https://doi.org/10.14400/JDC.2017.15.11.261

# 웹 크롤링 이용한 크레페 검색 시스템 설계

김효종*, 한군희**, 신승수*
동명대학교 정보보보학과*, 백석대학교 정보통신학부**

# Crepe Search System Design using Web Crawling

**Hyo-Jong Kim*, Kun-Hee Han**, Seung-Soo Shin***
**Department of Information Security, Tongmyong University***
**Division of Information & Communication Engineering, Baekseok University****

요 약 본 연구의 목적은 광역 네트워크로 연결된 다수의 봇을 활용한 방식이 아닌 단일 네트워크에서 정보의 최신성을 보장하기 위해 데이터베이스 서버를 사용하지 않고 실시간으로 웹에 접속하여 정보를 불러오는 방식을 사용한 검색 시스템을 설계하는 것이다. 연구의 방법은 크레페 시스템에서 신속하고 정확한 인물과 키워드 검색을 할 수 있는 시스템을 설계하고 분석한다. 크레페 서버는 본문 태그 매칭 변환 과정은 사용자가 정보를 등록할 경우 글자체, 글자 크기, 색상등과 같이 사용자마다 여러 스타일이 적용되어 그 자체가 정보가 되기 때문에 모든 정보를 그대로 저장하게 된다. 크레페 서버는 본문 태그 매칭 문제점이 발생되지 않는다. 그러나 크레페 검색 시스템을 실행할 때에는 사용자들의 스타일 및 특성을 정형화할 수 없다. 이러한 문제점을 html_img_parser 함수와 Go언어의 html 파서 패키지를 사용하면 해결할 수 있다. 특정 사이트를 대상으로 하는 웹 크롤러 설계가 아닌 범용 웹 크롤러에 큐와 다중 스레드를 적용하여 다양한 웹 사이트를 빠르고 효율적으로 탐색, 수집한 빅 데이터를 다양한 응용 분야에 활용될 수 있을 것이다.

주제어 : 디지털 큐레이션, 콘텐츠, 웹 크롤러, 검색 시스템, 키워드 검색, 모듈

**Abstract** The purpose of this paper is to provide a search system using a method of accessing the web in real time without using a database server in order to guarantee the up-to-date information in a single network, rather than using a plurality of bots connected by a wide area network Design. The method of the research is to design and analyze the system which can search the person and keyword quickly and accurately in crepe system. In the crepe server, when the user registers information, the body tag matching conversion process stores all the information as it is, since various styles are applied to each user, such as a font, a font size, and a color. The crepe server does not cause a problem of body tag matching. However, when executing the crepe retrieval system, the style and characteristics of users can not be formalized. This problem can be solved by using the html_img_parser function and the Go language html parser package. By applying queues and multiple threads to a general-purpose web crawler, rather than a web crawler design that targets a specific site, it is possible to utilize a multiplier that quickly and efficiently searches and collects various web sites in various applications.

**Key Words :** Digital Curation, Contents, Web Crawler, Search system, Keyword search, Module

Received 2 October 2017, Revised 2 November 2017
Accepted 20 November 2017, Published 28 November 2017
Corresponding Author: Seung-Soo Shin
(Tongmyong University)
Email: shinss@tu.ac.kr

ISSN: 1738-1916

## 1. 서론

ICT(Information Communication Technology)의 발달로 인해 인터넷상의 정보는 급격하게 생성되고, 다양한 분야에 걸쳐 디지털 정보자원에 대한 서비스가 이루어지고 있다. 정보통신 기술을 접목한 교육을 통하여 효율적인 교수법이 제시되고, 학생들의 자기 주도 학습법이 가능하게 되고 있다[1,2,3].

기존의 ICT를 접목한 다양한 교육 중 생성되는 활동자료, 수업교재 등 여러 가지 데이터들은 매번 사용자들의 수와 그에 대한 정보들로 비례하여 그 양은 방대해진다. 이런 방대한 데이터들은 빅 데이터라고 불리며 일정한 데이터양과 주기에 따라 분류 및 관리를 하여 디지털 큐레이션을 구성할 수 있다. 디지털 큐레이션은 교육 사이트에 저장된 데이터를 기반으로 분석하고 분석된 정보는 학습법 연구의 중요성을 높일 수 있다[4].

정보통신기술 발달과 함께 도래한 지식정보사회는 급변하는 사회에서 창의성과 인성은 21세기 미래 교육에서 매우 중요한 역량으로 강조되고 있다. 디지털 큐레이션을 활용한 ICT 활용 교육으로 미디어를 이용하는 다양한 활동을 통해 창의적 인성교육이 활발히 진행되고 있다. 이 과정에서 생성되는 여러 종류의 비정형화 데이터들을 디지털 큐레이션 시스템을 적용하여 수집, 보존, 아카이빙, 제공하여 교수자들과 학습자들에게 시각적, 조작적으로 편의를 제공 할 수 있는 시스템을 구축했다[5].

창의・인성 교육기반의 디지털 큐레이션 시스템은 작성자의 성취와 성공에 대하여 보다 다양한 표현이 가능하므로 작성자의 독자적으로 하여금 보다 깊은 통찰을 가능하게 한다. 또한 콘텐츠의 공유 및 수정, 확장 등이 용이하다는 장점을 가지고 있다. 창의・인성 교육기반의 디지털 큐레이션 시스템(http://www.crepe.kr/)을 구축하고, 디지털 큐레이션 시스템에 대한 웹 공격의 유형과 취약점을 분석했다[6].

이러한 시스템들의 정보 검색 향상을 위해 고성능 웹 크롤러의 중요성이 부각되고 있다. 인터넷에서 제공하는 수많은 웹 문서를 일정한 간격마다 자동으로 탐색하고 수집하는 기술을 웹 크롤러(Web Crawler)라 한다[7]. 현재 상용중인 웹 크롤러는 주제별 웹 크롤러, 래퍼기반 웹 크롤러, 범용 웹 크롤러 등으로 분류하고 검색엔진의 상황에 따라 웹 크롤러를 선택한다. 검색엔진들은 사용자가 정보를 요청할 경우 웹 크롤러에서 가공된 데이터를 저장한 데이터베이스에서 색인과정을 통해 재구성되어 효율적으로 정보를 보여준다. 이러한 정보들을 탐색·수집·가공하는 웹 크롤러들은 다양한 데이터 처리방식 중에서 대부분 분산처리를 기반으로 동작한다[8].

본 논문에서는 디지털 큐레이션 시스템[6]을 대상으로 하는 웹 크롤러가 아닌 범용 웹 크롤러에 큐와 다중 스레드를 적용하여 빠르고 효율적으로 탐색, 수집 그리고 분석한 결과를 학습자들에게 편의를 제공 할 수 있는 크레페 검색 시스템을 설계하고 구현한다.

본 논문의 구성은 다음과 같다. 2장에서는 디지털 큐레이션, 웹 크롤링에 관한 기존연구를 분석하고, 3장에서는 크레페 검색 시스템을 설계하고 구현한다. 그리고 4장에서는 분석하고, 마지막으로 5장에서 결론을 맺는다.

## 2. 관련연구

### 2.1 디지털 큐레이션

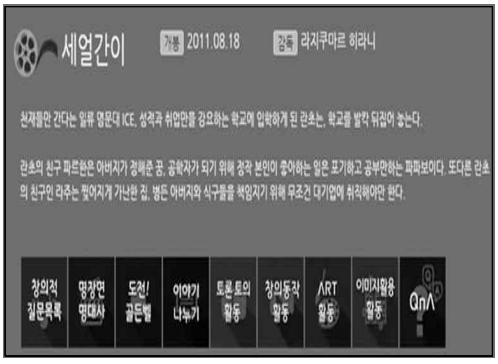
디지털 큐레이션은 디지털 자원을 수집, 보존, 아카이빙, 제공하는 것을 지칭하는 것으로 정보의 결핍에서 과잉으로 흐르는 과정에서 필연적으로 나타난다[9]. 이렇게 과잉 생산된 정보에 대해 디지털 큐레이션은 질적으로 양호한 정보와 적절한 내용을 제공한다. 최근 큐레이션과 정보기술이 융합된 연구가 활발히 진행되고 있는데 콘텐츠 큐레이션, 소셜 큐레이션, 디지털 큐레이션 등으로 이름들도 다양하게 사용되고 있는데, 비슷한 의미를 가지고 있으므로[10, 11], 본 논문에서는 디지털 큐레이션으로 지칭하도록 한다. 특히 큐레이션이 가능하도록 하는 도구로 웹 서핑을 통해 큐레이션하고 싶은 이미지나 텍스트를 클리핑하여 자료를 정리하고 보여주는 기술이 중요하게 대두되고 있다[12]. 디지털 큐레이션의 대상은 기능별로 메일서비스, 블로그, 쇼핑, 뉴스 등 여러 가지 종류가 있다. 디지털 큐레이션 시스템은 학습자간의 의견교환이 쉽고 디지털 자료들을 보기 쉽게 전시해주는 것은 성공하였지만 아직 개선의 여지가 남아있다. 개선의 일환으로 학습자마다 작성한 개인적인 글을 체계적으로 정리하여 자신의 포트폴리오로 출력할 수 있도록 개선이 필요하다. 그리고 모바일 환경에서도 창의・인성 활동이 가능하도록 확장하여 학습자들이 더욱 편리하게

이용할 필요가 있다.

### 2.2 크레페 시스템

지식정보사회로 급변하는 사회에서 창의성과 인성은 21세기 교육에서 매우 중요한 역량으로 강조되고 있다. 창의성은 글로벌 시대 핵심 역량으로 지속적으로 꾸준히 강조된 개념이었는데, 최근 이러한 창의성과 인성의 두 교육의 유기적 결합을 통해, 인성개발이 곧 창의성 개발로 이어지는 상호동반 효과에 대한 통합적 논의가 활발하게 진행되고 있다. 기존의 디지털 큐레이션 서비스 제공 사이트를 분석하여 각각의 특성을 파악하고, 창의・인성 교육에 확대・적용, 분석하여 창의・인성 교육기반의 디지털 큐레이션 시스템 구축이 절실히 요구된다.

창의・인성 교육을 목적으로 만든 디지털 큐레이션[6] 사이트에서 전문가의 필터링을 받아 5개의 콘텐츠를 8가지의 주제에 따라 교육하는 시스템이 구축되었다. 각각의 콘텐츠에 대하여 창의・인성 교육에 적합한 8가지의 활동을 수행하도록 [Fig. 1]과 같이 구성되어 있다.



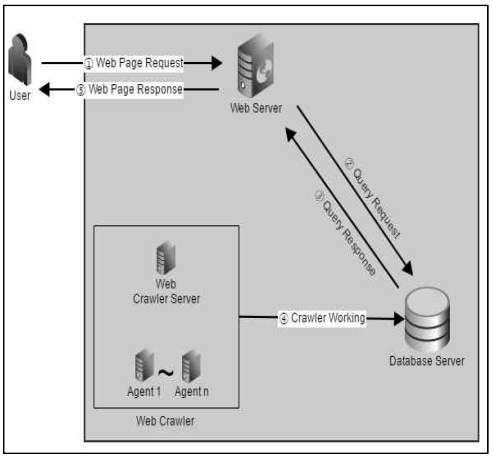

[Fig. 1] Sub Topic Menu

디지털 큐레이션 시스템에서 콘텐츠를 감상한 후 가장 먼저 수행하는 사전 활동은 “창의적 질문 목록”과 “명장면&명대사”, “도전 골든벨”이 있다. 그리고 수업과 동시에 진행되는 본시 활동으로 “이야기 나누기”, “토론・토의 활동”, “Art활동”, “창의동작 활동”이 있다. 사후 활동으로는 이미지 활용한 “이미지 활용 활동”이 있으며 그 외의에는 질문을 할 수 있는 “Q&A”가 있다[5]. 8가지의 활동 중 첫 번째로 창의적 질문 목록은 창의・인성 교육의 사전 활동으로 학생들이 미디어를 시청한 이후에 미디어의 내용에 대해 여러 가지 방식으로 접근하기 위하여 창의적 질문을 하는 활동이다. 두 번째로 명장면&명대사는 창의・인성 교육의 사전활동으로 미디어를 시청하면서 가장 인상 깊었던 장면이나 대사, 문구를 캡처한 그 장면에 대한 설명과 그 장면이 왜 인상 깊었는지에 대해 서로 논의하는 활동이다. 세 번째로 도전 골든벨은 미디어를 보고난 뒤 교수가 미리 등록해 둔 문제들을 풀어보는 활동이다. 네 번째로 이야기 나누기는 교수가 대화방을 만들어 대화 주제를 정해주면 학생들이 활동들에 대한 이야기를 하는 활동이다. 다섯 번째로 토론 토의는 학생들이 활동들에 대한 각자 자신의 생각을 토론하는 활동이다. 여섯 번째로 ART활동은 미디어의 내용과 관련되어 그림을 그려 창의적으로 표현하는 활동이고 일곱 번째의 창의동작 활동은 ART활동의 그림을 토대로 직접 만들어보거나 체험해보는 활동이다. 여덟 번째로 이미지 활용 활동은 교수가 미디어에서 골라낸 8개의 이미지를 이용해 학습자들에게 순서와 상관없이 상황에 맞는 이야기를 만들어내는 활동이다. 그 외의 활동인 Q&A는 학생들이 궁금한 점이나 알고 싶은 것을 질문하여 교수가 그 질문에 답하는 활동이다.

### 2.3 웹 크롤러

웹 크롤러는 인터넷에서 많은 양의 웹 문서를 일정한 간격으로 탐색, 수집, 가공, 저장한다. 인터넷에 연결된 서버를 대상으로 일정한 주기로 탐색하고. 사용자가 필요로 하는 데이터를 분석하여 비정형화된 데이터로 수집한다. 비정형화로 수집된 데이터를 정형화된 데이터로 일정한 규칙에 의해 작업을 하여 가공한 후, 정형화된 데이터를 데이터베이스에 저장하여 사용자가 검색을 요청할 경우 색인과정을 통해 정보를 제공한다[13,14,15].

웹 크롤러는 웹 서버, 데이터베이스 서버, 웹 크롤러로 구성된다. 웹 크롤러는 [Fig. 2]과 같은 순서로 동작한다. ①의 동작은 사용자 웹 서버에게 웹페이지 요청을 한다. ②의 동작은 사용자로부터 수신 받은 웹 서버는 요청이 정상일 경우 데이터베이스 서버로 정보를 요청한다. ③의 동작은 웹 서버로부터 수신 받은 정보를 기반으로 데이터베이스 색인 과정을 통해 정보를 검색한다. ④의 동작은 데이터베이스 서버에서 색인과정을 마친 정보를 웹 서버로 전송한다. ⑤의 동작은 데이터베이스 서버로부터 수신한 정보를 가공하여 사용자에게 전송한다[4].



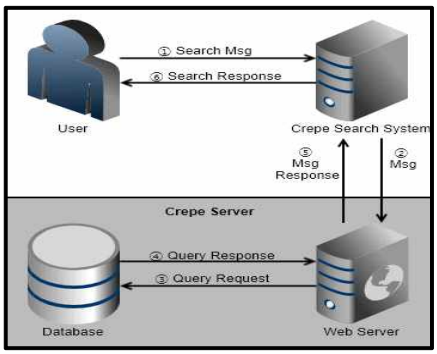

[Fig. 2] Basic principles of operation of the Web crawler

## 3. 크레페 검색 시스템

본 장에서는 정보의 최신성을 보장하기 위해 데이터베이스 서버를 사용하지 않고, 실시간으로 웹에 접속하여 정보를 불러오는 방식을 사용한 크레페 검색 시스템을 설계하고 구현한다.

### 3.1 시스템 구성

웹 크롤링을 이용한 크레페 검색 시스템은 User, Crepe Search System, Web Server, Database로 구성된다. 사용자가 크레페 검색 시스템으로 검색어를 질의하면 사용자로부터 수신한 검색어를 유효 검색어 판별 결과에 따라서 유효하지 않은 검색어일 경우는 사용자에게 에러 메시지를 송신한다. 유효한 검색어일 경우 크레페 검색 시스템은 웹 서버로 해당 사용자에 대한 정보를 요청한다. 웹 서버는 수신한 정보의 권한에 대한 유효성을 검사하고 크레페 서버의 데이터베이스로 Query를 요청한 해당 정보에 대한 색인 과정을 수신 후 JSON (JavaScript Object Notation) 형태로 가공하여 크레페 검색 시스템으로 송신한다. 웹 서버로부터 사용자에 대한 정보를 수신한 크레페 검색 시스템은 동적 웹 문서를 생성하여 사용자에게 송신한다. 이와 같은 과정은 [Fig. 3]과 같다.




[Fig. 3] Crepe Search System

### 3.2 검색 시스템 설계

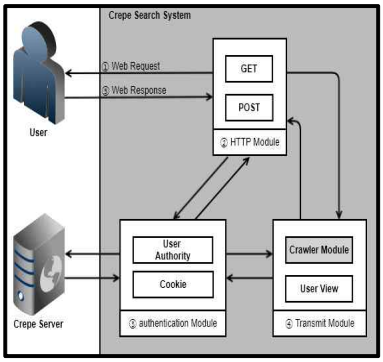
크레페 검색 시스템은 HTTP Module, Authentication Module, Transmit Module로 구성된다. 각각의 구성 별 기능은 다음과 같다. 먼저, HTTP Module은 GET 방식과 POST 방식으로 구성된다. GET 방식은 보안적인 요소와 관련이 없는 웹 페이지를 불러오거나 조회용으로 사용되고, 사용자가 서버에게 요청하는 파라미터의 길이가 제한적이며, 외부로 노출되는 방식이다. 그러나 보안적인 요소가 필요하고 데이터량의 제한이 없으며 사용자가 서버에게 파라미터를 송신할 경우 외부로 노출되지 않고 송신이 가능한 것이 POST 방식이다. 그리고 GET 방식보다 보안을 보장하지만 수동적 공격과 능동적 공격에는 취약하다. 이러한 취약점을 보완하기 위해 웹 서버는 SSL(Secure Sockets Layer) 또는 TLS(Transport Layer Security)를 사용하여 사용자와 서버간의 통신을 암호화한다.

Authentication Module은 관리자 권한과 사용자 권한이 있다. 관리자 권한은 모든 사용자가 작성한 게시글을 조회 가능하고 정보를 공유 할 수 있다. 그리고 사용자 권한은 본인이 작성한 게시글을 모두 확인이 가능하다. 본 논문에서 Authentication Module은 관리자 권한만 대상으로 한다.

Transmit Module은 사용자로부터 수신한 검색 옵션 및 검색어를 웹 서버에게 송신한다. 웹 서버는 해당 정보에 대한 권한을 검증 한 뒤, JSON 형태로 변환 후 크레페 검색 시스템에게 송신한다. 수신한 JSON을 웹 문서 형식으로 생성하여 크레페 검색 시스템은 사용자에게 송신한다. JSON은 속성과 값을 쌍으로 이루어진 정보를 담은

개방형 표준 포맷으로 대부분 서버에서 사용자로 수신할 때 사용하는 형태이다.

크레페 검색 시스템의 흐름은 다음과 같다. 사용자는 Authentication Module에서 크레페 검색 시스템으로부터 검색 권한이 주어지면, HTTP Module에서 검색 페이지를 수신한다. 사용자는 검색 페이지의 검색 옵션을 선택 후, 이름, 학번 등을 검색하면 POST 방식으로 검색어와 검색옵션을 크레페 검색 시스템으로 송신한다. 크레페 검색 시스템은 수신한 정보를 HTTP Module에 의해 파라미터를 분류하고 검색결과를 사용자에게 송신한다. 이러한 과정은 [Fig. 4]과 같다.




[Fig. 4] Modules of Crepe Search System

### 3.3 Crawler Module

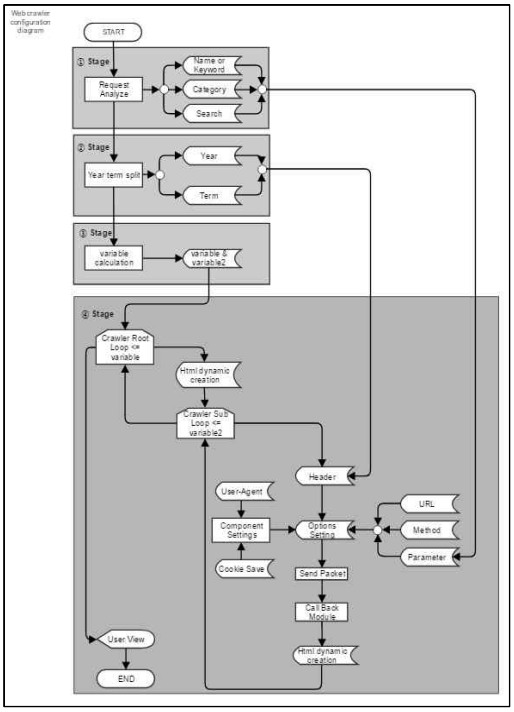
Transmit Module 내부에 있는 "Crawler Module"은 Request Analyze, Year term split, variable calculation, Crawler Root Loop로 구성된다. 크롤러 모듈에 있는 전송 모듈 내부의 각각의 기능은 다음과 같다.

첫 번째, Request Analyze는 사용자의 검색옵션인 이름 or 키워드, Category, Search로 분류한다. 이름 or 키워드는 사용자의 검색옵션을 이름 또는 키워드로 선택한다. Category는 사용자가 선택한 검색옵션을 지정하여 속도와 정확도를 높인다. Search는 사용자가 입력하는 검색어를 받는 변수로 입력할 경우 동작한다.

두 번째, Year term split는 사용자가 지정한 기간을 분석하여 연도와 학기를 나누어 변수로 저장한다.

세 번째, variable calculation는 Crepe Search System가 알고리즘에 의해 페이지 접속 횟수를 동적으로 계산하며 다음의 수식을 활용해서 값을 확인할 수 있다. x는 Topic의 개수이고, y는 전체 게시글의 개수를 의미한다. x의 값은 x <= y 이고, y의 값은 x >= y을 대입한 뒤, 올림을 하여 값을 나타낸다. 이후 계산한 값을 변수에 할당한다.

$$result = \lceil (y/x) \rceil$$




[Fig. 5] Crawler Module

네 번째, “Crawler Root Loop <= variable”는 크레페 검색 시스템의 Root 반복문이 동작한다. Root 반복문은 다른 페이지의 게시글을 조회하고 URL을 생성한다. Crawler Root Loop내의 “Crawler Sub Loop <= variable2”는 접속할 게시글에 대한 “HTTP Header”를 구성하고 "Options Setting"에 User-Agent, URL, Method, Parameter를 적용한다. User-Agent는 웹 서버에서 사용자의 디바이스 종류를 분류할 때 사용하는 정보이다. 이러한 디바이스 정보에 따라서 모바일과 PC로 분류되며, 본 논문에서는 디바이스 정보를 PC로 지정한다. URL은 접속할 게시글의 고유번호를 알고리즘에 의해 URL 형태로 생성하여 변수에 저장하고, POST 방식 또는 GET 방식을 지정한다.

그리고 Parameter은 “Request Analyze”에서 분류한

정보를 적용한 후, "Send Packet"으로 웹 서버에게 송신하고 "Call Back Module"으로 응답을 수신한다. 수신한 정보를 기반으로 html을 구성하고 모든 Root 반복문과 Sub 반복문이 정상적으로 종료되면 사용자에게 페이지를 송신한다. 이와 같은 과정은 [Fig. 5]과 같다.

## 4. 분석

본 논문에서 제안한 크레페 검색 시스템을 비교 분석한다.

### 4.1 구현 환경

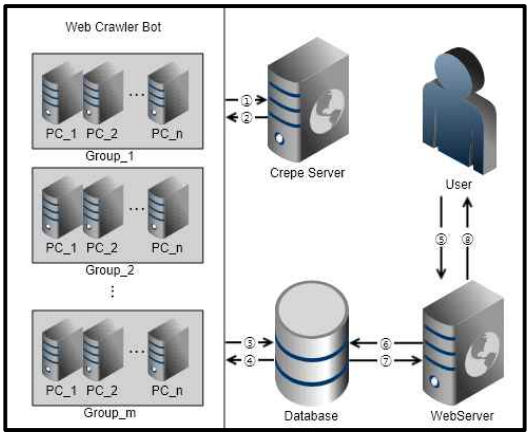
제안한 크레페 검색 시스템의 구현환경은 App, OS, Basic Application으로 구성된다. 소프트웨어는 GoLang, Windows 10 Pro을 사용하고, H/W Basic Application으로는 Apache, GO HTTP, PHP, MySQL, Go를 이용한다. 시스템 구현환경은 <Table 1>와 같다.

<Table 1> System Implementation Environment

| System Performance by Program | | | | |
|---|---|---|---|---|
| App | DB Server Web Server | | CPU | Intel(R) Core(TM) i7-4790 |
| | | | RAM | 8.00GB |
| | JSA | | CPU | Intel(R) Core(TM)2 Duo |
| | | | RAM | 4.00GB |
| | JCA | 1 | CPU | Intel(R) Core(TM) i5-6300U |
| | | | RAM | 8.00GB |
| | | 2 | CPU | Intel(R) Core(TM) i5-2320 |
| | | | RAM | 4.00GB |
| OS | Windows 10 Pro | | | |
| Basic Application | | | | |
| Apache , Go HTTP | | | PHP | MySQL Java |

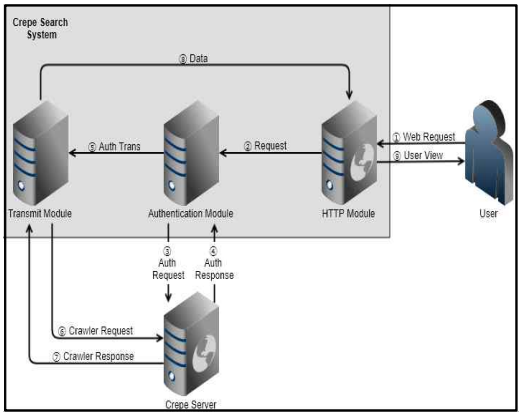
### 4.2 웹 크롤링 분석

웹 크롤링 시스템과 제안한 실시간 크레페 검색 시스템을 비교한다. 기존의 웹 크롤링 시스템은 데이터 수집하고 변환한 결과를 저장하는 단계로 진행되며, 시스템은 웹 크롤러, 웹 서버, 데이터베이스로 구성된다. 먼저, 웹 크롤러는 일정한 기간 동안 크레페 서버를 탐색하여 수집된 비정형 데이터를 정형화 데이터로 변환하여 데이터베이스에 저장한다. 그러면, 웹 서버는 사용자의 요청이 정상적인 경우, 해당 검색어를 데이터베이스로 송신하면 색인과정을 통하여 웹 서버로부터 수신한 검색어를 사용자에게 정보를 송신한다. 기존의 웹 크롤링 시스템의 흐름 과정은 [Fig. 6]과 같다.




[Fig. 6] Web crawling system

본 논문에서 제안하는 크레페 검색 시스템은 데이터를 수집한 정보를 변환하여 출력하는 단계로 진행된다.




[Fig. 7] Crepe search system

크레페 검색 시스템은 HTTP Module, Authentication Module, Transmit Module로 구성된다. HTTP Module은 사용자의 요청을 분석하여 정상적인 요청인 경우 Authentication Module로 검색어를 송신한다. 그러면,

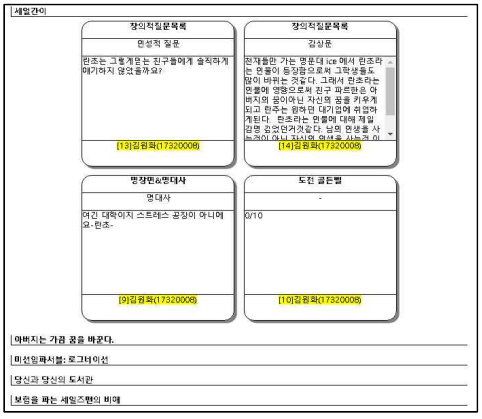
HTTP Module로부터 검색어를 수신한 Authentication Module은 Crepe System으로 검색 권한을 인증하고 Transmit Module로 웹 크롤링을 한다. 이후 Transmit Module은 HTTP Module을 통해 사용자에게 정보를 송신한다. 크레페 검색 시스템의 과정은 [Fig. 7]과 같다.

[Fig. 7]은 [Fig. 6]에서 사용하는 데이터베이스, 웹 서버 등을 사용하지 않기 때문에 실시간으로 사용자에게 정보를 제공함으로서 검색의 효율성을 높일 수 있다.

### 4.3 검색 시스템 분석

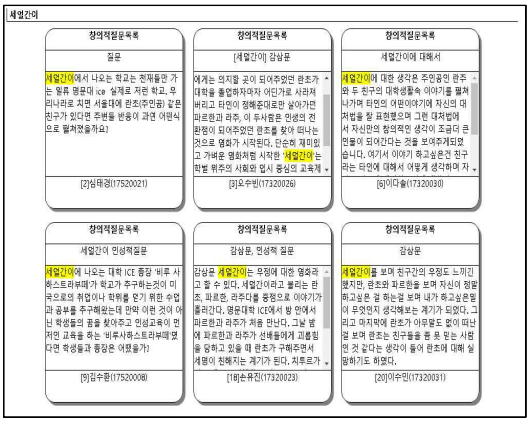
#### 4.3.1 검색 방법

웹 검색 시스템은 사용자가 웹 서버에게 검색어를 요청할 경우 웹 서버는 검색어를 재가공하여 데이터베이스 서버에게 전송한다[13]. 전송받은 데이터베이스 서버는 분류 및 저장된 정보 중에서 색인과정을 통해 정보를 사용자에게 보여준다. 크레페 검색 시스템은 인물과 키워드에 대한 검색이 있다. 인물 검색은 크레페 검색 시스템에서 학번과 이름을 입력하면 크레페 서버로부터 사용자의 정보를 받아 크레페 검색 시스템에서 필터링·색인·변환을 한 뒤 사용자에게 웹페이지를 생성하여 송신한다. 크레페 검색 시스템에서 “김**”에 대한 인물 검색을 사용하여 검색한 결과는 [Fig. 8]과 같다.




[Fig. 8] People search method result

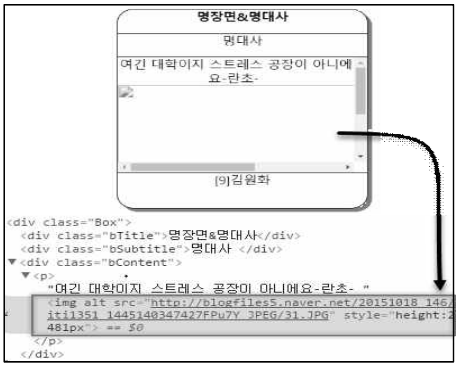
키워드 검색은 크레페 검색 시스템에서 사용자가 입력한 키워드 정보에 대한 색인 및 변환을 한 뒤 웹 페이지를 사용자에게 송신한다. 검색 조건은 사용자가 입력한 키워드를 기준으로 하며 사용자의 관점에서 신속한 정보를 확인할 수 있다. 크레페 검색 시스템에서 “세 얼간이”에 대한 키워드 검색을 사용하여 검색한 결과는 [Fig. 9]과 같다.




[Fig. 9] Keyword search results

#### 4.3.2 문제점 해결

크레페 검색 시스템에서 크레페 서버로부터 데이터를 불러올 경우, 비정형화된 이미지 주소의 검증, 본문 태그 매칭에 대한 문제점 등이 발생한다. 사용자가 모든 시스템에서 정보를 등록할 경우에는 정보와 이미지 파일이 정상적으로 저장되어야 한다. 그러나 비정형화된 이미지 주소 검증은 크레페 서버에서 이미지 파일이 정상적으로 저장되지 않아서 발생한다. 비정형화된 이미지 주소 화면은 [Fig. 10]과 같다.




[Fig. 10] Non-standardized image tag address

크레페 검색 시스템에서 본문 태그 매칭 변환 과정은 사용자가 정보를 등록할 경우 글자체, 글자 크기, 색상등과 같이 사용자마다 여러 스타일이 적용되어 그 자체가 정보가 되기 때문에 크레페 서버는 모든 정보를 그대로 저장하게 된다. 크레페 서버는 본문 태그 매칭 문제점이 발생되지 않는다. 그러나 크레페 검색 시스템을 실행할 때에는 사용자들의 스타일 및 특성을 정형화할 수 없다. 본문 태그 매칭 화면은 [Fig. 11]와 같다.


[Fig. 11] Body tag matching

위에서 제시된 문제점을 html_img_parser함수와 Go 언어의 html 파서(golang.org/x/net/html) 패키지를 사용하면 해결할 수 있다.

html_img_parser 함수는 크레페 서버로부터 수신한 html문을 정형화하여 웹 크롤러 서버에 적합한 형태로 변환한다. 함수의 기능은 html에서 정상적인 이미지 출력이 가능한 태그만 추출한다. 이미지 태그를 추출하는 의사코드는 loop함수와 for문을 사용한다. loop함수는 비정형 문자열 중에서 img 태그를 검색하는 재귀함수이고 for문은 loop 재귀함수를 호출하는 역할이다. loop함수 안에 if(html.Type == html.ElementNode && html.Data == "img")이 조건이다. if문은 비정형화된 문자열 중에서 img태그가 존재할 경우 이미지 태그 주소를 임시변수에 저장한다. 이후 비정형 문자열이 존재 하지 않을 경우 for문을 종료한다.

html 파서 패키지는 크레페 서버로부터 수신한 html문을 정형화하여 웹 크롤러 서버에 적합하게 변환한다. 패키지의 기능은 불필요한 태그를 제외하고 사용자가 입력한 정형화된 문자열만 추출한다. 정형화된 문자열만 추출하는 의사코드는 while문 안에 if(html.Next() == html.ErrorToken)이 조건이다. if문은 비정형화된 문자열이 while문에서 검색될 경우 “html.Text()” 메소드로 문장을 임시변수에 저장한다. 이후 비정형화된 문자열이 존재하지 않을 경우 “ErrorToken”을 반환한다. 반환된 ErrorToken과 Text의 값이 일치할 경우 break문을 통해 while문을 종료한다.

## 5. 결론

디지털 큐레이션은 디지털 자료를 전시하고 보여 주는 것과 동시에 자료들을 분류, 정리하고 다시 사용할 수 있도록 지원해주는 것을 말한다. 본 논문에서는 창의·인성 증진을 위한 교육용 디지털 큐레이션 시스템에서 정보의 최신성을 보장하기 위해 데이터베이스 서버를 사용하지 않고 실시간으로 웹에 접속하여 정보를 불러오는 방식을 사용한 검색 시스템을 설계했다. 광역네트워크로 연결된 검색시스템이 아닌 크레페 검색 시스템에서 인물과 키워드에 대한 검색으로 이루어진다. 데이터베이스, 웹 서버 등을 사용하지 않기 때문에 실시간으로 사용자에게 정보를 제공함으로서 검색의 효율성을 높일 수 있다. 향후에는 특정 사이트를 대상으로 확대하여 빠르고 효율적으로 탐색, 수집 그리고 분석한 결과를 빅 데이터 응용 분야에 활용될 수 있을 것이다.

## REFERENCES

[1] Jung-In Kim, Byung-Man Kim, Jung-Ju Kim, “A Development of Digital Curation System for Creativity and Personality Education”, Journal of Korea Multimedia Society, Vol. 19, No. 9, pp.1710-1722, 2016.

[2] Young-Hee Ahn, Ok-Wha Park, “Development of a Framework for Digital Curation Policy”, Journal of Korean Library and Information Science Society, Vol 41, No. 1, pp.167-186, 2010.

[3] Kang Soon Lee, “Development of Elementary Dance

Education Program Using ICT", Korean Society For The Study Of Physical Education, Vol. 18, No. 2, pp.77-89, 2013.

[4] H.K. Kim, Digital Curation Framework Research for Analyzing Issues Based on Big- Data, Master's Thesis of Chung-Ang University of Technology, 2014.

[5] Jung-In Kim, Byung-Man Kim, Jung-Ju Kim, "A Development of Digital Curation System for Creativity and Personality Education", Journal of Korea Multimedia Society, Vol. 19, No. 9, pp. 1710-1722, 2016.

[6] S.S. Shin, J.I. Kim, and J.J. Youn, "Vulnerability Analysis of the Creativity and Personality Education Based on Digital Convergence Curation System," Journal of Korea Convergence Society, Vol. 6, No. 4, pp.225-234, 2015.

[7] Kwang-Young Kim, Won-Goo Lee, Hwa-Mook Yoon, Sung-Ho Shin, Min-Ho Lee, "Development of Web Crawler for Archiving Web Resources," Journal of the Korea Contents Association, Vol. 11, No. 9, pp.9-16, 2011.

[8] Wan-Sup Cho, Jeong-Eun Lee, Chi-Hwan Choi, "Refresh Cycle Optimization for Web Crawlers," Journal of the Korea Contents Association, Vol. 13, No. 6, pp.30-39, 2013.

[9] N.E. Han and S.H. Kim, "Comparative Analysis on Digital Curation Process in Foreign Academic Libraries," Journal of Korean Library and Information Science Society, Vol. 45, No. 2, pp. 93-116, 2014.

[10] H.H. Lee and W.J. Lee, "A Study on the Design of Curation System of Customized Sport Convergence Contents for Activation of Sport for All," Journal of Korea Multimedia Society, Vol. 19, No. 2, pp. 396-404, 2016.

[12] B.H. Cho, "The Trend of Digital Curation Service," Week Technology Trends, Vol. 2013, No. 42, pp. 1-10, 2013.

[13] Myoung-sil Choi , "A Study on the Improvement of the Web-Crawler Performance based on Weighted Directed Graph," Department of Computer Science, Graduate School, Kyungpook National University, 2010.

[14] Dae Yu Kim, Jung Tae Kim, "Efficient Design of Web Searching Robot Engine Using Distributed Processing Method with Javascript Function," The journal of the Korea Institute of Maritime Information & Communication Sciences, Vol. 13, No. 12, pp.2595-2602, 2009.

[15] Kwang Hyun Kim, Joon Ho Lee, "A Methodology for Performance Evaluation of Web Robots," Information Processing Society, Vol. 11, No. 3, pp.563-570, 2006.

김 효 종(Kim, Hyo Jong)

- 2016년 2월 : 동명대학교 정보보호학과(공학사)
- 2017년 2월~ 현재 : 동명대학교 일반대학원 정보보호학과 석사과정
- 관심분야 : 웹 크롤링, IoT, 빅데이터 분석
- E-Mail : khj47561404@gmail.com

한 군 희(Han, Kun Hee)

- 2001년 3월 ~ 현재 : 백석대학교 정보통신학부 교수
- 관심분야 : 멀티미디어, 유비쿼터스, DB보안, 암호 프로토콜/알고리즘
- E-Mail : hankh@bu.ac.kr

신 승 수(Shin, Seung Soo)

- 2001년 2월 : 충북대학교 수학과(이학박사)
- 2004년 8월 : 충북대학교 컴퓨터공학과(공학박사)
- 2005년 3월 ~ 현재 : 동명대학교 정보보호학과 교수
- 관심분야 : 암호프로토콜, 무선 PKI, 네트워크 보안, U-헬스 케어, IoT
- E-Mail : shinss@tu.ac.kr