting, etc.[1].

Contrary to the popular believe, the side effects of herbal medicines are greater compared to conventional drugs, regardless of the generalization of 'natural means safe' due to the lack of appropriate quality control, inadequate labeling, and lack of appropriate patient information. [2,3]. Of lately, various plant-derived products are being incorporated into cosmetics and natural products. These products contain active phytochemicals in a range of unstandardized preparations (i.e. tablets, capsules, sachets, or pills). Some of the sports drinks, supplements and energy bars contain ingredients that have been mixed with herbs and medicines. The effect of a mixture of homemade medicines used where the patient begins to manage it on their own without supervision and advice from a doctor, therefore, offer an increase to the rate of drug-herbal interactions (DHIs) [4]. DHIs refer to the interactions between conventional drugs and herbal medicines [5]. DHIs commonly occur during the pharmacokinetic and pharmacodynamic interactions in prescribed drugs, dietary supplements, or a small portion of food items [6]. In oncology studies, pharmacokinetics interactions can metabolize enzymes like cytochrome P450 (CYP) and P-glycoprotein (P-gp), while pharmacodynamics interactions refer to drugs that influence each other's effects directly. However, excessive DHIs can lead to unexpected Adverse Drug Reactions (ADRs). For instance, a herb that interacts with cisplatin to cure cancer is the Black Cohosh [7].

Information on DHIs can be obtained from the World Wide Web (WWW). However, medical professionals like doctors, pharmacists, medical researchers, and others require an automatic solution to gather the information from articles, databases, and other websites. Therefore, a web crawler is proposed as a solution to accumulate all the information. Thus, the focus web crawler seeks pages that satisfy the relevant information related to the search topics [8,9]. The focused crawler retrieves the maximum number of relevant pages simultaneously and transverse the minimum number of irrelevant pages on the website [10–12]. The focus web crawler also indexes the website entry where the users can

send the index via query and provide the results of the website that matches with the query.

Information on DHIs are scattered since many databases are available to store the information, including Medical Literature Analysis and Retrieval System Online (MEDLINE), and PubMed [13]. A majority of the healthcare professionals prefer to search for research and case reports on DHIs in databases like MEDLINE, PUBMED, EMBASE, and COCHRANE libraries using the following search terms or combinations thereof: "drug–herb interaction," "herb-drug interaction," "interaction," "cytochrome P450," "plant," "extract," "medicinal," "concurrent administration," and "herbal and orthodox medicines." Appropriate search terms were used to represent numerous medicinal herbs used in Africa, America, Asia, Europe, and Australia. This study searched and compiled interaction reports between orthodox medications and their mechanisms of action. The searches were not restricted by publication date or location, but only considered publications in the English language [14,15]. PubMed and MEDLINE contain journals and articles on experiments and studies conducted by medical professions, while the other resource websites such as WebMD, HerbMed, and Natural Medicines Comprehensive Database provided information related to DHIs. Meanwhile, some journal articles like PubMed and MEDLINE require an account subscription to be able to download and read the papers. Therefore, medical professionals face limited time and access to information on DHIs from various websites, as some websites require a purchased subscription.

There are various types of supplements and pharmaceutical companies in Malaysia that manufacture supplements by mixing drugs and herbs regardless of the interactions and ADRs, prescribed and approved by the Drug Control Authority (DCA) and National Pharmaceutical Regulatory Agency (NPRA). Whereas, some people consume drugs and herbs as alternative treatments without consulting their doctors. People are unaware that they can obtain information on DHIs from websites and databases. Hence, a web crawler is proposed in this study to help extract relevant information on DHIs. This study allows readers and researchers from different backgrounds to explore more on the DHIs and web crawlers. Moreover, the web crawler can also download and extract information efficiently and faster. Therefore, a web crawler is the best solution to be implemented in the medical field. This study aims to perform web crawling from several herbal medicine websites related to DHIs and to evaluate web crawler algorithms for DHIs.

The most crucial part of a focused crawler is the selection of the URLs. The primary goal of an effectively focused crawler is to locate relevant web pages and guide them to those pages. Here, classification is widely accepted as the most common method for determining relevant and irrelevant pages. However, classification is not used when DHI websites and databases mostly contain DHI related information and even related URLs also require an indexing algorithm to sort the most preferred websites and databases. Therefore, a focus web crawler approach and indexing algorithms were proposed in this study to identify the most important websites and databases. Page URLs were divided into two categories by indexing algorithms namely primary websites and hyperlinks. To improve the indexing algorithms, this study set different depths for various main pages based on their content ability. The greater the number of pages and the higher weights for the main website can be achieved with a higher number of depths. Meanwhile, newly improved pseudocode was developed by indexing algorithms to improve the algorithm convergence. The performance of focused crawling is directly influenced by the method used to select URLs. This strategy allows the crawlers to find relevant web pages. This study picked sites from the unvisited list, and sorted them in an ascending manner relevant to the given topic of the page being visited. When determining the link weights during crawling, the current page's anchor text, context, and URL string are all considered [16–18]. The most frequently accessed links' features indicate the user's current location to assess their trends and patterns towards a site. After dividing the web page into sections, we evaluated each section as a single content block. Previously unvisited URLs were also extracted and added to the frontier where applicable, with the weight assigned based on its importance. Then, all of the content block links were removed.

This study specifically seeks to make three key contributions. Firstly, the study assessed the topic of ADRs, DHIs and Web Crawler, while prior studies on web crawling largely focused on its benefits for healthcare professionals. It is very important to study the associated costs thoroughly before initiating research. This study focused on the conflicts and pressures created by DHIs which could impact public health. It also focused on the adaptation of web crawlers in reducing costs and improving the efficiency of the web crawlers in DHIs. Therefore, issues and phenomena unique to DHIs are focused on in this study along with their interactions between each other and the outcomes. It also assessed the existence of side effects or ADRs, the implementation of web crawlers in DHIs, sorting of the heterogenous websites and databases, highlighting our extant understanding of DHIs, web crawler processes and outcomes derived from the ADRs and web crawler literature. Secondly, this study adds to the understanding of role theory especially regarding focused web crawling because it is the latest technology adopted by researchers. Thirdly, we contributed to the growing evidence of selecting the best indexing algorithms by comparing their performances. Although previous literature investigated other moderators for DHIs, they often focused on DHIs' research methods by obtaining random information manually. Instead, this study focused on DHIs, an aspect of self-regulation by medical professionals that is easily accessible, therefore could reduce the cost and time spent for obtaining information about DHIs.

The rest of the study is organized as follows: Section 2 describes related studies on ADRs, DHIs, and web crawler methods. Section 3 elaborates the research methodology in terms of the implementation and

experimental design, while Section 4 discusses the analyses outcomes. Lastly, Section 5 concludes this study.

2 Background and Related Work

ADRs originate from the term Adverse Events. Adverse Events are defined by National Care Institute as unexpected symptom that occurs during treatment or therapy (Figure 1) [19]. Adverse Events that occur when a patient consumes a drug excessively, might be defined as ADRs. ADRs refer to the negative or harmful responses due to medication [20]. ADRs which occur due to excessive consumption of conventional medicine with conventional medicine or herbal medicine with conventional medicine is also known as overdose. ADRs could affect adults, children and infants too. The World Health Organization (WHO) defines medication errors as failed treatments that could harm the patients [21,22]. Medication errors could occur during the medication process, choosing a medicine, errors in writing the prescriptions, using the wrong formula by the manufacturer, etc. [23–26].

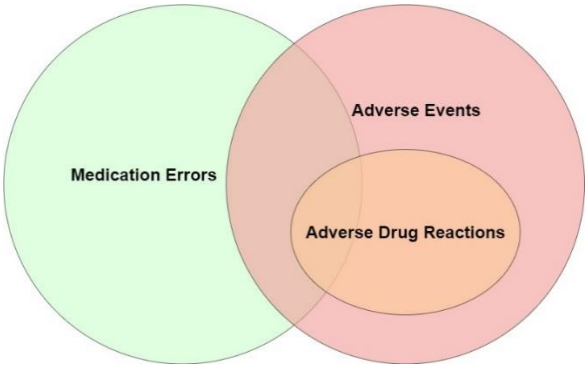

Figure 1: Relationship between medication errors, adverse events, and ADRs.

Several recent studies on ADRs were published by researchers from the United States, Malaysia, and other developed countries. Most of the ADR reports were obtained from MEDLINE, PubMed, etc. [27]. However, the studies revealed that ADRs do not occur only due to the overdose of drugs but also due to discontinuing drug therapy. Another study assessed the pervasiveness of ADEs and reported that 5.1% of the clinic affirmations were expected to be ADRs [28]. They reported that 5.3% of the admissions were caused by ADRs [28]. The percentage of patients who got admitted to the hospital increased due to ADRs [29].

In Malaysia, ADR reports are submitted to the Malaysian Adverse Drug Reactions Advisory Committee (MADRAC) [30,31]. The study also grouped causality into five, namely Certain, Probable, Possible, Unlikely, and Unclassifiable, to classify the ADR reports from MADRAC.

For some medications, the DHIs could also lead to ADRs, especially with herbal products. According to a study, 39 out of 46 herbal medical products interacted with conventional drugs [32]. For instance, Perforate St John's-wort is known as Hypericum perforatum, a prevalent natural item highlighted for its administration as an anti-depression have also been broadly considered for pharmacokinetic DHIs. Hypericum perforatum can be purchased over the counter and is consumed by patients with different pathologies. Hence, it is discouraged. In clinical preliminaries, Hypericum perforatum is known to better stimulate movement compared to fake treatment with a lower dose. Even though the concentrates of Hypericum perforatum (a leafy herb that come from the Hypericaceae family) contains a few phytochemicals and hyperforin, the dynamic energizer specialist is involved in interceding DHIs [33]. Compared to different phytochemicals, hyperforin is bioavailable in humans, where the amount should be halved every 12 hours. It should also take into account the aggregation that deserves attention in many parts of the human body. Hyperforin enacts the Pregnane X Receptor (PXR), an atomic receptor found solely in the liver and digestive tract. Having demonstrated an EC50 estimation of 23 nM with a 380 nM top plasma focus, and a 200 nM enduring state plasma levels reachable in people ingesting the standard routine of 900 mg hypericum perforatum (typically in three separated dosages), hyperforin activates PXR under the fixation to be achieved in human plasma. Similarly, different preclinical investigations have revealed the capacity of *Ginkgo biloba* concentrates to repress human metabolic proteins including CYP1A2, CYP2C9, CYP2E1, and CYP3A4 [34,35]. Most modern medicines contain expected effects including their toxicity when interacting with herbs. Drug toxicity can occur when someone consumes multiple drugs at a time. It specifically occurs when the dose consumed by the individual exceeds the prescribed dose (intentionally or accidentally).

Table 1: Effects of toxicity when interacting with DHIs

No.	Effects of toxicity when interacting with DHIs.
1.	Dizziness
2.	Low/high blood pressure
3.	Low/high sugar level
4.	Eczema
5.	Bruise

Table 1 lists some examples of toxic effects that may occur through the interaction of drugs and herbs. Patients who are in the age range of 33 to 78 years might experience bleeding while consuming certain drugs (i.e. aspirin, warfarin, acetaminophen, and ergotamine-caffeine) with the *Ginkgo biloba*. Most of them experienced major or minor bleeding, while some died due to massive cerebral haemorrhage. Meanwhile, patients who consume products containing Vitamin E with excessive *Ginkgo biloba* could experience an increased rate of platelet activity. The related side effects include blurred vision, headache, and dizziness.

Focus web crawler which is also known as topical web crawler is a technique that only collects web pages that satisfy specific properties. Focus web crawler can analyze the crawler's boundary to determine the most relevant links and avoid unnecessary or irrelevant regions

of the web. Focus web crawler aims to find pages that satisfy the relevant information related to particular topics [8,9]. The focused crawler retrieves a maximum number of relevant pages simultaneously and excludes the minimum number of irrelevant pages on the web [10,11]. Previous studies on focus web crawlers used keyword matching or regular expression matching. The focus web crawler was introduced using the Fish Search algorithm. Fish Search algorithm is an algorithm that stimulates crawling using a group of fish that migrates with the web [36]. Each of the crawled URLs is compared to a fish because survivability relies on visited page pertinence and remote server speed. Page importance is assessed utilizing a double order by employing straightforward catchphrase or customary articulation matches. The fish dies if it navigates through a specific number of unrelated pages. However, studies have tried to improve oriented crawlers to effectively collect related information. The depth parameter could restrict the crawler to not visit a site, which is not important for the searching fish. The relationship between URLs were given importance or priority value based on the similarity found on Shark Search. The priority value is estimated based on the degree of similarity. Although most studies attempt to improve targeted crawlers to gather related information effectively, the parameter depth will limit the visitor to not visit a site that is not meant for searching fish. Based on the similarities identified on Shark Search, the relationships are assigned through the significance or priority status. The target interest is determined by the degree of similarity. Hence, a keyword-focused crawler was proposed by Agre et al. [37]. The study implemented an approach that dealt with domain ontology to find the most relevant pages according to the user requirements. Domain ontology is used to filter out the repository information. The advantages of keyword web crawler over traditional web crawler are that it works intelligently and efficiently without requiring relevant feedback. Consequently, the crawler workload is reduced. On the other hand, a focus crawler system for automation webpage classification was proposed by Goyal [38]. This study aims to determine whether the web pages consist of information of Indian original faculty working in foreign universities. They introduced the automation webpage of the Indian faculties through methods of URL filtering using feature extraction, a genetic algorithm (GA)-based classification. A mutation algorithm was employed to calculate the number of feature extraction. NetBeans IDE 6.9.0 was chosen as the software to execute the implementations. The URLs were selected from the faculties and university websites using keywords. The tags and terms used as the feature extraction were named as chromosomes. The genetic algorithm-based classifier that was implemented used six steps, which are coding, generation of the initial population, evaluation of initial population, selection, crossover, and mutation. Their performance was analyzed in terms of document matrix (ranging from 0 to 1). A precision score higher than 0.8 considers a page to be relevant. An efficient focused web crawler searches for medical plants and relevant diseases using several algorithms such as Naïve Bayes Classifier Algorithm, Decision Tree Algorithm and Multilayer Perceptron [12]. Naïve Bayes classifier was employed to determine whether the current web page was relevant or is not related to the medicinal plant information. This method was proven to perform better. The three types of lexical features, which are title-feature, meta-description, and anchor text, were also implemented for the Naïve Bayes classification. Moreover, a simple decision tree algorithm was developed to determine the relevancy of the medicinal plant URLs. Three different techniques were applied and analyzed ("yes" represents related medicinal plant and "no" represents not related medicinal plant) for each of the medicinal plant URL. The accuracy of Naïve Bayes produced 90% accuracy compared to the other algorithms.

Meanwhile, another study proposed a technique called keyword focused web crawler [39]. This study aimed to improve the performance of web crawler by exploring in depth of the relevant web to the topic. This approach was proposed to extract keywords based on URLs or criteria regarding Indonesian recipes to obtain the best search. This scheme only downloaded URLs which contained Indonesian recipes from the searched keywords. They also used some metrics such as link analysis algorithm including Breadth First Search (BFS) and other URL prioritizing techniques to rank the URLs. The technique did not find the relevant web pages through any other branches as the parent node was related to "milk tea recipe" and "tilapia recipe". The results indicated that the resultant data was much higher than the information path which contained the word "fried chicken recipe".

Another study proposed the essentials of the pre-processing task in social network user behavior [40]. This study aimed to analyze the user's structural behaviour by implementing network link-based properties only. This study employed the BFS algorithm to traverse a range of nodes of the entire social network, where the vertices were put into a specific database format. The result indicated that the BFS successfully sorted the links according to their priority. Meanwhile, a different study employed a novel edge-based parallel algorithm based on the Shiloach-Vishkin (SV) approach for distributed memory systems [41]. This study aimed to reduce the data volume and balance the load as the iterations progress. They implemented a Hybrid Approach, consisting of Parallel SV and Parallel BFS. The graph nodes of parallel BFS were grouped using power-law degree distribution before being implemented into the parallel SV algorithm. This study also implemented Edison, a Cray XC30 machine. The speedup of the machine was measured. The results indicated that the speedup achieved a maximum speed up by 8 times higher than the default value of 16. In another study, a measurement graph of Swarm social, an Online Social Networks (OSN) application on the mobile phone was proposed [42]. The purpose of this study was to provide a comprehensive view of a mainstream OSN that consists of tens of millions of nodes. They created the social graph for Swarm, calculated the key graph metrics, clustering coefficient, assortative, PageRank, connected components and communities. They also implemented BFS in this study to queue the selected user and their

profile information. Moreover, they also implemented Metropolis-Hastings Random Walk (MHRW) and BFS with different subgraphs. The tool used for the implementation was the C++ programming language. Case studies were also performed by sampling the 1%, 5% and 10% of all nodes to calculate the mean and variance. The results indicated that the number mean and variance for BFS was higher than that of MHRW.

Authors in a different study improved the network models design by implementing a local PageRank algorithm [43]. They aimed to measure the influence of a single article regardless of the specialities of the field. They modified the PageRank algorithm by defining the value of $\lambda = 0.1, 0.15, 0.2$, and 0.25. This algorithm was used to rank the co-citation graphs in scientometrics. In co-citation networks, nodes represent articles and the edges represent the citation of articles. Then, the score of the PageRank is computed for each of the nodes. A score of more than 0.8 indicated that the articles ranked by PageRank were relevant. Similarly, another study reviewed a PageRank algorithm [44]. In the study, PageRank searched the web pages based on inbound and outbound hyperlinks.

Another study assessed the improvements of weighted PageRank [45]. This study aimed to determine the ranking popularity of the pages based on the user's usage trends and browsing behavior. They implemented the weighted PageRank based on the links visited. This algorithm assigned a higher-ranking value to the outbound links for the most node visits due to higher popularity compared to inbound links. Hence, they introduced three methods, weight calculator, relevance calculator, and weighted PageRank algorithm based on content and link visits (WPRCLV) calculator in the study. The result demonstrated the efficiency of PageRank in ranking the relevant pages. Whereas, a novel approach to finding topical authorities on Twitter named FAME was proposed by another group of researchers [46]. This study implemented personalized PageRank to deploy a variant query-dependent on FAME. This study chose a suitable feature such as Twitter users' relationships to perform PageRank. The experiment exhibited the improvement of FAME for the authorities from Twitter.

As such, many studies attempted to improve oriented crawlers for the effective collection of related information. However, the depth parameter may restrict the crawler to not visit a site that is not important for searching fish. The URLs can be given the importance or priority value based on the similarity found on Shark Search. The priority value can be estimated based on the degree of similarity.

Based on the literature, the focus web crawler, BFS and PageRank are active research topics in gathering crawling the text information. Since DHIs can lead to pros and cons to our body and health, people need to know relevant information regarding the DHIs. The Internet is a very powerful tool because it is the main element of information growth and dissemination. It is used by billions of people every day from around the world for opportunities and information by people from all walks of life, especially researchers. There are various databases and websites on DHIs available on WWW. This study employed the focus web crawler as the suitable type of web crawler because it only crawls information regarding a specific topic. Tin line with a previous study [39], this study also implemented the addition of an extension to the web-crawling algorithm called PageRank. PageRank was chosen for this study because it was described to be able to rank relevant information, produce more accurate results, and take less time to execute the program.

3 Research method

The focus web crawler was enhanced to search for information on DHIs. This study aimed to ease the algorithms to index the URLs. As such, BFS and PageRank indexing methods were employed to index the URLs according to their priority. Figure 2 illustrates the proposed methodology of the efficient focused web crawler.

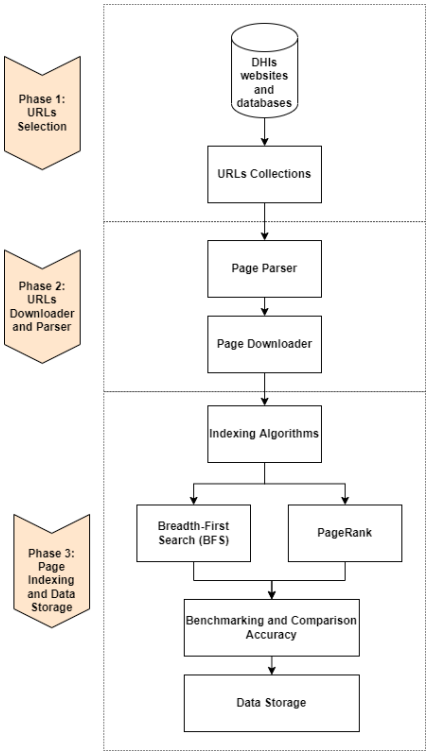

Figure 2: Proposed methodology of focus web crawler.

3.1 URLs selection and collection

There are various websites and databases holding information on DHIs. In phase 1, 5 different published websites or databases on the internet were randomly checked one by one whether or not they contained information on DHIs through the Google search engine (Table 2). Medical professionals commonly use most of these websites to retrieve information on DHIs. Some of the websites listed in Table 2 have a list of databases related to DHI and not related to DHIs. URLs related to DHIs will be collected from the main websites and databases.

Table 2: Lists of websites and databases.

No.	Name of Websites and Databases

1.	Medline Plus Database
2.	Western Botanical Medicine
3.	Global Information Hub on Integrated Medicine (GlobinMed)
4.	Chinese Med Digital Projects
5.	Countway Library of Medicine

3.2 Page parsers

Upon completing page collection, the HTML tags parsing the URLs were fetched from the page collections, followed by a page downloader to extract relevant information before storing the contents of those pages into the disk. DHI information is submitted to crawler downloader that containing the name, type of interaction etc. by a collection of pages used for data download, indexing and storage processes. The page parsers also submitted the information to determine the BFS and PageRank of the last crawled page. Page parsers information index the URL to check whether the URLs have enough crawl information based on its priority.

3.3 Page downloader

The page downloader fetches the URLs and puts them in the URLs queue to download the corresponding relevant pages from the web. The page downloader contains a domain to download the relevant pages. This domain is used to send the domain request and proceed with downloaders. The domain needs to set a timeout to ensure that it does not take too much time to read large pages or wait for the response of web servers. Robot Exclusion Protocol is an important step that needs to be considered for crawl page files because it provides a mechanism to the webserver. Thus, the webserver administrator can determine which pages cannot be accessed by the web crawlers. Meanwhile, the crawler used to exclude robots from a server is called robot Exclusion Protocol. This method creates a file on the server, where, this created page file must be accessible via local URLs or "robots.txt". A crawler can only check whether the pages can be downloaded or not with the approval of robots.txt. Figure 3 displays the examples of crawler.txt implemented in a web crawler. This example indicated that crawlers of other pages and public files are allowed to specify the address of the folder:/other/public/folder to facilitate the crawler's search to crawl relevant information. The page files contain cache to increase the efficiency of crawling. Therefore, it can avoid re-duplicating the page files when downloading the main pages from the same server.

```
User-agent: *
Disallow: /private/
Disallow: /confidential/
Disallow: /other/
Allow: /other/public/
```

Figure 3: Examples of crawler.txt

3.4 Page indexing and data storage

In this phase, the URLs are indexed by implementing two algorithms (BFS and PageRank) and data storage from the page extraction stage.

3.4.1 Indexing algorithms

3.4.1.1 BFS

BFS uses the frontier as a First in First Out (FIFO) queue in which the URL collection was arranged in the order they were encountered. When the FIFO queue is full with URL collection, the crawler added only one link from a crawled page. The BFS crawling method is illustrated in Figure 4. The pseudocode of BFS is summarized in Table 3 [47].

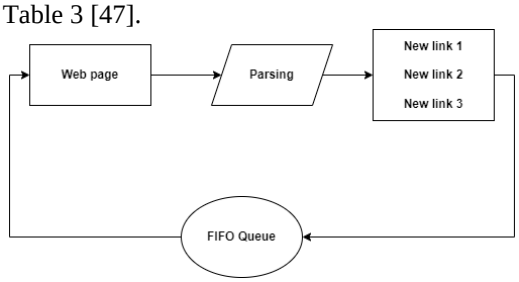

Figure 4: BFS crawling methods.

Table 3: BFS pseudocode.

PSEUDOCODE 1: INDEXING THE URLS USING BFS

```
Input: get_source
Output: links, urls, rank
if depth=4
    return
    print source, depth
    page_source = get_source(source)
    links = Set(findId(page_source))
else
    print 'some error encountered'
    return
end if
Repeat for link in links:
    if link not in urls:
        urls = urls ([link])
    end if
end for
Repeat for link in urls:
    Rank= (link,depth+1)
end for
Print output.
end
```

3.4.1.2 PageRank

Table 4: PageRank pseudocode for focus web crawler

PSEUDOCODE 2: INDEXING THE URLS USING PAGERANK ALGORITHM FOR FOCUS WEB CRAWLER	
	Input: urls, pages
	Output: hyperlinks, rank
1	initialize array of urls
2	initialize pages in integer
3	fetch hyperlinks from urls
4	**if** pages >0
5	Fetch hyperlinks from urls
6	**Repeat for** hyperlinks in urls
7	**if** hyperlinks exists
8	put hyperlinks in queue for rank
9	rank the number of hyperlinks
10	**end if**
11	**end for**
12	**end if**
13	print output
14	**end**

Besides BFS, PageRank algorithm is popularly used to index websites and to determine a page's relevance. The PageRank algorithm was introduced by the founders of Google, Brin and Page. PageRank uses probability distribution to represent the user's behavior [48]. PageRank can be calculated for collections of URL pages of different sizes. PageRank needs several passes or also known as "iterations" via the collection to adjust the approximate PageRank values to reflect the accurate theoretical value. The PageRank in the web crawler is calculated as a sum of the PageRank of all the pages that are linked to each other divided by the number of links on each of those pages. Equation 1 represents the formula for PageRank [48]. The pseudocode for PageRank is highlighted in Table 4. [47].

$$PR(A) = (1-d) + d(\frac{PR(T1)}{C(T1)}) + \cdots + PR(\frac{Tn}{C(Tn)})$$

Equation 1:PageRank equation.

Where

- PR(A) is the PageRank of page A (main page).
- PR(T1) is the PageRank of pages T1 which is linked to page A (child page).
- C(T1) is the number of outside links from page T1.
- d is a damping factor with the range 0 < d < 1, and is usually set to 0.85.

3.4.2 Data storage

Data storage is also an important process for search engines for future use. There are two types of data storage for crawled data namely memory-based and disk-based storage. Once the page extraction and page parsers are conducted, the data storage is performed to store all the extracted information from the websites on the disk. There are several ways to store the data including the non-relational databases depending on the structures of data, JavaScript Object Notation (JSON) files, Comma-separated values (CSV) files or Extensible Mark-up Language (XML) files. While data that consists of DHIs information (herbal name, herbal URLs, herbal description, herbal interactions, and the levelness of DHIs) are stored on disk in JSON file format for future references and usages. Each herbal name in the databases and websites contain the levelness of interaction, for instance, American ginseng interacts majorly with warfarin, hence, needs to be consumed with caution. This American ginseng data will be collected and stored in a database. The data containing URLs along with the rank numbers computed by BFS and PageRank are stored in SQLite for visualization purposes.

3.4.3 Evaluation

The quality of the system was evaluated based on certain criteria, that is accuracy. Accuracy is measured between two indexing algorithms that contain information of DHIs based on their priority and effectiveness. Accuracy determines the best algorithm for this study. The number of pages that contain related information about the interaction of drugs and herbs along with the number of pages that contain relevant and irrelevant DHIs information is required to calculate the accuracy of the indexing algorithms that has been indexed. Accuracy for each indexing algorithm indicates the number of downloaded DHI websites and databases apart from the number of relevant websites for each keyword. Then, the accuracy of all crawlers were compared to each other for all keywords. Accuracy indicates the efficiency of the web crawler in crawling the pages. The calculation of the number of pages retrieved from the main website and the number of pages retrieved from hyperlinks must be done first before calculating the accuracy of the indexing algorithms. Equation 2 represents the formula for accuracy.

$$Accuracy = \frac{number\ of\ related\ pages}{number\ of\ retrieved\ pages}$$

Equation 2: Accuracy equation.

4 Experimental analysis and results

In the proposed method, the DHIs thesaurus information was used for query expansion to induce more web content associated with the user query. The herbal dataset was extracted from an authorized website, downloaded, indexed, and stored in a structured format.

This medicinal plant database consisted of information on 4,744 herbals (Figure 5) from 24 websites and databases (Table 5). The effectiveness of the focus web crawler for DHIs was compared using two algorithms, BFS and PageRank.

Table 5: Herbals extracted from the DHIs databases.

No.	Names of DHIs databases	Number of herbals extracted that are related to DHIs
1.	MedlinePlus.	167
2.	National Center for Complementary and Integrative Health.	52
3.	Chinese Herbal Medicine Database.	408
4.	Western Herbs.	79
5.	Medicinal Herbs & Plant Database (Consumers)	79
6.	South Africa Herbs.	14
7.	Ayurveda Herbs.	25
8.	Native American Herbs.	31
9.	Essential Oils.	74
10.	Alternative Nature Online Herbal.	68
11.	Chinese Medicine Specimen Database.	859
12.	Medicinal Plant Images Database.	1159
13.	Chinese Medicinal Material Images Database.	420
14.	A Modern Herbs.	44
15.	The Raintree Tropical Plant Database.	180
16.	Longwood Herbal Tasks Force.	71
17.	Memorial Sloan Kettering Cancer Centre.	274
18.	HerbMed Pro	130
19.	HerbClip Online.	6
20.	Healthy Ingredients.	107
21.	The Commission E Monographs	120
22.	Herbal Medicine: Expanded Commission E.	107
23.	South Central America Herbs	100
24.	Medicinal Herbs and Plant Database.	200
	Total:	4,744

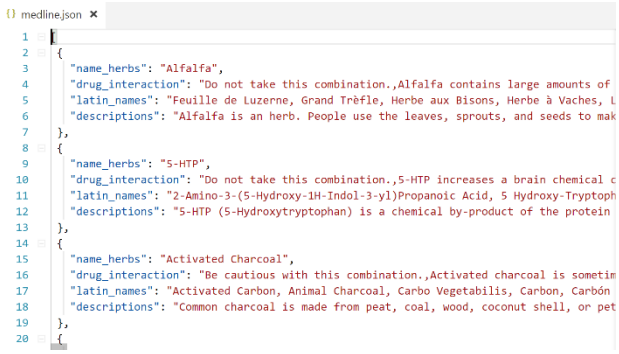

Figure 5: Sample of medicinal plant database consisting of herbal information.

Table 6: Lists of keywords consisting of main element biological transport and biological action.

List of Main Keywords (Keywords 1)
Drug-herb interactions
Drug-food interactions
Botanical medicine
Herbal medicine
Plant medicine
Traditional medicine
Alternative medicine

List of Biological Transport Keywords (Keywords 2)
P450 Cytochromes
Organic anionic transporters
Organic cationic transporters
P-glycoprotein
Drug transporters
Organic anion transporting polypeptide
ABC: ATP binding cassette transporter superfamily
SLC: solute-linked carrier transporter family
MDR1: multi-drug resistance
BCRP: breast cancer resistance protein

List of Action Keywords (Keywords 3)
Substrate, inhibitor, inducer
extracts, bioactive compounds

Keywords 1 contained main elements to perform user query, while keywords 2 consisted of biological transporter and keywords 3 contained action keywords to perform as a 'bridge' to link keywords 1 and keywords 2. For instance, Drug-herb interaction inhibitor P450 Cytochromes as tabulated in Table 6. Each keyword underwent a stemming process, to ease the root words and synonym words, for instance, Drug-herb interaction inhibits P450 Cytochrome. Logical rules, AND and OR were also implemented to crawl the exact DHIs, for instance, Drug-herb interaction AND inhibit AND P450 Cytochrome.

The working of BFS is very simple. It operates on the first come first serve basis. It starts with the https://medlineplus.gov/druginfo/herb_All.html

hyperlinks and is printed as 0 as it is the parent node. https://www.nlm.nih.gov/m, https://nccih.nih.gov/health/echinacea/ataglance.htm, and other hyperlinks are the child nodes. BFS queues all the hyperlinks and continues to update the hyperlinks until there are no more hyperlinks inside the website.

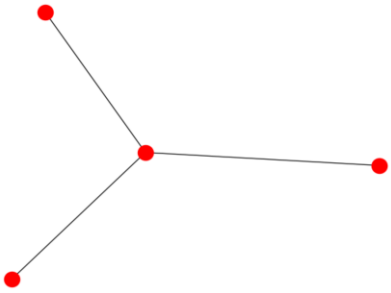

Figure 6: BFS visualization.

Based on Figure 6, BFS is visualized based on the results of the rank. For instance, https://medlineplus.gov/druginfo/herb_All.html is the parent node (right side). The hyperlinks in https://medlineplus.gov/druginfo/herb_All.html are the child nodes (middle and left side). The only limitation of BFS is that it cannot draw nodes if there are more than one similar ranking when the hyperlinks are different.

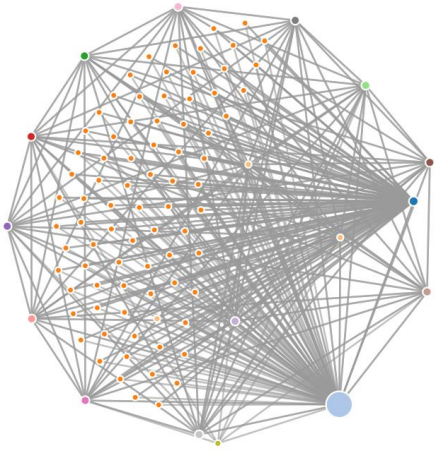

Figure 7: PageRank visualization.

Figure 7 illustrates the graph of PageRank. The biggest circle in light blue color represents the parent node. The other circles in orange color are the child nodes consisting of hyperlinks of the DHIs information. The size of the circle (big or small) depends on the computation page weightage for each of the hyperlinks. The page weightage is calculated based on the priority of hyperlinks. The bigger the size of the circle, the higher the priorities of the hyperlinks. PageRank also has some limitations. Firstly, the graph will become messier if the number of hyperlinks increases. Secondly, PageRank can only draw one big graph for one website or database. So, it cannot calculate the page weightage of more than one URL of a website or database. All the extracted herbs information is stored in JSON file format for future references and usage. All the ranking webpages are stored in SQLite.

Based on

Figure 8, the accuracy of each DHIs main website was different due to the different number of hyperlinks in the websites. Chinese Med Digital Projects and MedlinePlus recorded the same highest accuracy for both the indexing algorithms with 98% for PageRank and 71% for BFS. This is due to the number of hyperlinks of Chinese Med Digital Projects and MedlinePlus are higher than the other main websites. Whereas, Western Botanical Medicine website recorded the lowest accuracy, with 88% for PageRank and 58% for BFS. The lowest accuracy was achieved because of the missing hyperlinks that map into the other websites due to the removal of the domain name or unavailable domain name for the public. Based on Figure 9, PageRank for the GlobinMed website took the shortest time to execute the program compared to BFS. Meanwhile, the American Botanical Council website took the longest time for BFS compared to PageRank. In general, PageRank is faster than BFS as the number of hyperlinks in main websites increases. Based on the results, it can be concluded that PageRank has higher accuracy and is faster compared to BFS due to a few factors. 1) PageRank is generated using the entire internet graph, rather than a small set, it is less susceptible to localised linkage than other ranking systems, 2) a single indicator of a page's quality at the time of the crawl is coupled with a standard information retrieval score for the period, and 3) ranking is based on a page's popularity, therefore, it delivers the most relevant results. Therefore, PageRank is a better performer than BFS.

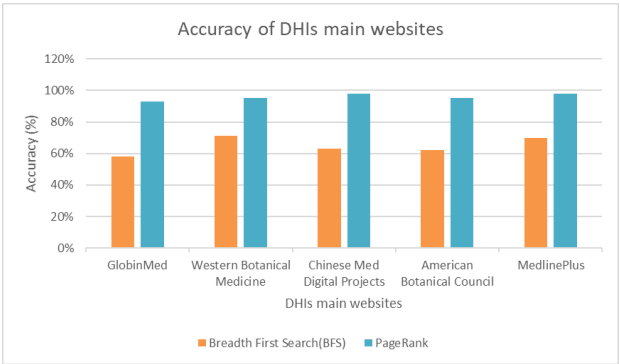

Figure 8: Accuracy of BFS and PageRank for DHIs main websites.

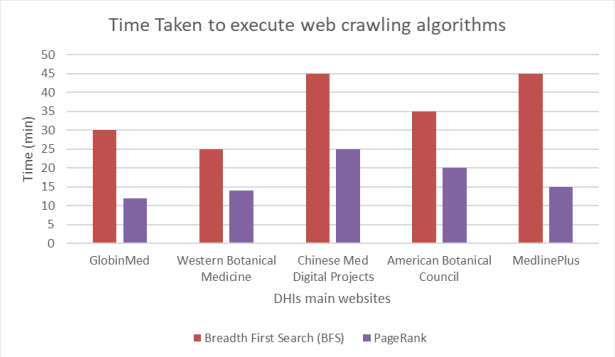

Figure 9: Time taken for web crawling algorithms to execute the program.

5 Conclusion and future work

This study proposed a framework for an efficient focus web crawler to organize and manage a large number of different herbal information and their interactions with drugs. In this proposed framework, the higher the accuracy of the indexing method, the higher the priority of relevant information of DHIs collected from databases and websites. In this study, BFS and PageRank were utilized on the dataset, where accuracy was also calculated. PageRank was more accurate for all the main websites. This focused web crawler provided accurate information while gathering relevant information on DHIs.

Based on the outcome, the indexing algorithms, time is taken for a crawler, and the visualization graph for indexing algorithm can be improved in the future. Firstly, more indexing algorithms could be used as this study only utilized two algorithms. Secondly, the BFS algorithm needs to be modified in terms of increasing the indexing depth of websites and databases. Next, the time taken to execute has to be reduced especially for modified BFS compared to modified PageRank to suit focus web crawler. Besides that, BFS consumes a lot of memories when more URLs are added into the queue as it also ranks some of the irrelevant information of DHIs. Hence, to overcome this limitation, the number of URLs of the DHIs webpages needs to be limited to increase efficiency, produce a better visualization, and rank results.

6 Acknowledgement

This project was supported by Smart Challenge Fund (SMART Fund) - SR0917Q1027 Ministry of Science, Technology and Innovation Malaysia (MOSTI).

7 References

[1] A. Fugh-Berman, E. Ernst, Herb-drug interactions: Review and assessment of report reliability, Br. J. Clin. Pharmacol. 52 (2001). https://doi.org/10.1046/j.0306-5251.2001.01469.x.

[2] R. Hooda, Herbal drug interactions - a major safety concern, Res. Rev. J. Pharmacogn. Phytochem. 4 (2016).

[3] B. Li, B. Zhao, Y. Liu, M. Tang, B. Lüe, Z. Luo, H. Zhai, Herb-drug enzyme-mediated interactions and the associated experimental methods: a review, J. Tradit. Chin. Med. 36 (2016). https://doi.org/10.1016/s0254-6272(16)30054-1.

[4] J.J. Bruno, J.J. Ellis, Herbal use among US elderly: 2002 National Health Interview Survey, Ann. Pharmacother. 39 (2005). https://doi.org/10.1345/aph.1E460.

[5] I. Meijerman, J.H. Beijnen, J.H.M. Schellens, Herb-Drug Interactions in Oncology: Focus on Mechanisms of Induction, Oncologist. 11 (2006). https://doi.org/10.1634/theoncologist.11-7-742.

[6] I. Cascorbi, Drug interactions - Principles, examples and clinical consequences, Dtsch. Arztebl. Int. 109 (2012). https://doi.org/10.3238/arztebl.2012.0546.

[7] N.C. for C. and I. Health, Herb-drug interactions, 355 (2015). https://doi.org/10.1016/S0140-6736(99)06457-0.

[8] M. Diligentit, F.M. Coetzee, S. Lawrence, C.L. Giles, M. Gori, Focused crawling using context graphs, in: Proc. 26th Int. Conf. Very Large Data Bases, VLDB'00, 2000.

[9] B. Novak, a Survey of Focused Web Crawling Algorithms, Proc. SIKDD. 5558 (2004).

[10] C. De Groc, Babouk: Focused web crawling for corpus compilation and automatic terminology extraction, in: Proc. - 2011 IEEE/WIC/ACM Int. Conf. Web Intell. WI 2011, 2011. https://doi.org/10.1109/WI-IAT.2011.253.

[11] R. Gaur, D.K. Sharma, Review of ontology based focused crawling approaches, in: ICSCTET 2014 - Int. Conf. Soft Comput. Tech. Eng. Technol., 2016. https://doi.org/10.1109/ICSCTET.2015.7371191.

[12] N. Pawar, K. Rajeswari, A. Joshi, Implementation of an Efficient web crawler to search medicinal plants and relevant diseases, in: Proc. - 2nd Int. Conf. Comput. Commun. Control Autom. ICCUBEA 2016, 2017. https://doi.org/10.1109/ICCUBEA.2016.7860006.

[13] Y. Qian, X. Ye, W. Du, J. Ren, Y. Sun, H. Wang, B. Luo, Q. Gao, M. Wu, J. He, A computerized system for detecting signals due to drug-drug interactions in spontaneous reporting systems, Br. J. Clin. Pharmacol. 69 (2010). https://doi.org/10.1111/j.1365-2125.2009.03557.x.

[14] H. Ibrahim, A. Saad, A. Abdo, A. Sharaf Eldin, Mining association patterns of drug-interactions using post marketing FDA's spontaneous reporting data, J. Biomed. Inform. 60 (2016). https://doi.org/10.1016/j.jbi.2016.02.009.

[15] M.L. Rethlefsen, MEDLINE: A Guide to Effective Searching in PubMed and Other Interfaces, J. Med. Libr. Assoc. 95 (2007). https://doi.org/10.3163/1536-5050.95.2.212.

[16] Review of various web page ranking algorithms in web structure mining, Int. J. Adv. Eng. Res. Dev.

3 (2015). https://doi.org/10.21090/ijaerd.ncrretcs20.

[17] C.J. Luh, S.A. Yang, T.L.D. Huang, Estimating Google's search engine ranking function from a search engine optimization perspective, Online Inf. Rev. 40 (2016). https://doi.org/10.1108/OIR-04-2015-0112.

[18] A.E. Wibowo, K.M. Lhaksmana, M. Isd, Perbandingan Peformansi Terhadap Algoritma Breadth First Search (BFS) & Depth First Search (DFS) Pada Web Crawler, E-Procceding Eng. 6 (2019) 9905–9914.

[19] NCI, NCI Dictionary of Cancer Terms, (n.d.). https://www.cancer.gov/publications/dictionaries/cancer-terms/def/adverse-event (accessed October 19, 2018).

[20] S. Scott, J. Thompson, Adverse drug reactions, Anaesth. Intensive Care Med. 15 (2014). https://doi.org/10.1016/j.mpaic.2014.02.008.

[21] J.K. Aronson, Medication errors: Definitions and classification, Br. J. Clin. Pharmacol. 67 (2009). https://doi.org/10.1111/j.1365-2125.2009.03415.x.

[22] WHO, Electronic Tools: Technical Series on Safer Primary Care., WHO Press. (2016) 1–21.

[23] J.K. Aronson, Medication errors: What they are, how they happen, and how to avoid them, QJM. 102 (2009). https://doi.org/10.1093/qjmed/hcp052.

[24] I.R. Edwards, J.K. Aronson, Adverse drug reactions: Definitions, diagnosis, and management, Lancet. 356 (2000). https://doi.org/10.1016/S0140-6736(00)02799-9.

[25] A.M. Mayo, D. Duncan, Nurse perceptions of medication errors what we need to know for patient safety, J. Nurs. Care Qual. 19 (2004). https://doi.org/10.1097/00001786-200407000-00007.

[26] M. Shamna, C. Dilip, M. Ajmal, P. Linu Mohan, C. Shinu, C.P. Jafer, Y. Mohammed, A prospective study on Adverse Drug Reactions of antibiotics in a tertiary care hospital, Saudi Pharm. J. 22 (2014). https://doi.org/10.1016/j.jsps.2013.06.004.

[27] J.R. Nebeker, P. Barach, M.H. Samore, Clarifying Adverse Drug Events: A Clinician's Guide to Terminology, Documentation, and Reporting, Ann. Intern. Med. 140 (2004). https://doi.org/10.7326/0003-4819-140-10-200405180-00009.

[28] S. V Taché, A. Sönnichsen, D.M. Ashcroft, Prevalence of Adverse Drug Events in Ambulatory Care: A Systematic Review, Ann. Pharmacother. 45 (2011). https://doi.org/10.1345/aph.1p627.

[29] A. Krähenbühl-Melcher, R. Schlienger, M. Lampert, M. Haschke, J. Drewe, S. Krähenbühl, Drug-related problems in hospitals: A review of the recent literature, Drug Saf. 30 (2007). https://doi.org/10.2165/00002018-200730050-00003.

[30] USER929, Introduction to MADRAC, (n.d.). https://www.npra.gov.my/index.php/en/about/malaysian-adverse-drug-reactions-advisory-committee-madrac/madrac-introduction (accessed October 19, 2018).

[31] H. See Lei, A.A. Fatah Rahman, A.M. Haq Syed Haq, Adverse drug reaction reports in Malaysia: Comparison of casuality assesmentds, Malaysian J. Pharm. Sci. 5 (2007).

[32] P. Posadzki, L. Watson, E. Ernst, Herb-drug interactions: An overview of systematic reviews, Br. J. Clin. Pharmacol. 75 (2013). https://doi.org/10.1111/j.1365-2125.2012.04350.x.

[33] A. Nahrstedt, V. Butterweck, Lessons learned from herbal medicinal products: The example of St. John's wort, J. Nat. Prod. 73 (2010). https://doi.org/10.1021/np1000329.

[34] P.C. Chan, Q. Xia, P.P. Fu, Ginkgo biloba leave extract: Biological, medicinal, and toxicological effects, J. Environ. Sci. Heal. - Part C Environ. Carcinog. Ecotoxicol. Rev. 25 (2007). https://doi.org/10.1080/10590500701569414.

[35] C. Gaudineau, R. Beckerman, S. Welbourn, K. Auclair, Inhibition of human P450 enzymes by multiple constituents of the Ginkgo biloba extract, Biochem. Biophys. Res. Commun. 318 (2004). https://doi.org/10.1016/j.bbrc.2004.04.139.

[36] P. de Bra, G.-J. Houben, Y. Kornatzky, R. Post, Information Retrieval in Distributed Hypertexts, RIAO. (1994).

[37] G.H. Agre, N. V. Mahajan, Keyword focused web crawler, in: 2nd Int. Conf. Electron. Commun. Syst. ICECS 2015, 2015. https://doi.org/10.1109/ECS.2015.7124749.

[38] N. Goyal, R. Bhatia, M. Kumar, A genetic algorithm based focused web crawler for automatic webpage classification, in: IET Conf. Publ., 2016. https://doi.org/10.1049/cp.2016.1546.

[39] G.A.F. Alfarisy, F.A. Bachtiar, Focused web crawler for Indonesian recipes, in: Proc. - 2017 Int. Conf. Sustain. Inf. Eng. Technol. SIET 2017, 2018. https://doi.org/10.1109/SIET.2017.8304134.

[40] K. Das, S.K. Sinha, Essential pre-processing tasks involved in data preparation for social network user behaviour analysis, in: Proc. Int. Conf. Intell. Sustain. Syst. ICISS 2017, 2018. https://doi.org/10.1109/ISS1.2017.8389423.

[41] C. Jain, P. Flick, T. Pan, O. Green, S. Aluru, An Adaptive Parallel Algorithm for Computing Connected Components, IEEE Trans. Parallel Distrib. Syst. 28 (2017). https://doi.org/10.1109/TPDS.2017.2672739.

[42] Y. Chen, J. Hu, H. Zhao, Y. Xiao, P. Hui, Measurement and Analysis of the Swarm Social Network with Tens of Millions of Nodes, IEEE Access. 6 (2018). https://doi.org/10.1109/ACCESS.2018.2789915.

[43] A. London, T. Németh, A. Pluhár, T. Csendes, A

local PageRank algorithm for evaluating the importance of scientific articles, Ann. Math. Informaticae. 44 (2015).

[44] A. Vishwakarma, R. Saxena, M. Awasthi, M. Yamuna, Comparative analysis of PageRank and hits: A review, Int. J. Pharm. Technol. 8 (2016).

[45] R. Prajapati, S. Kumar, Enhanced weighted PageRank algorithm based on contents and link visits, in: Proc. 10th INDIACom; 2016 3rd Int. Conf. Comput. Sustain. Glob. Dev. INDIACom 2016, 2016.

[46] P. Lahoti, G. De Francisci Morales, A. Gionis, Finding topical experts in twitter via query-dependent personalized PageRank, in: Proc. 2017 IEEE/ACM Int. Conf. Adv. Soc. Networks Anal. Mining, ASONAM 2017, 2017. https://doi.org/10.1145/3110025.3110044.

[47] D. Shestakov, Intelligent Web Crawling, IEEE Intell. Informatics Bull. 14 (2013) 5–7. http://www.comp.hkbu.edu.hk/~iib/2013/Dec/article1/iib_vol14no1_article1.pdf (accessed August 1, 2019).

[48] I. Rogers., The Google PageRank Algorithm and How It Works, (n.d.).