Enhancing Information Retrieval System through Web Crawler for Design and Build Company

Panot Khongkhaluang[1] and Phannachet Na Lamphun[2]

[1,2]Faculty of Engineering and Technology, Panyapiwat Institute of Management
Nonthaburi, Thailand
E-mail: 6272100116@stu.pim.ac.th, phannachetnal@pim.ac.th

Received: July 19, 2021 / Revised: August 10, 2021 / Accepted: August 11, 2021

***Abstract*—The purposes of this project are to develop a prototype online retrieval system to alleviate the need for regular labor for personnel who perform redundancy tasks on a daily basis and to obtain accurate information to reduce errors that may occur during the course of their jobs. The prototype system was created by combining various software applications such as the Python programming language, PHP, and the MySQL database system to store various data that will be used in fascinating projects.**

The users inside the corporate business development department have given the prototype a positive rating and were satisfied with the trials. The prototype may be further enhanced for practical use by substantially decreasing processes that support screening information for online procurement projects that match the organization's requirements and guidelines. A report may be sent to management for review and may result in new business opportunities as well as the possibility of working with suitable business partners. The results of user satisfaction overall satisfaction have a mean of 4.46.

***Index Terms*—Information Retrieval, Automate Document Generate, Data Extraction, Design and Build, Progress Report**

I. Introduction

Collaboration is an important key to success for any Design and Build business. Design and Build businesses are construction businesses that provide a range of services, including design services for buildings, facilities, utilities, and construction services for projects. The service also includes management, monitoring, and control of projects. Due to the size of some Design and Build projects, such as the construction of buildings, airports, or mass rapid transit systems, sub-contractors also work together to continue projects. For example, one sub-contractor may be responsible for ground clearing and pavements; a second sub-contractor may be responsible for office building construction; a third may be responsible for depot station operations and construction, and yet a fourth sub-contractor may have responsibility for providing and integrating systems for operations between office and depot. As completion of a project involves many parties, the collaboration between all parties is very important to enable each party to know about the status of the project and the continuity of actions.

The collaboration to complete the work might be simple, but it can be difficult to achieve because the sub-contractors can be from different companies in various places. People working together in one company can be complicated. However, it is different when people have to work together with others in different companies.

Document formatting is also a problem when working together. Documents must be accessible to all parties and must comply with the project owner's requirements. Document reports are crucial in any project because they show the progress of work, which is used to withdraw funds from the project owners. These documents must be well detailed and prepared, which can be difficult when working with more than one party. The main contractor must gather necessary information from sub-contractors and prepare the final report to present to the project owners. Regarding collaboration, time is another factor affecting many aspects of the project. The major effect lies in the development progress report to funding the project. Each month the main contractor must report to the project owners about the progress of the work. If the project proceeds to a certain level, the main contractor will be able to withdraw the budget. If they fail to present the progress report, the budget cannot be withdrawn. If this process is delayed, it means that the whole system will be slowly progressed, resulting in the delay of payment.

Therefore, collaboration is crucial in working together in Design and Building businesses. Sending a paper report to the main contractor for gathering together to develop a progress report is not effective enough because the process can be time-consuming,

and the information from sub-contractors might be incomplete or missing some details which can affect the whole report.

One of the keys to success in increasing collaborative performance is technology. This research aims to study the need for systems to help increase collaboration and alleviate problems regarding time consumption in the development of reports for Design and Build Businesses. Developing automatic information retrieval to gather information from the progress reports of sub-contractors and automatically developing the final progress report and presentation will help reduce working time, while the system helps monitor progress and informs each party with any necessary information.

II. Literature Review

A key aspect of automatic information retrieval on the web is Web Crawler. A web crawler is an internet bot that runs on a website. It can be called a web spider. The crawler is used by search engines and some websites to improve online content and collect data from other sources. It has the capability of copying pages that pass through it. Because of this, indexes can be used to obtain results much more quickly. The crawler can retrieve data faster than regular users [1]-[4]. While the usage of the crawler is advantageous for a variety of activities, the cost of the crawler comprises both network and computing resources. Moreover, much work over a longer duration can lead to an excessive strain, especially when the access frequency is high, resulting in the failure of the server or the router.

The task of web crawlers includes visiting internet applications, collecting data, and learning about new web pages from visited pages. Prior to this, the web crawlers started by collecting statistics of the web. Then it was developed to perform accessibility and vulnerability checks on the application [5]-[7].

A good web crawler must-have aspect as follows [1]:

Architecture: The basic requirements for a good web crawler are speed and efficiency. So, the web crawler is needed to be a well-defined architecture in order to function efficiently.

Intelligent Recrawling: Websites can be different in terms of formats. Some of the websites might be updated; thus, the format can be changed. It is crucial to develop the crawler intelligently, so it will be able to analyze the format and frequency so that the pages can get updated on the targeted websites.

Thorough/ Efficient Algorithms: The process of web crawler usually goes by Last in First out (LIFO) or First in First out (FIFO) which theoretically seem to work well. However, in some cases, data need to be traversed per crawl sprawls larger and deeper than anticipated. Then the crawler can be stuck in that situation and unable to process any further. To avoid such a situation, the web crawling system should be able to analyze the crawling time and divide the tasks among all data crawlers equally so that there are no vacant resources.

Scalability: In some processes, the data might be large in terms of size and number. This can result in the crawler taking more time than necessary to be able to extract and store. Therefore, the crawler should be able to compress the data before fetching or using a bounded amount of storage for storage-related scalability.

Language Independent: A web crawler needs to be language-neutral and extract data in all languages as requested across the globe.

Politeness: The use of a crawler should be in a polite manner. By sending a badly timed or poorly structured data crawler to crawl on the web, a DoS crash on the internet might occur. Due to this, the user should avoid sending the crawler repeatedly, for it might result in a site crash from the overcrowded traffic of the server.

The web crawler can be divided into several types as follows [8]:

Email Crawling: This crawler extracts email addresses from designated locations. However, this kind of crawler is illegal as it violates personal privacy. This type of crawler, therefore, cannot be used without users' permission.

News Crawling: This crawler extracts news from all over the world through the internet.

Image Crawling: This crawler extracts the images or visual representations that are available on the Internet. This type of crawler helps people find relevant pictures in a plethora of images across the web.

Social Media Crawling: This crawler extracts information from social media platforms. However, some platforms are legal with crawling, whereas others might be considered illegal because of the use of personal violations.

Video Crawling: This crawler extracts VDO and its locations.

Put simply, the capabilities of crawlers are not only limited to information extraction, but also an application from various fields. However, legal and privacy compliance of the crawlers should be strictly considered.

Regarding the process of a web crawler, it starts from Frontier or the system gets a set of seed URLs [6], [7], [9]. The seed URL is the web location where the user wants the crawler to process and extract information. The seed URLs are passed to a module called Fetcher that will retrieve the contents of the pages. Then these contents are passed through the Link Extractor to extract content and information. Then the process continues to store the processor

which interacts with the database and stores the discovered links and information. In addition, the Page Filter will filter any URLs that are not interesting to the web crawler to reduce time and any irrelevant data. The process is shown in Fig 1.

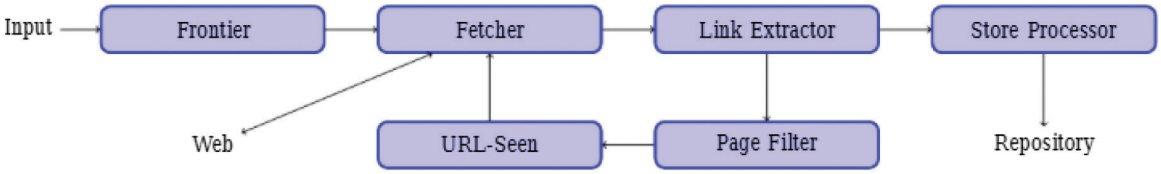

Fig. 1. Process of web crawler
From: https://www.researchgate.net/publication/262071338_A_Brief_History_of_Web_Crawlers)

The web crawler can be applied in various fields including the network information field which the crawler will obtain data, process data, or analyze data and feed it back to users. So, the users will have the most up-to-date information. In the financial sector field which the crawler will get the financial data on the web to help support user decision making [10]. The crawler is the key to success in design and build business. This is because it is a unique business that contains all services of the construction project for the customer. The cost for the project can be from one million Baht to even more than one hundred million Baht. This can be very competitive. So, the crawler can give the edge to the company to get the information fast and act fast to be able to complete the proposal and submit it to the process [11].

III. Methodology

The system was developed based on the web crawler concept including the front-end and back-end. In the front-end, the process starting from the system will display information that is used as a criterion for crawlers to consider extracting information. The users can adjust criteria to suit their needs. Any changes occurring will be saved back to the database.

Users can be selected from storing keywords or adding new keywords into the system. This is because there might be new interesting keywords that users would like the crawlers to find and extract information from the internet. The prototype system was developed in the type of check-box which users can select any interesting keyword as shown in Fig 2. The second process is to assign the website where users would like the crawler to extract information. The design is similar to the keyword page which contains the check-box where users can select the websites. Besides, users can add new websites into the system, if necessary, as shown in Fig 3.

Apart from the front-end, the back-end system will process to the listed websites and start extracting data based on the saved criteria. In case the data are found, they will be extracted and stored in the database for further usage. After the crawlers complete the process, users can view and select the information that matches the users' needs from the database that crawlers have been collected as shown in Fig 4.

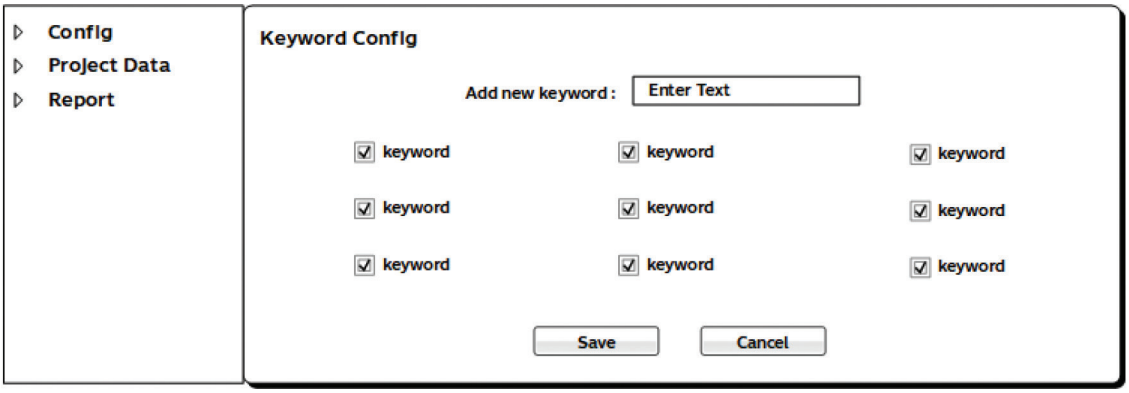

Fig. 2. Keyword setting page

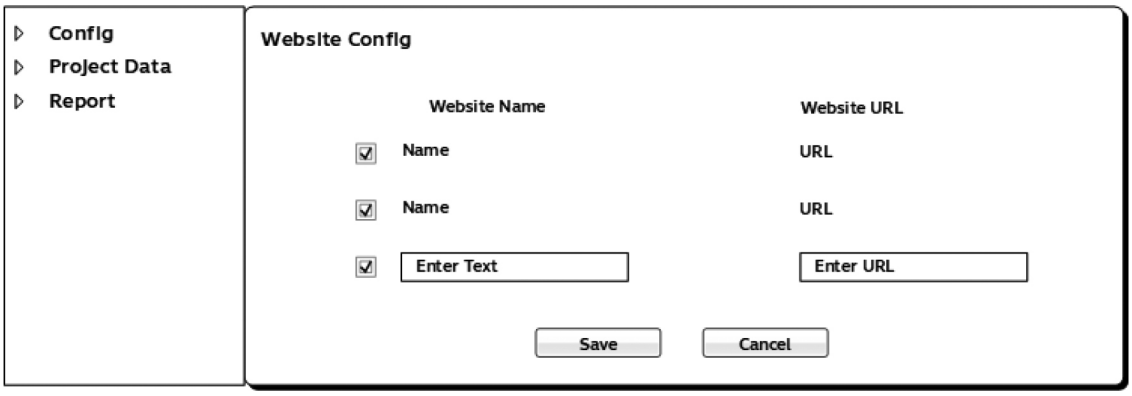

Fig. 3. Web setting page

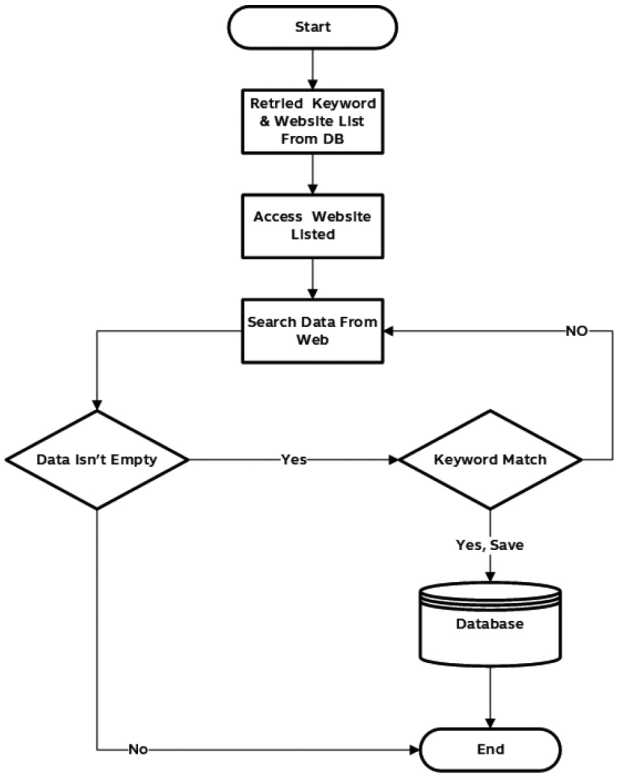

Fig. 4. Process of a prototype system

The database of the prototype system includes several fields as follows:

Web_Type: The web type is used to keep the type of websites either state or private ones.

Web_List: The web list stores URL links of the website.

Keyword: The keyword is used to collect settings from users.

Filter_Permission: The filter permission can be used to select their projects of interest at levels 1 and 2.

Project_Data: The project data are used to store information that the crawler collects from the website.

The information can be used further to meet users' needs. The system will collect and display data in a dashboard view which details' summary. On this page, the user can select interesting information to export in the report format as shown in Fig 5.

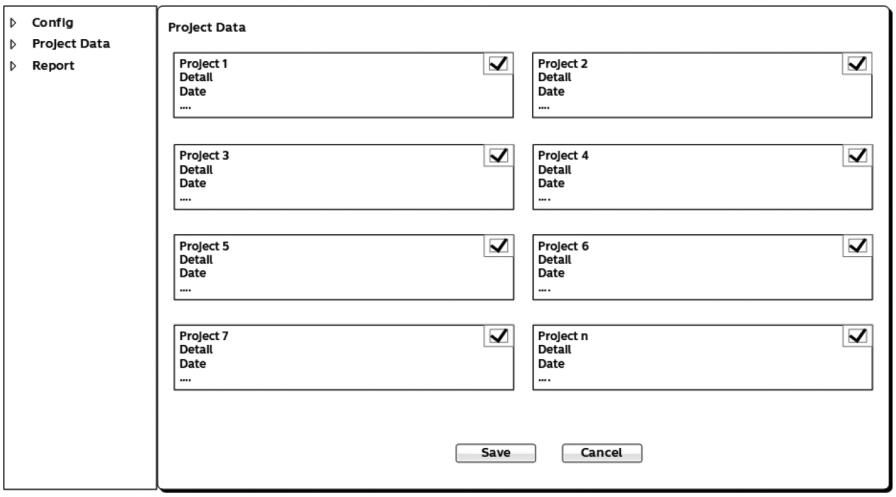

Fig. 5. Result selection page

IV. Experiment

The prototypes were tested in the real environment by crawling to designate websites and extract information based on the assigned keywords. The system was assigned to start the process at 3.00 a.m. where the internet traffic was flowing to avoid having any unwanted problems. The keywords used in the experiment were involved in the design and build project which covered civil duties. The keywords are:

- Master Plan: The concept designed for the development of the project is compliant with engineering standards and can be used to develop the project's construction.
- Survey: Its purpose is to explore and analyze various factors to assemble and develop projects.
- Detailed Design: Its purpose is to design project details, such as a structural design, the construction, an engineering drawing and, a structural model.
- Design-Build: Its purpose is to carry out the project in a design-oriented manner in parallel with the project's construction.
- Consultant: Procurement of engineering professionals to be project advisors, supervising, giving advice, and suggesting guidelines for project implementation.
- Project Management: Its purpose is to control the implementation of the project as well as to control costs and conditions under TOR to be completed within the project period or with a faster specified time frame.
- Construction Supervision: Its purpose is to control the construction works according to the detailed drawings, use of building materials according to the requirements, and the schedule of construction.

The website for testing was developed based on the real web structures of that site with the test data assigned to each site. This is to avoid any issues related to legality and damages that might occur while testing. The interface of the prototype system for keywords and a website selection is shown in Fig. 6 and Fig. 7.

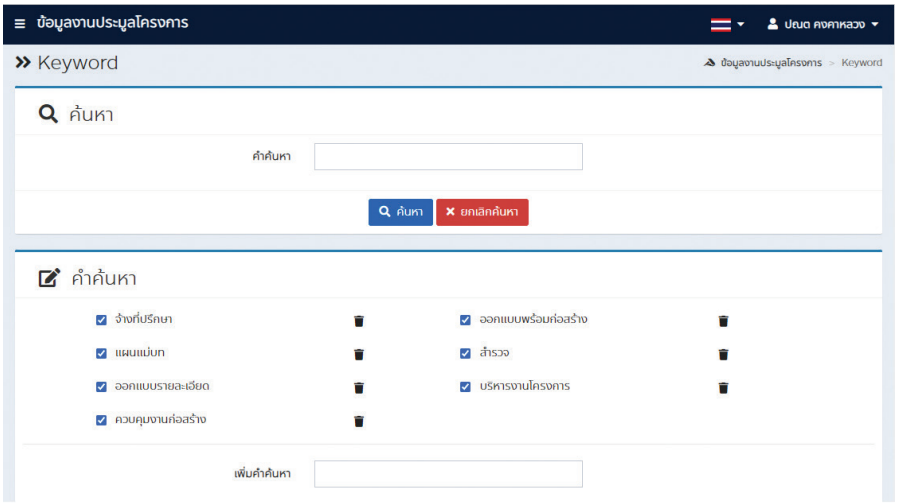

Fig. 6. Keyword setting page of a prototype system

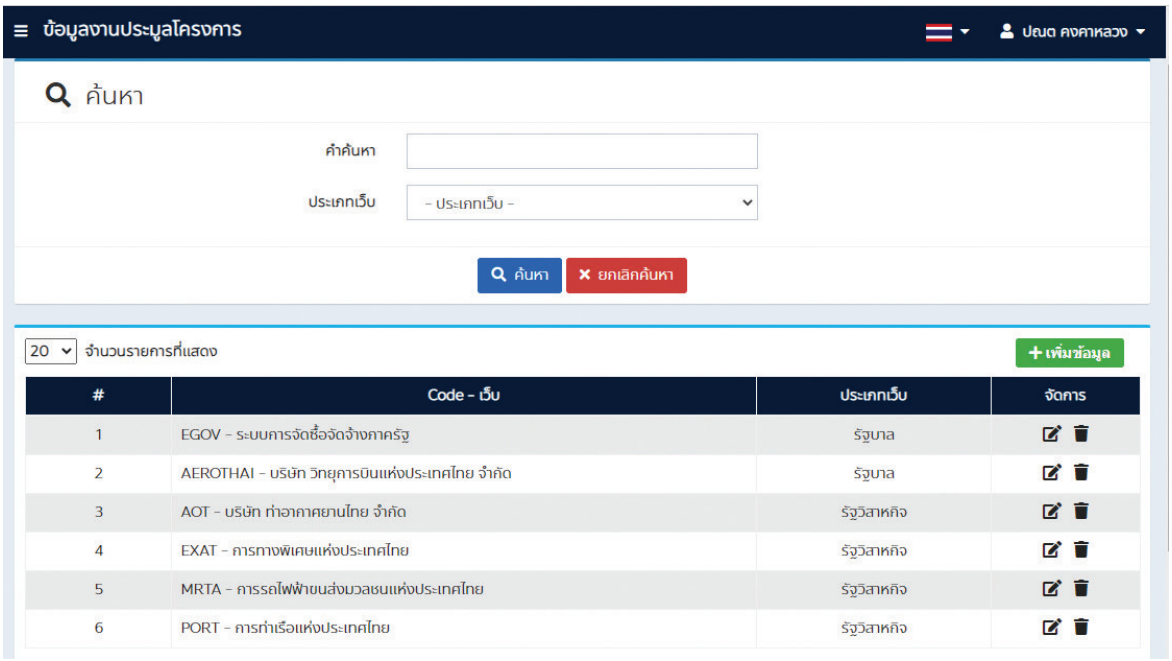

Fig. 7. Website setting page of a prototype system

After the keywords and websites were assigned, the prototype system was activated at 3.00 a.m. and started collecting the data. The process showed no errors during this phase. Once the data were collected in the database, during the working periods, the users could open the system and check the collected data.

The prototype system displayed the information as shown in Fig. 8. Then the users could select any data that might relate to the company's interest and export those data to report the and can be sent to the business development department for further work as shown in Fig. 9.

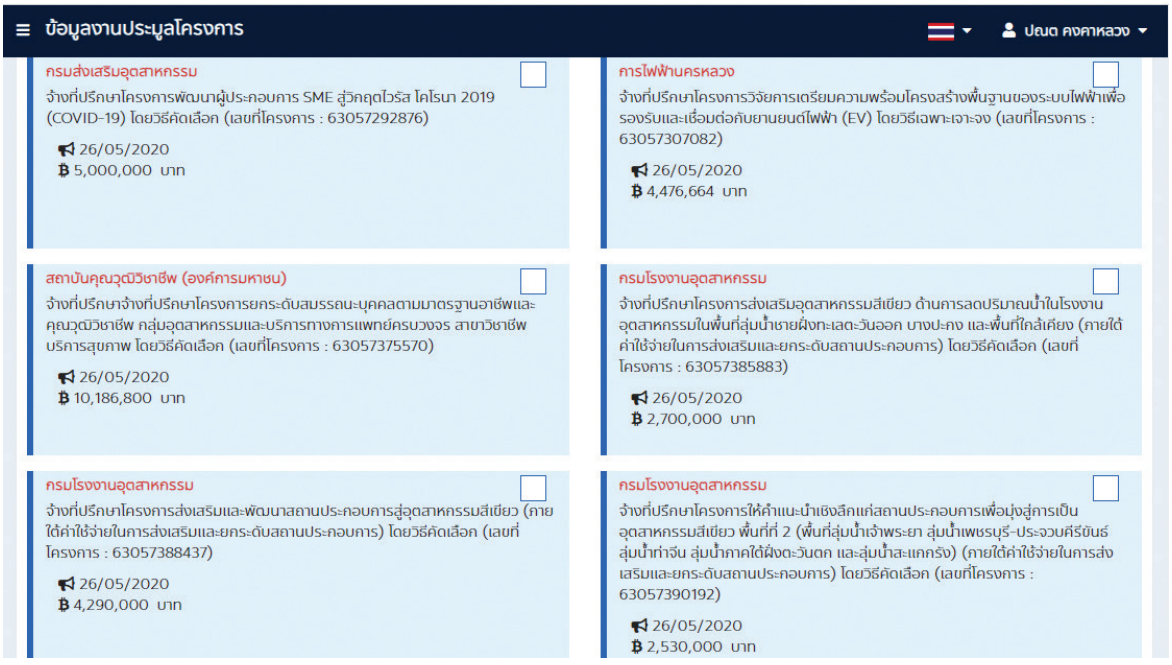

Fig. 8. An information display and a selection of prototype systems

	A	B	C	D	E	F
1	ที่	รหัสโครงการ	หน่วยงาน	โครงการ	มูลค่า	ประกาศเมื่อ
2	1	63057292876	กรมส่งเสริมอุตสาหกรรม	จ้างที่ปรึกษาโครงการพัฒนาผู้ประกอบการ SME สู่วิกฤตไวรัส โคโรนา 2019 (COVID-19) โดยวิธีคัดเลือก (เลขที่โครงการ : 63057292876)	5,000,000	26/5/2020
3	2	63057307082	การไฟฟ้านครหลวง	จ้างที่ปรึกษาโครงการวิจัยการเตรียมความพร้อมโครงสร้างพื้นฐานของระบบไฟฟ้าเพื่อรองรับและเชื่อมต่อกับยานยนต์ไฟฟ้า (EV) โดยวิธีเฉพาะเจาะจง (เลขที่โครงการ : 63057307082)	4,476,664	26/5/2020
4	3	63057375570	สถาบันคุณวุฒิวิชาชีพ (องค์การมหาชน)	จ้างที่ปรึกษาจ้างที่ปรึกษาโครงการยกระดับสมรรถนะบุคคลตามมาตรฐานอาชีพและคุณวุฒิวิชาชีพ กลุ่มอุตสาหกรรมและบริการทางการแพทย์ครบวงจร สาขาวิชาชีพบริการสุขภาพ โดยวิธีคัดเลือก (เลขที่โครงการ : 63057375570)	10,186,800	26/5/2020
5	4	63057385883	กรมโรงงานอุตสาหกรรม	จ้างที่ปรึกษาโครงการส่งเสริมอุตสาหกรรมสีเขียว ด้านการลดปริมาณน้ำในโรงงานอุตสาหกรรมในพื้นที่ลุ่มน้ำชายฝั่งทะเลตะวันออก บางปะกง และพื้นที่ใกล้เคียง (ภายใต้ค่าใช้จ่ายในการส่งเสริมและยกระดับสถานประกอบการ) โดยวิธีคัดเลือก (เลขที่โครงการ : 63057385883)	2,700,000	26/5/2020
6	5	63057388437	กรมโรงงานอุตสาหกรรม	จ้างที่ปรึกษาโครงการส่งเสริมและพัฒนาสถานประกอบการสู่อุตสาหกรรมสีเขียว (ภายใต้ค่าใช้จ่ายในการส่งเสริมและยกระดับสถานประกอบการ) โดยวิธีคัดเลือก (เลขที่โครงการ : 63057388437)	4,290,000	26/5/2020

Fig. 9. A report format of a prototype system

The prototypes system also provided the report. This is to inform the user of how the process was carried on. If there was no error, then the status would be complete. However, in some cases, when the website changes or was adjusted in its format, an interruption to the process can occur. Interrupted, the system will indicate the status as incomplete for that website. Moreover, the system counts the number of related data that we're able to extract from that web. Therefore, the users will have information that they can recheck for accuracy as shown in Fig. 10.

Report Project Retrieval Data — Date: xxxx-xx-xx

Web Name	Quanity (Project)	Status
Web 1	20	Complete
Web 2	10	Complete
.	.	.
.	.	.
.	.	.
Web n	30	Complete

More Project

Fig. 10. Process report of a prototype system

The results showed that the crawler was able to extract the assigned data from the website accurately. However, when confirmed with the Business Development department, 75% of the collected data were related to a company's interest. This occurred because of the unclear keywords of the Thai language. Nevertheless, the data that were corrected can be used for a company to apply for bidding and starting more businesses.

V. Evaluation

As for the results of user satisfaction, the overall satisfaction is 4.46. The details of the evaluation are displayed in Table I. According to the evaluation results, the highest average level of user satisfaction from the online retrieval system for a design-build project, "The system has fast and precise processing as needed" received an average level of satisfaction of 4.80

at 96%, which can be interpreted as the most suitable. These statistics also showed that the developer made a good planning process and developed suitable tools to meet the users' requirements.

For the readiness and overall satisfaction of the system, it was revealed that both lists achieved the satisfaction level of 4.70 or 94%, implying that the developer understands the needs of users and pays to recognize the work significance for business operations. With these results, it is possible to design and develop a prototype system ready for use.

TABLE I
ASSESSMENT RESULTS FROM THE COMPANY'S BUSINESS DEVELOPMENT UNIT

No.	Assessment lists	Level	S.D.	%	Levels of satisfaction
1	Be able to understand and learn the system immediately.	4.50	0.53	90	Most suitable
2	Information obtained by retrieving information on the website are complete and correct	4.30	0.82	56	Most suitable
3	Various components on the system screen are clear and understand easily.	4.40	0.52	88	Most suitable
4	The system has fast and precise processing as needed.	4.80	0.42	96	Most suitable
5	The system presents information correctly which meets user needs.	4.30	0.48	86	Most suitable
6	The report can be exported in an Excel format correctly.	4.30	0.67	86	Most suitable
7	The system can help reduce work processes.	4.40	0.52	88	Most suitable
8	The system is stable in operation.	4.20	0.63	84	Most suitable
9	The system is ready for service.	4.70	0.48	94	Most suitable
10	Overall satisfaction with the use of the system	4.70	0.48	94	Most suitable
Total average		**4.46**	**0.56**	**89**	**Most suitable**

VI. CONCLUSION

Findings of this research revealed that the prototype system facilitates searching and provides the project information to personnel in order to simplify storage and minimize time spent on regular activities. This prototype system can also help the organizations to find and gather the project information to ensure fast data collection and completion. Moreover, the system reduces the number of personnel required to perform routine tasks without affecting their works. The uses of the system also increase business opportunities for the organization by obtaining job information or collaborating with a new business organization while providing up-to-date information.

VII. FUTURE WORK

Further studies can be done such as developing a subscription system that allows membership applications to gain access to the online procurement search system for design and construction projects and packages at various prices, as well as the system's capabilities at various levels. Besides, researchers can continue studying the development of a useful add-on for the system in the field of analysis by using data collected from the prototype system's storage in conjunction with the aforementioned packages or to be used as a separate package that is specialized for analysis capabilities of the prototypes system.

REFERENCES

[1] R. Singla. (2021, Jun 10). *Six Must-Have Qualities of a Good Web Crawler.* [Online]. Available: https://bigdata-madesimple.com/six-must-have-qualities-of-a-good-web-crawler/

[2] A. Nemeslaki and K. Pocsarovszky. (2021, Jun 1). *Web Crawler Research Methodology*. [Online]. Available: https://www.researchgate.net/publication/254460232_Web_crawler_research_methodology

[3] N. V. Dasari. (2021, Jun 1). *Web Crawling and IR*. [Online]. Available: https://www.cfilt.iitb.ac.in/resources/surveys/CrawlingAndIR-Varun-May14.pdf

[4] C. Olston and M. Najork, *Web Crawling, Foundations and Trends in Information Retrieval*. Boston, USA: Now Publishers Inc, 2010, pp. 1-86.

[5] V. Shkapenyuk and T. Suel, "Design and Implementation of a High-Performance Distributed Web Crawler," in *Proc. The International Conference on Data Engineering*, 2002, pp. 357-68.

[6] S. M. Mirtaheri, M. E. Dinçktürk, S. Hooshmand et al. (2021, Jun 10). *A Brief History of Web Crawlers.* [Online]. Available: http://ssrg.eecs.uottawa.ca/docs/CASCON2013.pdf

[7] M. S. Ahuja, J. S. Bal, and Varnica, "Web Crawler: Extracting the Web Data," *International Journal of Computer Trends and Technology*, vol. 1, no. 8, pp. 629-632, Jul. 2015.

[8] T. V. Udapure, R. D. Kale, and R. C. Dharmik. (2021, Jun 10). *Study of Web Crawler and Its Different Types about Types of Web Crawler*. [Online]. Available: http;//www.iosrjournal.org/iosr-jce/papers/Vol16-issue1/Version-6/A01616010 5.pdf

[9] B. M. Jasani and C. K. Kumbharana, "Analyzing Different Web Crawling Methods," *International Journal of Computer Trends and Technology*, vol. 107, no. 5, pp. 975-8887, Dec. 2014.

[10] L. Yu, Y. Li, Q, Zeng et al. (2020, Nov.). Summary of Web Crawler Technology Research. *Journal of Physics: Conference Series,* [Online]. *1449*, pp. 5-7. Available: https://iopscience.iop.org/article/10.1088/1742-6596/1449/11/012036/pdf

[11] P. Khongkhaluang, P. Na Lamphun, and A. Kheaksong. (2021, Jun.). Increasing Performance Through Information Retrieval System for Design and Build Company. *International Scientific Journal of Engineering and Technology*. [Online]. *5(1)*. pp. 32-40. Available: https://Ph02.tci-thaijo.org/index.php/isjet/article/view/243422/165745

Panot Khongkhaluang received the Bachelor of Science (Information Technology), Panyapiwat Institute of Management, Thailand. Past work experience on acting IT Manager at Index International Group Public Company Limited. (IND).

Phannachet Na Lamphun received the Master of Science (Computer Engineering), Polytechnic institute of New York University, New York, USA, and Doctor of Engineering (Information and Communications Technologies), Asian Institute of Technology. Past work experience is at Index International Group Co., Ltd. as a system engineer for Project Management Consultation of MRTA Blue Line, Dust free Road Project.