Editor-in-Chief of the IJSEIA Journal

Haeng-Kon Kim,
Catholic University of Daegu, Daegu, Korea

Jinan Fiaidhi
Lakehead University, Canada

General Information of IJSEIA

Bibliographic Information
- ISSN: 1738-9984
- Publisher: SERSC
  Science & Engineering Research Support Society

Contact Information
Science & Engineering Research Support Society
- Head Office: 20 Virginia Court, Sandy Bay, Tasmania, Australia
- Phone no.: +61-3-9016-9027
- Email: ijseia@sersc.org

Journal Topics

IJSEIA aims to facilitate and support research related to software engineering technology and the applications.

To bridge the gap of users who do not have access to major databases where one should pay for every downloaded article; this online publication platform is open to all readers as part of our commitment to global scientific society.

The topics covered by IJSEIA include the following:
- Component-Based Software Engineering
- Software Requirements Engineering
- Software Architectures & Design
- Information Engineering
- Mobile/Wireless Computing & Applications
- Communication Systems & Networks
- Visual and Multimedia Computing
- Intelligence
- Economic and Financial Systems
• Software Security
• Modeling with UML
• Reverse Engineering & Reengineering
• Use Cases, Scenarios
• Middleware Design Techniques
• Software Economics & Cost Modeling
• Programming Languages
• Software Tools and Development environments
• Software Policy and Ethics
• Theory, Formal Methods and Tools
• Design Patterns and Frameworks
• Software Safety and Reliability
• Web Engineering, & Web-Based Applications
• Software Process Models
• Parallel and Distributed Computing
• Data Mining, Knowledge Discovery, Artificial Intelligence
• Information Management Systems, MIS
• Software Testing & Analysis
• Software Components and Reuse
• Case Studies
• Object-Oriented Techniques
• Agent Technology
• Project Management and Maintenance
• Embedded and Real-Time Software
• Empirical Software Engineering and Metrics
• Aspect-Orientation and Feature Interaction
• Software Development Methodology
• Software Process Improvement
• Project Management
• Quality Model
• Cost Estimation
• Numerical (Statistics) Analysis Development of Algorithms
• Simulation and Design Methodology for H/W and S/W System
Advisory/Editorial Board

- Abdelwahab Hamou-Lhadj, Concordia University, Canada
- Abhik Roychoudhury, National University of Singapore, Singapore
- Alok Mishra, Atılım University, Turkey
- Andrea De Lucia, Università di Salerno, Italy
- Ashesh Mahidadia, University of New South Wales, Australia
- Aymen Issa Zreikat, Mutah University, Jordan
- Bernard Wong, University of Technology, Sydney, Australia
- Bivas Dinda, Mahishamuri Ramkrishna Vidyapith, India
- Cornelia Boldyreff, University of Lincoln, UK
- D.Subbarao, Chitkara University, India
- Danilo Caivano, University of Bari, Italy
- Debnath Bhattacharyya, Heritage Inst. of Technol., Kolkata, India
- Devendra K. Tayal, Jaypee Institute of Information Technology, India
- Elizabeth Burd, University of Durham, UK
- Emilia Mendes, University of Auckland, New Zealand
- Félix García, University of Castilla-La Mancha, Spain
- Frank Houdek, Daimlerchrysler, Germany
- Giuseppe Visaggio, University of Bari, Italy
- Gregory M. Kapfhammer, Allegheny College, USA
- Harvey Siy, University of Nebraska, USA
- Hironori Washizaki, National Institute of Informatics, Japan
- Hongji Yang, De Montfort University, UK
- Houari Sahraoui, University of Montreal, Canada
- Jemal H. Abawajy, Deakin University, Australia
- Jianjun Zhao, Shanghai Jiao Tong University, China
- Jin Song Dong, National University of Singapore, Singapore
- Jonathan Lee, National Central University, Taiwan
- Juan Fernandez-Ramil, The Open University, UK
- Juan Jose Garcia Adeva, University of the Basque Country, Spain
- Jun Bi, Tsinghua University, China
- Justin Zha, Carnegie Mellon University, USA
- Karl Leung, The Chinese University of Hong Kong, Hong Kong
• Khademul Islam Molla, University of Tokyo, Japan
• Kouich Sakurai, Kyushu University, Japan
• Kousalya G, Lealta Media, India
• Luigi Buglione, Engineering.IT / Nexen, Italy
• M. Govindarajan, Annamalai University, India
• Manoel Gomes Mendonça, Salvador University, Brazil
• Maricel Balitanas, University of San Agustin, Philippines
• Mokhtar Beldjehem, Sainte-Anne's University, Canada
• N. M. Nahar, Central Arid Zone Research Institute, India
• Nabor Chagas Mendonça, Salvador University, Brazil
• Oded Cohn, IBM Research Lab- Haifa, IL
• Pavan Kumar Gorakavi, IPMA-USA Young Crew/PMI-ISSIG, USA
• PAWAN JINDAL, Jaypee University of Engineering and Technology, India
• Pornsiri Muenchaisri, Chulalongkorn University, Thailand
• Randy Tolentino, Hannam University, Korea
• Robert L. Glass, Griffith University, Australia
• Rosslin John Robles, University of San Agustin, Philippines
• Rudolf Ferenc, University of Szeged, Hungary
• Sandro Morasca, University degli Studi dell’Insubria, Italy
• SeokSoo Kim, Hannam University, Korea
• Shamim Ahmad, University of Rajshahi, Bangladesh
• Silvia Abrahão, Universidad Politécnica de Valencia, Spain
• Somlal Das, University of Rajshahi, Bangladesh
• Stan Jarzabek, National University of Singapore, Singapore
• Tibor Gyimóthy, University of Szeged, Hungary
• Tien Nhut Nguyen, Iowa State University, USA
• Vitus Sai Wa Lam, University of Hong Kong, Hong Kong
• Xiang Fu, Georgia Southwestern State University, USA
• Yann-Gaël Guéhéneuc, University of Montreal, Canada
• Yijun Yu, The Open University, UK
• Yong-Kee Jun, Gyeongsang National University, Korea
IJSEIA is indexed by:

- EBSCO
- ProQuest
- ULRICH
- DOAJ
- OpenJ-Gate
- SCOPUS
- Cabell

Editorial Secretary

- Ronnie D. Caytles
Foreword and Editorial

International Journal of Software Engineering and Its Applications

We are very happy to publish this issue of an International Journal of Software Engineering and Its Applications and Pattern Recognition by Science and Engineering Research Support Society.

This issue contains 20 articles. Achieving such a high quality of papers would have been impossible without the huge work that was undertaken by the Editorial Board members and External Reviewers. We take this opportunity to thank them for their great support and cooperation.

The research paper “A Reference Framework of Smart Library”, proposes functional requirements and an architectural reference framework of Smart Library as a prototype of future library information systems that is a seamless convergence of the library information systems and the cutting-edge information technologies. The functional requirements of Smart Library Systems by considering the users’ demands and the eco-systems of information technology are described. The referential framework that can accommodate functional requirements and provide smart knowledge services to the users through the diverse smart devices is also shown.

The primary objective of paper “An Approach to Evaluate Object Oriented Class Structure using Score Carding Framework” is to study and analyze the various metric suites for object oriented systems, and hence projects the existing parameters of the suite and their contribution to software quality. Then design and develop a software prototype called “Class Break point Analyzer (CBA)” for extracting the parameters of the studied metric suites, and frame “Class Quality Scorecards” to study the contribution of these parameters to software quality.

The paper “A Study on the Short Range Wireless Communication of Smart Clothing for Elderly and the Infirm” explores short range wireless communication that can be easily applied to smart clothing that the elderly use. The study was conducted through implementing a system using wireless communication and through implementing a system combining Bluetooth, ZigBee, and WiFi.

Authors of the paper “Applying Machine learning for Configuring Agile Methods”, introduces an Agile Method engineering approach to find the degree of suitability of agile methods for a particular situation. The introduced process uses associative clustering for finding the cluster of appropriate methods against the organisational requirements-in-hand. The specially designed fuzzy logic controller is used to extract the most appropriate methods from the cluster of appropriate methods. Fuzzy logic controller works in coordination with the databases that have been formed using the previous results and is being trained with the new knowledge. Finally two practical case studies have been discussed to describe how these concepts are applied in practice with industry specified requirements and results are being explored.
Paper “Effect of Gyro Sensor to the Position Error in Mobile Robot” focuses on the effect of gyro sensor by measuring position error which is comprised of angular error and distance error. In the experimentation, the position errors of mobile robots for five different courses are measured for the robot with a gyro sensor and those without it. The closed courses include three, five, seven, nine and eleven direction changes which cause to generate considerable position errors. The experimental result shows that the angular error is reduced to about 14.3%~41.4% after a gyro sensor is used. Further, the position errors in distance are reduced to 22.4~31.8% when a gyro sensor is used. The remarkable difference between the two kind of errors is that the angular error increases steadily whereas the distance error increases rapidly but does not increase anymore over a certain level as the number of direction change increases.

Paper “Utilizing CSI to Improve Distance Estimation Precision in the Indoor Environment”, deals with some experiments to improve the indoor distance estimation precision by taking advantage of the new RF channel property, CSI (Channel Status Information). It mainly aims to show how much the multipath effect can be mitigated with CSI rather than RSSI, then how it influences to the distance estimation.

Author of “Software Quality Assurance – E-commerce Customers Satisfaction in Requirements Engineering Process” describes the quality attributes that are gained from the requirements elicitation, requirements documentation, requirements validation, negotiation and the requirements management planning activities. The results show that functionality, security, usability, reliability and efficiency affect e-commerce customers’ satisfaction. Most of the online shopping websites comply with customer requirement and requirement expectations.

The paper “Improvement of Matrix Factorization-based Recommender Systems Using Similar User Index” present a new similar user index-based matrix factorization approach for large scale recommender systems. Finding similar users is the most time-consuming phase in large scale recommender systems. To reduce time to find the similar users, it proposes a similar user index in matrix factorization. This paper describes the index structure and algorithms. Several experiments are performed.

In the work “Optimization of the Smoothing Parameter of the Adaptive Kernel Estimator used in Bayes Classifier - Application to Microarray Data Analysis” authors focus on nonparametric kernel methods for estimating the probability density function (pdf). The convergence of a kernel estimator depends crucially on the choice of the smoothing parameter. They present in this paper, a new method for optimizing the bandwidth of an estimator of the probability density function: the adaptive kernel estimator. This optimized estimator is used to construct the Bayes classifier. In this sense, they have proposed a new approach to optimize the pdf based on the statistical properties of the probability distributions of random variables. They adopt the maximum entropy principle (MEP) in order to determine the optimal value of the smoothing parameter used in the estimator. In the proposed criterion, the estimated probability density function is called optimal in the sense of having a minimum error rate of classifying data. Finally, they illustrate the robustness of the optimization process of the kernel estimation methods by using a set of DNA microarray data showing that the approach effectively improves the performance of the classification process.
The paper “A Linux Filesystem Tracing Method Using the Kprobes Linux Dynamic Instrumentation System” presents an integrated performance analysis tool for Linux file systems. The tool provides actual time information for Linux file system functions. In contrast to other existing tools, the tool provides a filtering mechanism, a graphical interface, and system-level analysis information without a heavy load of measurement. It may be used by Linux developers or end-users for analyzing file system layers or measuring software performance to find bottlenecks in Linux file systems.

The study “Language Engineering for Creating Relevance Corpus” presents a software engineering solution for the process of creating relevance corpora that achieves reusability, flexibility, multilinguality and modularity, in order to respect the experimental nature of IR field. The software engineering solution is presented as UML models. This paper then shows how the proposed design model was used to implement the process of building an open source relevance Arabic corpus based on the Clue Web 2009 data set for the purpose of supporting research evaluating and improving search engines for Arabic language.

The Authors of paper “Fast Noise Tolerant Vision-based Defect Detection on Flight Deck Surface from Non-destructive Testing Images” propose a method to locate such defective areas automatically by computer vision. The proposed method applies a series of image processing techniques and strong against noise as verified in the experiment.

Paper “The Impact of Agile Methodology on Software Team’s Work-Related Well-Being” aims to investigate empirically the effect of agile methodology on software development team’s work-related well-being. To achieve this goal, a comparison study was carried out in an academic setting. A quantitative approach using statistical analysis was used to investigate the effect. Results showed that agile does not significantly affect work-related well-being. Nonetheless, the team that is able to apply the agile practices as closely as possible experienced higher level of enthusiasm during software project. This study provides additional empirical data in software engineering research and practices specifically on human aspects. Further investigation needs to be carried out on the software projects with higher task complexity.

The purpose of the paper “Optical Illusion using Histogram Analysis#” proposes a fuzzy based histogram equalization method. It adopts Kaufmann’s measure of fuzziness concept for histogram equalization. The main idea of this paper is to use and assess fuzziness. The proposed histogram equalization method is applied to Y channel of YUV, which is transformed signal from RGB image. The obtained and improved Y channel is rejoined to the other I and Q signals and retransformed to RGB image. The objective and the visual performance are compared in simulation results section.

The study entitled “An Empirical Application of an Information System to Relieve Chronic Obstructive Pulmonary Disease” attempted to investigate an empirical application of an information system to relieve chronic obstructive pulmonary disease. The subjects of this paper were 128 patients who had visited a general hospital which located in Chungnam area. As a result, first, for gender, the experimental group in a female with 64.1% showed higher than female with 60.9% in the control group. There was a significant difference in subjects who had intaked daikon after information intervention (t=−2.75, p=.000). In conclusion, a comprehensive adoption of an information system to minimize the damage of COPD will contribute effectively to the COPD recovery and prevention.
In the paper “Integrated MARTE-based Model for Designing Component-Based Embedded Real-Time Software” an integrated component model is proposed, by integrating MARMOT and a Unified Modeling Language (UML) profile known as Modeling and Analysis for Real-Time and Embedded (MARTE), which can be adapted to the CBSE approach standard modeling. In addition, the proposed model can improve the existing MARMOT software process.

Authors of the paper “A Study on Distributed Crawling-Based Overhead Optimization” investigates that crawlers can operate in one system, decentralizing a crawler’s main functions. By using/implementing a proposed system, the efficiency is compared, analyzed and verified with the existing problems.

Authors of the paper “Gamification on Phrase Building Training Application” show a way to add elements by gamification of an application for practicing building phrase which is mostly difficult to attract learners’ interests in spite of its importance in language learning.

The paper “Development of an Indoor-Outdoor Positioning Android App for Anapji Tourist Guides” introduces the development details of an Indoor and Outdoor Positioning Android app for Anapji, an artificial pond in Gyeongju National Park in Korea, tourist guide apps. There is a pathway around the pond, with around 10 viewpoints and Imhaejeon, a gazebo-like Korean traditional house, along this pathway. In Imhaejeon, it can find an Anapji bird's-eye view at the center and 12 cultural artifacts unearthed at Anapji displayed along the four edges of Imhaejeon. Tourists walk along the pathway and stop at a viewpoint to enjoy the scenic beauty at their wish. Most of them also enter Imhaejeon, walk along the edge, and stop in front of an artifact to watch it. Whenever a tourist stops to watch something, tourist guide apps show a content that is closely related to the thing that is being watched. The content can be a video clip describing how the artifact was crafted or a video game that mimics an ancient game played by the people in the 8th century. This kind of tourist guide app cannot be realized unless they can recognize the position where the user is located. This paper develops an indoor-outdoor positioning Android app that is an essential part of tourist guide apps.

March 2015

Haeng-kon Kim, Catholic University of Daegu, Daegu, Korea
Jinan Fiaidhi, Lakehead University, Canada

Editors of the March Issue on
International Journal of Software Engineering and Its Applications
# Table of Contents

**A Reference Framework of Smart Library**  
1  
*Hee-Kyung Moon, Ju-Ri Kim, Sung-Kook Han and Jin-Tak Choi*

**An Approach to Evaluate Object Oriented Class Structure using Score Carding Framework**  
9  
*P. Edith Linda, Dr. E. Chandra and Dr. J. Sharmila*

**A Study on the Short Range Wireless Communication of Smart Clothing for Elderly and the Infirm**  
17  
*Jooyoung Ko, Jaechang Shim and Hyenki Kim*

**Applying Machine learning for Configuring Agile Methods**  
29  
*Rinky Dwivedi and Daya Gupta*

**Effect of Gyro Sensor to the Position Error in Mobile Robot**  
41  
*Kyujung Lee and Choong Ho Lee*

**Utilizing CSI to Improve Distance Estimation Precision in the Indoor Environment**  
49  
*Jeong-Woo Yang and Gi-hwan Cho*

**Software Quality Assurance – E-commerce Customers Satisfaction in Requirements Engineering Process**  
57  

**Improvement of Matrix Factorization-based Recommender Systems Using Similar User Index**  
71  
*Haesung Lee and Joonhee Kwons*
Optimization of the Smoothing Parameter of the Adaptive Kernel Estimator used in Bayes Classifier - Application to Microarray Data Analysis  

Yissam Lakhdar and El Hassan Sbai  

79

A Linux Filesystem Tracing Method Using the Kprobes Linux Dynamic Instrumentation System  

Sang-Young Cho and Hyunsu Ha  

95

Language Engineering for Creating Relevance Corpus  

Nuha H. El-Khalili, Bassam Haddad and Haya El-Ghalayini  

107

Fast Noise Tolerant Vision-based Defect Detection on Flight Deck Surface from Non-destructive Testing Images  

Kwang Baek Kim and Doo Heon Song  

117

The Impact of Agile Methodology on Software Team’s Work-Related Well-Being  

Mazni Omar and Sharifah Laillee Syed Abdullah  

123

Optical Illusion using Histogram Analysis  

Gwanggil Jeon, Eunju Kim and Nam Kook Kim  

137

An Empirical Application of an Information System to Relieve Chronic Obstructive Pulmonary Disease  

Seong-Ran Lee  

147

Integrated MARTE-based Model for Designing Component-Based Embedded Real-Time Software  

Mohd Z. M. Zaki, Dayang N. A. Jawawi and Mohd Adham Isa  

157
A Study on Distributed Crawling-Based Overhead Optimization  175

Han-Gil Kim, Jong-Won Lee, Tae-Hak Ban and Hoe-Kyung Jung

Gamification on Phrase Building Training Application  183

Takayuki Morita and Mitsuhiro Kimura

Development of an Indoor-Outdoor Positioning Android App for Anapji Tourist Guides  195

Kyubark Shim, Jaegeol Yim and Junri Jeon
A Study on Distributed Crawling-Based Overhead Optimization

Han-Gil Kim, Jong-Won Lee, Tae-Hak Ban and Hoe-Kyung Jung*

PaiChai University, Doma2-Dong, SeoGu, DaeJeon, Korea
khg0482@yahoo.co.kr, starjwon@naver.com, thban@pcu.ac.kr, hkjung@pcu.ac.kr

Abstract

Network technology has been in the process of development with IT. Data on the network is increasing rapidly and there are also various data searching methods. One of these methods is a crawler. It finds data that users want and collects it. However, as data on the network increases exponentially, potential problems of the crawler are constantly emerging.

This paper investigates that crawlers can operate in one system, decentralizing a crawler’s main functions. By using/implementing a proposed system, the efficiency is compared, analyzed and verified with the existing problems.

Keywords: Crawler, Distribute, Overhead, Optimization

1. Introduction

With the development of IT technology, Internet data has been used exponentially. Processing data that users need in each field is important. Conventional management method alone was not able to manage the increasing data. The tool that emerged as a way to solve it is ETL (Extraction Transformation Loading). The ETL is the concept of using means to generate a value as information after the processing to extract the required data. It is a web crawler that has diverse functions, including collecting data [1-3].

Conventional crawlers can be roughly classified into four kinds: General, Focused, Topical, and Wrapper-based crawler. Crawlers that are currently/widely used a lot are General and Focused crawler.

Since the most widely used open source crawler is Crawler4j, it is a small crawler that individuals are capable of operating.

It has all the features of Universal, Focused, and Topical crawlers, and it is also possible to set the acquisition target in various ways. However, the length of the source is long and complex with a high memory usage since all the functions operate in one crawler and overhead problems occur frequently.

Due to overlapping collection and frequent error occurrence, it is not efficient; thus open source crawlers work as a basic reference in crawler studies.

In this paper, the new crawler system is designed/implemented to replace an open source crawler containing complexity, high memory usage, frequent overhead occurrence, and overlapping collection.

3 units of URLs were randomly selected and verified excellently in terms of memory usage and speed through the experiments and analysis. The efficiency was verified by setting the extension of collecting files, suppressing the collection of files that users do not want, and minimizing the data collection.

*Corresponding Author

ISSN: 1738-9984 IJSEIA
Copyright © 2015 SERSC
2. Related Researches

2.1. General Crawler and Focused Crawler

General and Focused crawlers search on a relatively large range, setting various targets and collecting files. To collect accurate data from the situation that the field/theme is not set, crawlers can be used in all fields and collect a lot of data. The disadvantages of them are overlapping and lack of accuracy [4,5].

2.2. Topical Crawler

If you need to set the field or topic, or you want search and accurate retrieval and data collection, topical crawlers are suitable for the use. They narrow the range for precise search compared to General crawlers/ Focused crawlers. Tropical crawlers are fast because there are not many collected targets and overlapping collection. They are often used in the medical field. Since the search range is limited, small amount of data can be collected. It is difficult to ensure the integrity of the data [6, 7].

2.3. Wrapper-based Crawler

Wrapper-based crawlers pre-analyze the website desired to collect, set up required web pages and collect data only on the web page. Since they set a search range and collect data, wrapper-based crawlers are faster and more accurate than general purpose/ focused crawlers. But they are unsuitable for collecting a plurality of different non-specific data [8].

2.4. Crawler4j

Open source Crawler4j is a crawler with a variety of functions. It can collect large amounts of data by widening the scope of the search like General crawlers and Focused crawlers, and anointing a target for accurate search just like topical crawlers. Accurate retrieval can also be appointed for the crawlers to search. However, the problem is it uses a lot of memory due to the excessive use of global variables, and unnecessary duplication occurs in collection. In addition, it is difficult to modify that the amount of crawler source is massive and there are also frequent errors and stuttering occurrences.

3. Distributed Crawling System Design and Implementation

Distributed crawling system has wide search range, which is one of the disadvantages of General /Focused crawlers. In order to increase the accuracy, it uses a way of setting conditions for searching targets just like Topical crawlers do.

One of the disadvantages of Topical crawlers is that a fewer amount of data is collected due to narrow search range. Distributed crawling system solves this problem by widening the search range just like General crawlers, resulting in a complementary system that overcomes the disadvantages of previously used crawlers. Crawler4j can perform all the functions of the general purpose crawler and Topical crawler. The structure of the system will be compared and described.

3.1. Distributed Crawling System Design

Problems with the Crawler4j are memory waste due to the excessive use of global variables and occurrence of unnecessary duplication during collection. Additionally, the amount of the crawler is massive that it is difficult to modify.
There are also frequent errors and dropouts. Distributed crawling system is focused to increase the running speed in order to reduce the memory usage by classifying the search string and the data collection module to solve the problems with the Crawler4j. We implemented two crawlers in Java to do so. First is the URL that is responsible for the crawler search string, and the second is responsible for data collection HTML crawler.

We have also implemented a database to store the data in an HTML crawler collects MySQL. Figure 1 is a flow diagram of a data processing Crawler4j. Operation process is the same as the data processing flow chart of the system proposed in Figure 2.

![Figure 1. Crawler4j Data Processing Flow](image1)

![Figure 2. Proposed System Data Processing Flow](image2)
3.2. Distributed Crawling System

3.2.1. URL Crawler Implementation

URL crawler refers to the results collected after the URL value corresponding to the desired search string exists a string, run screen of URL crawler is as shown in Figure 3, the execution result is shown in Figure 4.

![Figure 3. Crawler4j Data Processing Flow](image)

![Figure 4. Proposed System Data Processing Flow](image)

3.2.2. Build Database

A database to store the data collected by the crawler has three fields: Link, which is URL address, content to store HTML tags, and visited to determine whether there is a visit.
3.2.3. HTML Crawler Implementation

HTML crawler moves through the URL that is classified by the URL crawler, checks whether there is a visit, and collects data. HTML crawler collects data using multi-threading.

Distributed crawling system is shorter in length than Crawler4j sources and reduces the memory by less use of global variables. And the speed was increased through optimization of multi-threading and the robot.

In order to minimize the overlapping collection, by checking whether there is a visit, it determines the data to collect and establishes the extraction range setting through URL numeric control.

The users can specify tags to extract as well, thus unnecessary HTML tags are not collected.

Tags you want to extract can also be specified by users, and unnecessary HTML tags are not collected.

Data collection is checked and confirmed through the extensive set method, which presets users’ desired extension and loads the contents of the Notepad file that you have written before.

If a user has set the image file, the file can be collected through the modification of the contents of the notepad. Figure 5 and 6 show the details of the URL stored in the crawler screen and Database while collecting the data.

```
- http://mie.pcu.ac.kr/subnav
- http://mie.pcu.ac.kr/footer++
- http://mie.pcu.ac.kr/content.php?db=login
- http://mie.pcu.ac.kr/onclick++
- http://mie.pcu.ac.kr/content.php?db=w1_1+++++
- http://mie.pcu.ac.kr/content.php?db=w2_2+++++
- http://mie.pcu.ac.kr/content.php?db=w2_1 Теппература+++++
- http://mie.pcu.ac.kr/content.php?db=w3_1+++++
- http://mie.pcu.ac.kr/content.php?db=w2_3+ Теппература+++++
- http://mie.pcu.ac.kr/content.php?db=w2_3_2+++++
- http://mie.pcu.ac.kr/content.php?db=w3_6
- http://mie.pcu.ac.kr/content.php?db=w1_2+ Теппература
- http://mie.pcu.ac.kr/content.php?db=w3_7
- http://mie.pcu.ac.kr/content.php?db=w3_9+++++
- http://mie.pcu.ac.kr/board.php?db=w4_1
- http://mie.pcu.ac.kr/board.php?db=w5_2
- http://mie.pcu.ac.kr/board.php?db=w5_4
- http://mie.pcu.ac.kr/board.php?db=w5_4
```

**Figure 5. Data Collection Screen**

```
http://mie.pcu.ac.kr/contents
http://mie.pcu.ac.kr/footer
http://mie.pcu.ac.kr/img
http://mie.pcu.ac.kr/library.php?db=_login
http://mie.pcu.ac.kr/library/user_agree.php
http://mie.pcu.ac.kr/onclick=
http://mie.pcu.ac.kr/rel=
http://mie.pcu.ac.kr/submit
http://mie.pcu.ac.kr/subnav
http://mie.pcu.ac.kr/table
```

**Figure 6. URL Stored in Database**
3.3. System Experiment

Table 1 shows that the number of all URLs from URL1 and URL2 may determine that there is no change for two weeks. July 9 data was collected at once, and since then have confirmed that there is no data collection. For URL3, it is not updated by the data on July 20 but updated after July 21, it can be seen that new data has been collected. There is no duplicate collection. Only if the new data is updated, data collection happens. Thus, the integrity of data is confirmed.

Table 1. Two Weeks Collection URL Results

<table>
<thead>
<tr>
<th>Date</th>
<th>URL1</th>
<th>URL2</th>
<th>URL3</th>
</tr>
</thead>
<tbody>
<tr>
<td>7/9</td>
<td>170</td>
<td>53</td>
<td>1163</td>
</tr>
<tr>
<td>7/10</td>
<td>170</td>
<td>53</td>
<td>1163</td>
</tr>
<tr>
<td>7/11</td>
<td>170</td>
<td>53</td>
<td>1163</td>
</tr>
<tr>
<td>7/12</td>
<td>170</td>
<td>53</td>
<td>1163</td>
</tr>
<tr>
<td>7/13</td>
<td>170</td>
<td>53</td>
<td>1163</td>
</tr>
<tr>
<td>7/14</td>
<td>170</td>
<td>53</td>
<td>1163</td>
</tr>
<tr>
<td>7/15</td>
<td>170</td>
<td>53</td>
<td>1163</td>
</tr>
<tr>
<td>7/16</td>
<td>170</td>
<td>53</td>
<td>1163</td>
</tr>
<tr>
<td>7/17</td>
<td>170</td>
<td>53</td>
<td>1163</td>
</tr>
<tr>
<td>7/18</td>
<td>170</td>
<td>53</td>
<td>1163</td>
</tr>
<tr>
<td>7/19</td>
<td>170</td>
<td>53</td>
<td>1163</td>
</tr>
<tr>
<td>7/20</td>
<td>170</td>
<td>53</td>
<td>1163</td>
</tr>
<tr>
<td>7/21</td>
<td>170</td>
<td>53</td>
<td>1164</td>
</tr>
<tr>
<td>7/22</td>
<td>170</td>
<td>53</td>
<td>1164</td>
</tr>
</tbody>
</table>

4. Conclusion

The main object of the crawler is that the user quickly/accurately collects the desired data. Each user wants different data, and has a different purpose for collection. For this reason, if the individual uses crawlers, not searching engine of enterprises, they use open source crawlers or crawlers they build up.

Distributed crawling system proposed in this paper aims to improve problems caused by Crawler4j. When it operates one time, it uses a lot of functions. By dividing the main features externally, the operating system was made.

Server overhead was minimized by reducing the number of global variables to reduce memory usage and limiting HTML tags to prevent indiscriminate internal collection.

In addition, users can set files that they want in order to write and collect as text forms and thus modify them in real time. Future studies are needed to verify the efficiency in various fields, including improvement of the usability of the proposed system and studies of module extension to a generalized system.

References


Authors

**Han Gil Kim**, he received the B.S., M.S. degrees from the Department of Electronic Engineering of Hanbat National University, Korea, in 2002, and 2011, respectively. Since 2005, he worked for the Department of Music & Sound Technology at Korea University of Media Arts as a professor. He is currently a Doctorate course in Department of Computer Engineering of Paichai University. His current research interests include information retrieval system, web 2.0, and the semantic web.

**Jong Won Lee**, he received the B.S degree from the Department of Computer Engineering of Paichai University, Korea in 2013. He is currently a Master course in Department of Computer Engineering of Paichai University. His current research interests include multimedia information processing, information retrieval system, and semantic web.

**Tae Hak Ban**, he received the B.S degree from the Department of Computer Engineering of Paichai University, Korea in 2011, and he received the M.S degree from the Department of Computer Engineering of Paichai University, Korea in 2013. He is currently a Doctorate course in Department of Computer Engineering of Paichai University. His current research interests include multimedia information processing, multimedia communication, and ubiquitous computing.

**Hoe Kyung Jung**, he received the B.S degree in 1987 and Ph. D. degree in 1993 from the Department of Computer Engineering of Kwangwoon University, Korea. From 1994 to 2005, he worked for ETRI as a researcher. Since 1994, he has worked in the Department of Computer Engineering at Paichai University, where he now works as a professor. His current research interests include multimedia document architecture modeling, information processing, information retrieval, and databases.