대학 홈페이지 활용도 향상을 위한 웹 로그 프로세스마이닝 분석

이용욱* · 최상현**

Web-log Process Mining Analysis for Improving Utilization of University Homepage

Yong Uook Lee* · Sang Hyun Choi**

Abstract

The purpose of operating the main homepage of University is to provide the related information about University resources to site visitors. In this study, we analyze website browsing patterns and extract characteristics of users in order to improve its utilization. The access log files to main homepage were used to analyze the browsing patterns and converted to process log files adaptable to a process mining tool, ProM. Finally we provide useful information about user friendly homepage design and suggest plans for improving its utilization to website operators.

Keywords : Process Mining, Web Log Analysis, Homepage Utilization

논문접수일 : 2014년 08월 13일 논문수정일 : 2014년 09월 30일 논문게재확정일 : 2014년 10월 01일

※ 본 연구는 미래창조과학부 및 정보통신산업진흥원의 대학 IT연구센터 지원사업의 연구결과로 수행되었음(NIPA-2013-H0301-13-4009). 이 논문은 2013년도 충북대학교 학술연구지원사업의 연구비 지원에 의하여 연구되었음.

* 충북대학교 비즈니스데이터융합학과, e-mail : lyudoong@naver.com

** 교신저자, 충북대학교 경영정보학과, BK21+BSO사업팀, e-mail : chois@cbnu.ac.kr

1. 서 론

인터넷과 스마트기기의 발달로 인한 온라인의 세상은 그림・음악・동영상・텍스트 등 다양한 형태의 수많은 콘텐츠들을 담고 있는 이른바 빅 데이터의 세상이다. 기업들은 지속적으로 인터넷 사용 인구가 증가함에 따라 웹사이트의 사용성 향상을 위해 다각적인 노력을 해오고 있다. 웹사이트 관리자 입장에서 웹사이트의 디자인, 웹 서버의 디자인, 웹사이트 네비게이션 설계 등의 업무는 매우 복잡한 작업분석이 필요하기 때문에 하나의 웹사이트가 어떻게 사용되는지를 분석하는 작업은 중요하다[Srivastava et al., 2000]. 사이트 개발자 또는 관리자는 웹사이트를 사용자 입장에서 설계하기 위해서는 그들의 다양한 특성들을 이해하고 이를 반영하여야 한다. 하지만 웹사이트 분석가의 입장에서는 분석이 곤란할 정도의 많은 양의 데이터들이 쏟아지고 있다. 많은 양의 데이터 중에서 의미 있는 가치를 찾아내려는 웹 데이터 마이닝 기술들이 많이 연구 중이다. 웹 데이터 중에는 사용자들이 직접적으로 데이터를 입력하는 사항도 많으나 갱신이 되지 않거나, 사실대로 제대로 기입되지 않는 경우도 많아서 데이터들을 마이닝을 활용하여 분석하여 사용자들의 특성을 파악하기 보다는 실시간으로 사용자들의 행동이 기록이 되는 웹 서버의 로그를 분석하는 것이 보다 객관적이고 정확한 결과를 도출해 낼 수 있는 분석이라고 할 수 있다.

본 연구에서는 웹사이트의 사용성 향상을 위해 웹 로그 파일을 분석하고자 한다. 로그 파일 분석을 통해 사용자의 페이지 뷰 분석, 방문 경로 파악 등을 통해 사이트 메뉴의 순서를 개편하거나 통합이 필요한 메뉴들을 찾아내고 웹사이트 내 필요 기능들을 도출하고자 한다. 본 연구에서는 우선 사용자의 특성을 이해하기 위해 그들의 웹사이트에서의 활동 흔적인 웹 로그를 프로세스 마이닝 기법으로 분석함으로써 그들의 행동 패턴들을 살펴보고 사용자 그룹 마다 어떠한 특성을 가지고 있는지 알아보고자 한다. 기존의 연구들에서는 사용자의 행동 패턴을 분석하는 방법으로 흔히 휴리스틱 알고리즘이나 퍼지 알고리즘을 많이 사용하였는데, 본 연구에서는 프로세스 마이닝 기법 중에 하나인 Performance Sequence Diagram 분석 기법을 사용하여 사용자들을 좀 더 세밀하게 분류 해보고자 한다.

2. 웹사이트 로그 및 패턴분석 관련연구

2.1 웹 로그 분석

웹 로그 분석은 웹 서버가 웹페이지에 대해 서비스를 제공하면서 실시간으로 생성되는 로그파일을 원천 데이터로 하여 분석을 하는 것으로, 웹 서버의 트래픽이나 오류 또는 방문경로 등의 대한 분석을 수행하여 웹 서버의 시스템적인 문제나 웹페이지 구성에 대한 문제 등을 분석하는 것을 말한다. 웹 로그 분석은 웹사이트의 관리를 하는 의사결정자에게 의사결정에 필요한 정보를 제공한다[김동곤, 2013]. 웹 로그 분석을 위해서는 서버 엑세스 로그에 대한 데이터 마이닝 기법을 적용하는 웹 사용 마이닝을 수행할 필요가 있는데, 이를 위해서는 전처리 과정이 중요하다. 웹사이트 분석가는 전처리된 정보를 이용하여 각 페이지의 접근 빈도, 자주 접속되는 페이지, 페이지 네비게이션 패턴 등의 유용한 정보를 활용할 수 있다. 이를 위해 다양한 전처리 방법에 관한 연구가 진행되어 왔다[Cooley et al., 1999]. 웹사이트 로그 데이터를 이용하여 사용 패턴을 도출하기 위한 연구가 다각적으로 이루어질 뿐 아니라 웹사이트 방문자의 사용성을 평가하거나 온라인 소비자의 성향을 파악하는 등의 분석업무에 적극 활용되고 있다

[Srivastava et al., 2000; 고석하 외 2인, 2005; 전종근, 박철, 2006]. 웹 로그란 웹 서버를 통해서 행하여지는 모든 작업들에 대한 기록으로 웹 서버는 웹 서비스에 대한 요청과 결과에 대한 것들을 모두 저장해 두기 때문에 웹 로그 파일을 열어보면 누가 언제 무엇을 했는지 알 수 있다. 웹 로그 파일은 웹 서버의 종류와 지원하는 웹 로그파일의 형식에 따라 많은 차이가 있지만 현재 가장 많이 보편적으로 사용하는 웹 로그 형식은 아파치 웹 로그 파일이나 IIS, W3C Extended 등이 있다[문경선, 2002].

웹 로그는 웹 서버에 따라서 한 개가 아닌 여러 개가 생성이 될 수도 있는데 그 종류는 크게 엑세스 로그파일, 에러 로그파일, 레퍼럴 로그파일 및 에이전트 로그파일 등 4가지로 분류할 수 있다[전재환, 2002]. 이중에서 본 연구에서 사용한 로그파일은 CLF로 보편적으로 많이 사용하는 파일 형식이며 로그파일의 종류는 엑세스 로그파일로 일반적으로 방문자가 웹사이트 내에서 메뉴를 클릭하여 웹 서버로의 요청 등에 대한 기록으로 구성이 되어 있다.

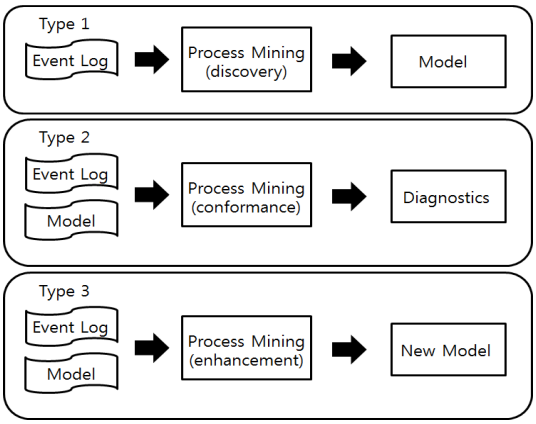
본 연구의 웹 로그 분석은 웹 데이터마이닝 분야 중에서 사용자들이 웹사이트를 브라우징 했던 기록을 남기는 엑세스 로그 파일로부터 접속 패턴 유형 발견을 목적으로 하는 웹 사용 마이닝 연구로 분류할 수 있다[김민정, 2002]. 이밖에 레퍼럴 로그나 CGI 스크립트 등에 의해서 획득한 고객 등록 정보나 설문데이터 등으로 개인화 맞춤서비스, 시스템 향상, 사이트 개선 등에 활용할 수 있다[김광용, 2002]. 웹 로그 마이닝에 대한 연구들을 몇 가지 살펴보면 학교 이러닝 효율화 향상[남궁영, 2005; 김정현, 2013], 공공기관의 웹사이트 운영에 대한 전략 수립[문현주, 2013], 온라인 마케팅 시스템설계[김재훈, 2011] 등 공공기관, 민간기업 구분 할 것 없이 웹 서버를 가지고 웹 서비스를 제공한다면 다양한 분야에서 활용이 가능하지만, 해당 연구들의 분석 알고리즘은 연관성규칙이나 퍼지 알고리즘, 휴리스틱 알고리즘처럼 다소 한정된 분석 기법들이 대부분 사용되고 있다.

2.2 패턴 분석을 위한 프로세스 마이닝 도구

비즈니스 프로세스 관리(Business Process Management : BPM)는 IT분야에서 비즈니스적 관점에서 사업적인 활용을 위한 하나의 기술적인 도구라는 개념에서 시작하여, 현재 기업의 경쟁력을 강화하기 위한 전략적인 방향으로 진화하고 있다[Garter, 2003]. BPM은 기업들의 생산성 향상을 위해 업무의 프로세스를 체계적으로 설계・관리・개선을 하는 데에 있어서 모든 활동을 지원하는 총체적인 관리 방법론이다[Smith, 2003]. BPM을 추진할 경우 기업들이 기대할 수 있는 가장 큰 효과는 업무의 수행이나 관리의 수준이 향상이 된다는 점으로써 업무들이 정형화가 됨에 따라 조직의 역량의 향상이 이루어지게 된다는 것이다. 업무와 관련된 명확한 규칙이나 절차를 표준적으로 제공하고 업무에 대한 상황을 실시간으로 파악을 하며 균일한 업무 수행능력의 조기 확보와 병목업무의 통계적인 파악 및 지속적인 개선가능, 업무누수시간 최소화 등을 가능하게 하므로 결과적으로 기업의 조직적 역량이 향상되는 것으로 연구되었다[정민수, 2010].

프로세스 마이닝은 ERP, CRM, SCM과 같은 정보시스템을 사용하는 과정에서 시스템에 저장되는 프로세스 로그를 분석하여 비즈니스 프로세스를 관리하기 위한 기법으로 분석목적에 따라 <그림 1>과 같이 세 가지 타입으로 분류될 수 있다. 첫 번째 타입인 discovery는 이벤트 로그를 사용하여 프로세스 모델을 만들어 주는 기술이다. 대표적인 discovery 기술에는 α-알고리즘, 퍼지마이닝, 휴리스틱 마이닝 등이

존재한다. 두 번째 타입인 conformance는 동일한 프로세스로부터 만들어진 프로세스 모델과 이벤트 로그를 비교하여 기존의 프로세스 모델이 기록된 이벤트 로그에 적합한 프로세스 모델인지 확인해 준다. 마지막 타입인 enhancement에서는 실제 프로세스 정보를 사용하여 기존 프로세스 모델을 확장하거나 향상시키는 작업을 수행한다. Enhancement는 repair과 extension 두 가지 타입으로 나누어지며, repair는 현실성을 반영하여 프로세스 모델을 수정하고, extension은 프로세스 모델과 로그를 상호 비교하여 시간 기록들(time stamps), 자원들(resources), 장애물(bottlenecks) 등과 같은 새로운 정보를 프로세스 모델에 추가해준다[van der Aalst, 2011].



〈그림 1〉 프로세스 마이닝 분석 타입

프로세스 마이닝에 대한 관련 연구로는 프로세스 마이닝에서 가장 기본적인 알고리즘인 α-알고리즘에 대한 소개[Aalst, 2004], 프로세스 마이닝 분야에서 많이 사용하는 퍼지 마이닝 [Gunther, 2007], 기존의 워크플로우 기반의 소셜 네트워크 기법을 활용한 연구[김광훈, 2012]들이 있으며, 활용 분야로는 공공기관, 금융, 의료, 제조 등 전 산업에 걸쳐 적용이 되지만 가장 활발히 활용이 되는 분야는 제조업으로 공정에 대한 주요 프로세스를 도출하고 프로세스분석과 성과분석 등을 통해 기업에 유용한 정보를 제공[송민석, 2012]하는데 많이 활용 되고 있다.

3. 연구 설계

3.1 홈페이지 접속 로그 분석 개요

본 연구의 분석대상은 C대학교에서 운영하는 학교 홈페이지를 방문하는 방문자들에 대한 브라우징 패턴이다. 분석에 사용하는 데이터는 홈페이지에 대한 웹 서버의 액세스 로그 기록으로써 방문자들에 대한 브라우징 패턴을 알아보고, 그 결과로 방문자들이 가지고 있는 특성을 이해하고 홈페이지에 대한 활용도가 향상이 되도록 할 수 있는 정보를 제공하고자 한다.

분석에 사용하는 대상 데이터는 C대학교 홈페이지에 대한 사용자들의 웹 로그기록이며 데이터에 대한 특성은 다음과 같다. C대학교 홈페이지 로그데이터는 홈페이지 웹 서버를 통해 접속했던 사용자들의 액세스 로그 파일이며 텍스트 형식으로 되어 있다. 로그데이터 수집기간은 2013년 7월 30일부터 동년 8월 6일까지 8일이다. 데이터의 총 레코드 수는 약 140만 건 정도이고 용량은 138MB 정도이다. 데이터의 탐색과 전처리를 위해 엑셀을 사용하였으며 데이터분석 도구의 특성상 날짜별로 나누어 전처리 작업을 수행하였다.

〈표 1〉 C 대학교 홈페이지 로그데이터 설명

속성	설명
Content	Web-server log record of University Homepage
Format	Access log file(text)
Volume	137,877 kb
Main Access Page	Mainpage, Introduction, College, Community, Notice, Academic administration, Campus life
Record Amount	1,445,463 record (2013. 07. 30~2013. 08. 06 , 8days)

3.2 홈페이지 로그데이터 탐색 및 전처리

로그데이터를 살펴보면 <표 2>와 같이 총 7개의 필드로 구성되어 있다. 본 연구의 목적은 로그에 대한 기초통계 및 사용자 유형별 브라우징 패턴을 분석하는 것이다. 우선, 연구 목적에 적합한 필드들을 규명하고, 불필요한 레코드들을 삭제하는 과정 등의 전처리를 어떻게 진행해야 하는지를 결정하였다.

〈표 2〉 로그데이터 필드 및 속성값 예시

필드	값(사례)
Host IP	203.225.78.23
Identification	203.225.78.23-01
User Authentification	-
Time Stamp	[30/Jul/2013:00:00:01 +0900]
HTTP Request Field	“GET /kor/index.jsp HTTP/1.1”
Status Code	200
Transfer Volume	60475

Host IP는 웹사이트 방문자에 대한 도메인이나 IP주소가 기록이 되는 곳으로 웹 서버 관리자가 둘 중 어느 것으로 기록하게 할지 결정을 하지만 보통은 IP주소를 남기도록 한다. 그 이유는 웹 서버가 항상 IP주소를 사용하기 때문에 도메인 주소로 기록을 하게 되면 모든 접속마다 서버는 그 도메인을 역추적 해야 하기 때문에 웹 서버에 상당한 부하를 주게 되므로 보편적으로는 IP를 기록한다. Identification는 접속한 사용자에 대하여 이름을 표시하는 곳으로 거의 사용하지 않기 때문에 “-”로 보이는 경우가 많은데 이 사례에서는 접속자의 IP와 접속번호를 연결하여 사용하기로 하였다. User Authentification는 등록된 사용자의 이름이 표시되는 곳으로 웹 서버의 특정 디렉토리를 지정하여 사용자 이름과 암호를 설정해 놓은 경우에 해당 디렉토리는 등록된 사용자에게만 자료검색을 허락하며 보통 때는 “-‘로 표시 된다. Time Stamp는 사용자가 웹 서버로 요청을 보낸 시간이 표시가 되는 필드로 날짜와 초단위까지 나타나며 시간뒤에 +, - 기호와 함께 그리니치 표준시와의 차이가 기록된다.

HTTP Request Field는 사용자가 서버에 보내는 요청이 기록이 되는 필드로 3부분으로 구성이 되어있다. 첫 번째 부분은 요청방식에 따라서 GET, POST, HEAD 등으로 나타나며 두 번째는 요청에 대한 실제 대상 파일의 이름이고 마지막은 전송 프로토콜 부분인데 현재는 대부분 HTTP를 사용하기 때문에 “HTTP.버전번호”로 기록이 된다. Status Code는 웹 서버는 접속자의 요청에 대해 세 자리 수로 된 오류코드와 함께 응답을 하는데, 이 숫자들은 접속 상태와 데이터에 대한 상태들을 표기한 것으로 제일 앞자리의 숫자에 따라 그 의미가 구분이 되어 있으며, 나머지 자리의 수들로 인해 세부적인 의미로 구분이 된다. 오류가 없이 정상적으로 전송이 되었을 시에는 200으로 표시되고, 400번 대는 클라이언트에 대한 오류이고, 500번대는 서버에 대한 오류를 나타낸다. Transfer Volume은 사용자가 웹 서버로부터 실제 가져간 데이터의 용량을 기록한 것으로 단위는 Byte단위로 기록이 되며, 상태코드가 200인 경우에만 기록이 되며 나머지에 경우에는 “-”으로 기록이 된다. 요청이 HEAD인 경우에는 요청이 성공을 했더라도 데이터의 양은 계산을 하지 않으므로 “-”으로 기록된다[문경선, 2002].

3.3 데이터 전처리

데이터의 전처리 단계는 분석의 질을 높이는 첫 단계라 할 수 있다. 불필요한 레코드들을 삭

제하고 새로운 변수를 생성하기도 하며, 결측치 등의 처리 여부 등 데이터의 잡음을 없애고 불완전성을 잡아가는 과정이다. 프로세스마이닝 도구를 이용하여 브라우징 패턴을 분석하기 위해서 로그데이터를 프로세스마이닝 필드로 변환하는 과정이 필요하다. 프로세스마이닝 분석에 필요한 변환을 위해서 원본데이터를 엑셀을 이용하여 불러들인 후에 각 필드를 접속자 IP는 IP로 처리시간을 Time Stamp로 HTTP Request Field를 Url로 상태코드를 Status로 Transfer Volume을 Volume으로 매칭시킨 후에 전처리를 진행하였다. 그리고 사용자인식 정보 필드와 패스워드 보호영역 필드를 보면 아무런 정보가 없이 모든 레코드가 "-"로 표기가 되어 있었기 때문에 삭제를 시켰고, Time Stamp 레코드들의 뒷부분은 공백 구분자로 +0900는 따로 필드가 나누어지는데 이 역시 삭제를 하였다.

공공장소에서 접속하는 경우가 많아서 IP를 사용하여 특정 사용자를 구분하기가 쉽지가 않다. 따라서 최소한의 구분을 위해 IP와 Time의 날짜를 매핑을 시켜서 같은 IP라도 날짜가 다르면 다른 사용자로 구분을 할 수 있도록 ID라는 새로운 변수를 생성하였고 패턴분석에서 고유 아이디로 IP대신 ID변수를 사용하였다. Url 필드를 살펴보면 요청에 대한 내용이 확장자 .js나 .swf와 같이 홈페이지의 디자인적인 요소들에 대한 것으로 분석에 필요없는 기록들이 존재하기 때문에 삭제하였다. HTTP Request Field에서 요청 메소드가 GET과 POST가 아닌 HEAD 요청의 경우 일반 사용자가 아닌 서버관리자들에 의한 기록이므로 삭제하였다. 원본데이터의 필드를 분석을 위해 새로운 변수로 매핑한 결과는 다음과 같다.

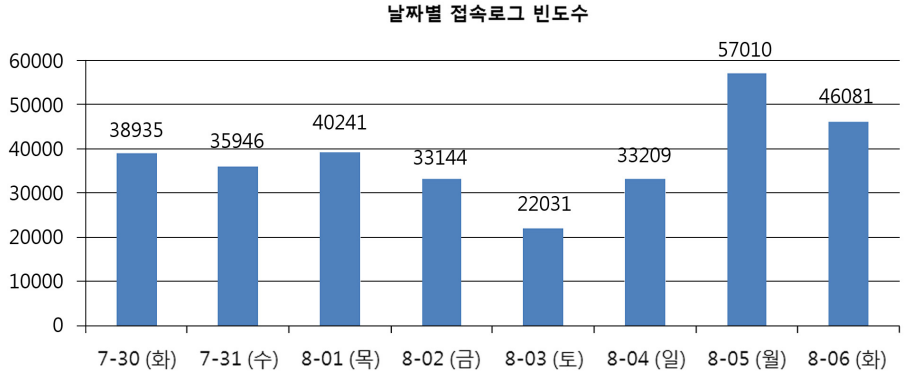
4. 분석결과

4.1 기초 통계 분석

IP 출처와 URL 접속단위를 명확하게 구분할 수 있는 로그들로 정리한 결과 총 306,597건이었다. IP 출처 분석을 통해 교외에서 접속한 로그는 267,796건, 교내는 38,801건 인 것으로 확인되었다. 사용자들이 접속한 페이지는 학교 홈페이지 URL로 구분하였으며, 메인메뉴와 빈도수가 높은 상위 메뉴의 세부메뉴를 정리하여 70개의 메뉴에 접속하였음을 확인하였다. 그림 2에서 보는 바와 같이, 데이터 수집기간 동안 날짜별 (요일별)로 접속빈도를 분석해 보았다. 8월 3일 토요일에 접속자가 가장 적고, 8월 5일 월요일에 대한 접속자가 가장 많으며, 월요일을 기점으로 주말로 갈수록 그 빈도는 점점 내려가다가 목요일에 다시 증가하는 추세로 나타났다.

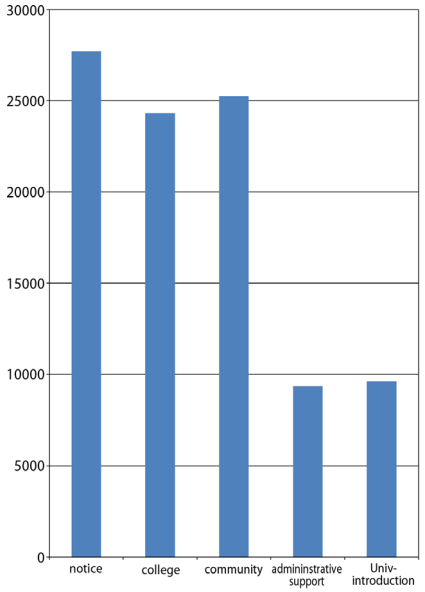
〈표 3〉 로그데이터와 프로세스마이닝 도구 간 필드명 매칭

원본데이터 필드 ➡	매칭결과(엑셀)	비고
Host IP	IP	사용자 분류
Identification	삭제	
User Authentification	삭제	
Time Stamp	Time Stamp	접속시간
HTTP Request Field	URL	Access page
Status Code	Status	HTTP 상태코드
Transfer Volume	Volume	해당 요청에 대한 용량
	IP+Time 활용하여 ID필드 생성	분석에서 Case ID로 사용



〈그림 2〉 날짜별 접속로그 빈도수

데이터 수집기간동안 홈페이지 방문자들이 어떠한 메뉴를 많이 방문 했는지를 분석하기 위해서 <그림 3>에서 보는 바와 같이 메인메뉴 및 하위메뉴에 대한 접속 빈도를 분석하였다. 홈페이지는 메뉴별로 하위메뉴들이 있고, 그 하위메뉴들 아래에는 또 다른 하위 메뉴들이 존재한다. 홈페이지를 처음 방문했을 때의 메뉴들을 메인메뉴라 정의하며 세부메뉴 레벨 1은 메인메뉴의 하위메뉴이며 세부메뉴 레벨 2는 해당 하위메뉴 아랫 단계의 하위메뉴를 나타내는 것으로 정의하였다. 그러므로 하위메뉴의 레벨이 높아질수록 해당 페이지를 접속하기 위해서는 더 많은 페이지를 거쳐야한다. <그림 3>의 오른쪽 부분은 방문자들이 자주 방문하는 하위메뉴를 상위빈도부터 내림차순으로 나타낸 것으로 이를 살펴보면 단대별/학과별 홈페이지로 접속할 수 있는 대학/대학원 메뉴와 공지사항, 커뮤니티 카테고리를 많이 이용함을 볼 수 있다. 가장 많은 빈도수는 첫 메인페이지였으나 분석의 편의상 해당 페이지 내용은 제거하였다.



메인 메뉴	하위메뉴-Lv1	하위메뉴-Lv2	빈도
대학/대학원	대학		13136
개신골관장	공지사항	공지사항	9760
학사행정	학사일정		8488
개신골관장	커뮤니티	학교생활	6932
개신골관장	커뮤니티	구인구직	6519
개신골관장	공지사항	학사/장학	5472
개신골관장	공지사항	공지사항	5197
개신골관장	공지사항	CBNU뉴스	5102
개신골관장	공지사항	공지사항	4753
메인페이지			2961
개신골관장	커뮤니티	구인구직	2929
학교안내	학교소개	대학헌장	2851
대학/대학원	대학원		2543
개신골관장	공지사항	학사/장학	2530
학교안내	학교조직	조직도	2490
대학/대학원	대학	공과대학	2351
개신골관장	커뮤니티	삽니다/팝니다	2351
대학생활정보	상담서비스	학사상담	2326
대학/대학원	대학	자연과학대학	2306
개신골관장	커뮤니티	학사/장학	2297
개신골관장	커뮤니티	학교생활	2230
개신골관장	커뮤니티	수업/스터디/리포트	2203
개신골관장	커뮤니티	삽니다/팝니다	2100
대학/대학원	대학	사회과학대학	2083
공지사항	커뮤니티	거주정보(하숙)	2005
대학/대학원	대학	농업생명환경대학	1898
학교안내	캠퍼스투어	캠퍼스맵(확대)	1808
학사행정	교육과정	교육과정이수관련규정	1752

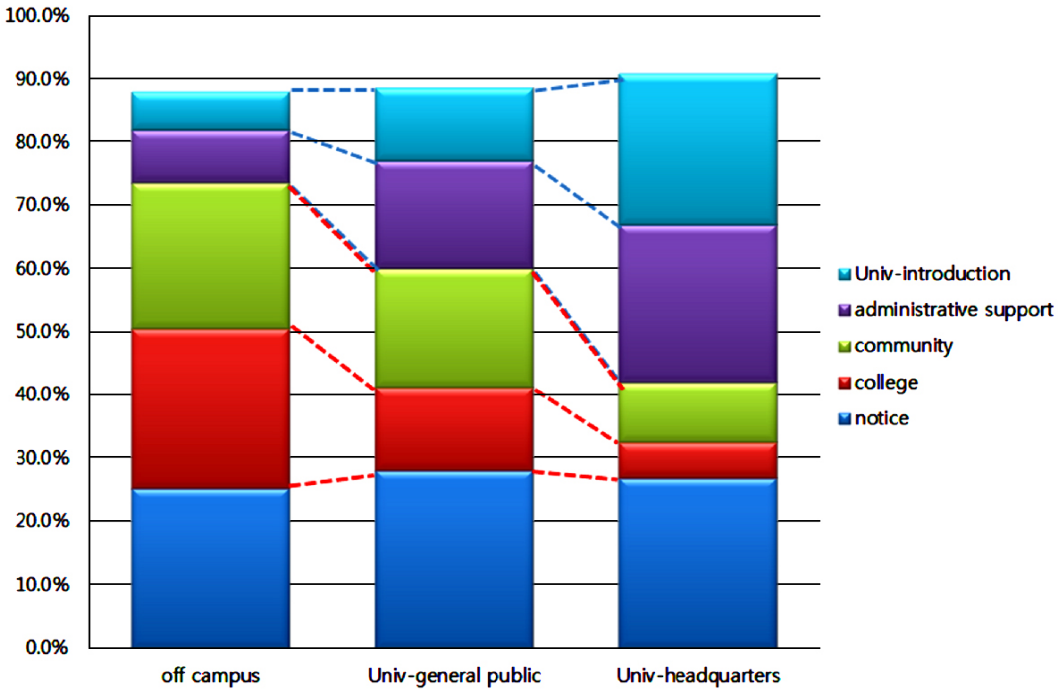
〈그림 3〉 URL 빈도표 - 전체사용자

기초통계 분석결과 시사점

기초통계 분석 결과를 몇 가지로 정리해보면 다음과 같다. 첫째, 교외접속자는 다른 사용자들에 비해서 취업관련 정보 열람의 비율이 높았다. 학교 홈페이지에서 제일 많이 이용하는 카테고리는 공지사항과 커뮤니티이며, 교외접속자와 교내접속자의 차이는 교외접속자의 경우에 공지사항 메뉴의 취업공지페이지와 커뮤니티 메뉴 중 구인/구직 페이지처럼 취업관련 정보페이지들의 열람빈도가 높게 나타났다. 둘째, 교내사용자들의 경우 커뮤니티 카테고리의 방문비율이 외부사용자에 비해 절반 정도이다. 교내본부 접속자들이 학교소개 메뉴와 단대별대학을 많이 방문하는 이유는 업무상 지원페이지 방문과 조직도나 각 단과대 페이지에 방문함으로써 업무의 용도로 연락처를 확인하는 것으로 예상해 볼 수 있다. 셋째, 홈페이지의 많은 페이지들 중에서 방문자들이 이용하는 페이지는 특정 카테고리들에 집중되어 있다. 학교 홈페이지 카테고리의 방문비율을 살펴보면 홈페이지 5개의 메뉴(공지사항, 단대별대학, 커뮤니티, 학교소개, 지원페이지)에 약 90%로 집중이 되어있음을 알 수 있었다.

4.2 IP 출처별 접속빈도 분석

홈페이지에 접속한 ip들을 기초분석과 교내 ip할당 현황표를 가지고 사용자들을 최소한으로 구분해보기 위해서 교외, 교내일반, 교내본부 접속자로 나누었다. 교내본부의 경우는 학교에서 행정업무를 담당하는 직원들이 주된 사용자이므로 다른 교내 사용자들과 브라우징 패턴이 다를 것이다라는 가정을 가지고 분류를 따로 하였고, 그 외에 교내 접속자들을 교내 일반으로 분류 하였다. 각 출처별로 홈페이지에 대한 메인메뉴 단위의 접속 비율을 상위 5개 메뉴로 정리를 하면 <그림 4>와 같다.

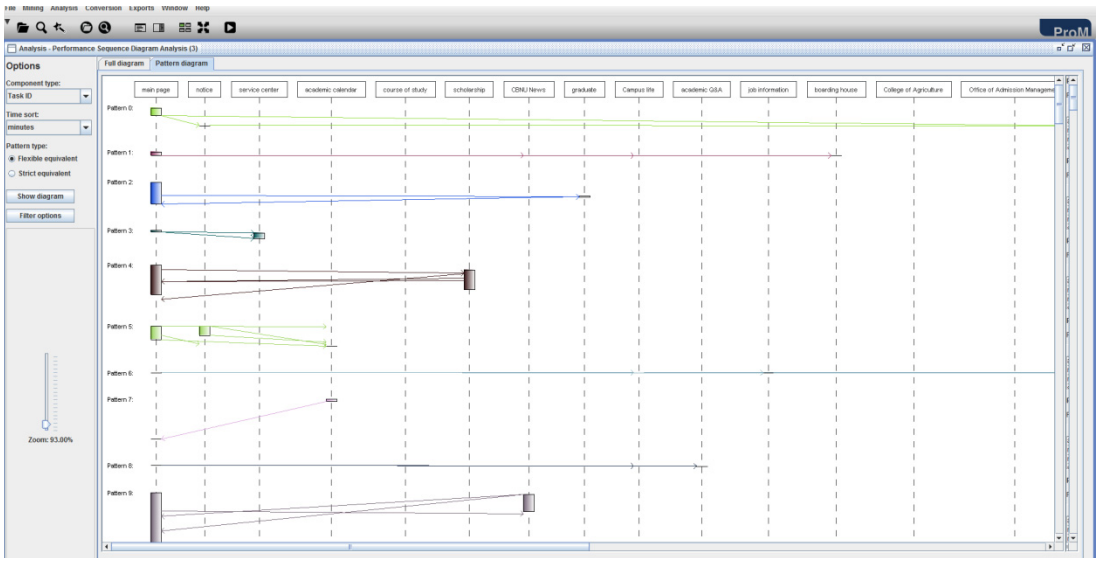


〈그림 4〉 IP 출처별 메인메뉴 단위 접속 비율 상위 5개

메뉴의 개수가 너무 많아 모든 페이지를 변수로 고려할 수가 없으므로 메인메뉴 단위를 기준으로 상위 5개의 메인메뉴 만을 표시하였다. <그림 4>는 IP출처별로 홈페이지 메뉴에 대한 접속 비율을 나타낸 것인데 3가지 사용자유형들 모두가 전체 카테고리의 상위 5개 메뉴의 접속비율이 전체의 약 90% 정도를 차지하는 것으로 나타났다. 방문자들이 홈페이지에서 활용하는 정보는 위 5개 메뉴에 집중되어 있음을 알 수 있었으며, 전체적으로 공지사항(notice) 메뉴는 방문비율이 비슷하였지만 대학/대학원(college) 메뉴와 커뮤니티(community) 메뉴의 방문비율은 교내일반 사용자와 교내본부 사용자로 옮겨갈수록 점점 내려가고, 행정지원(administrative support) 페이지와 학교안내(Univ-Introduction) 메뉴들은 점점 비율이 높아짐을 볼 수 있다. 데이터의 수집기간이 방학중이였기 때문에 해당 데이터 수집기간에 교내에서 접속한 사용자들은 주로 행정업무를 담당하는 직원이나 조교임을 예상할 수 있으며, 이에 비해 일반 학생들은 커뮤니티와 단대별 대학 메뉴에 접속하는 빈도가 높았다고 할 수 있다.

4.3 홈페이지 접속자 브라우징 패턴분석

사용자들의 홈페이지 브라우징 패턴 분석을 위해서 Performance Sequence Diagram 패턴분석을 수행하였다. Performance Sequence Diagram 패턴분석은 고유한 ID에 대해서 시간 순으로 어떤 활동을 했는지 분석을 하는 기법이다. 패턴 분석은 Url 속성의 수 70개를 가지고 IP출처별(교외・교내일반・교내본부)로 구분한 3가지 데이터로 나누어 분석을 진행 하였다. <그림 5>는 교외접속자들의 Performance Sequence Diagram 분석에 대한 실행화면의 결과이다. 기존의 휴리스틱이나 퍼지마이닝으로 해당 분석을 실시하면 변수가 많아질수록 분석에 어려움이 많은데 Performance Sequence Diagram 분석은 변수가 많이 있더라도 분석결과에 대한 가독성이 용이하며, 많은 변수를 사용하지 않는 패턴도 표현해 주기 때문에 패턴분석에 대한 결과의 질이 더 높다고 할 수 있다.



〈그림 5〉 Performance Sequnce Diagram 분석 화면

〈표 4〉 유형별 분류기준표

유형	주요패턴	패턴빈도수
공지사항 확인형	공지사항 ↻	1181
	메인페이지 → 공지사항	5390
	메인페이지 → 공지사항 → 취업공지	228
학사자료 열람형	학사일정 ↻	1038
	메인페이지 → 학사일정	1815
	메인페이지 → 학사/장학	1389
커뮤니티 확인형	구인/구직 ↻	851
	메인페이지 → 커뮤니티 → 구인/구직	1394
	메인페이지 → 학교생활 → 거주정보	274
단대별대학 방문형	메인페이지 → 대학 → 인문대학	531
	메인페이지 → 대학 → 공과대학	406
	메인페이지 → 대학 → 공과대학 → 학교안내 → 학교소개	384
지원페이지 방문형	메인페이지 → 학생과	398
	메인페이지 → 공지사항 → 학사과	266
	메인페이지 → 총무과	79
연락처 확인형	메인페이지 → 조직도	244
	메인페이지 → 조직도 → 단대별대학	238
	메인페이지 → 학교조직 → 전화번호안내	165

Performance Sequence Diagram 분석에서 나온 결과를 바탕으로 패턴을 정리해보고 브라우징 패턴이 비슷한 사용자들끼리 패턴을 분류하였다. <표 4>에서 보는 바와 같이, 패턴분석 결과 18가지의 주요패턴으로 정리되었으며, 3가지의 주요패턴을 하나의 유형으로 분류하여 총 6개 유형의 브라우징 패턴으로 분석되었다. 예를 들어, 공지사항(재방문 포함), 메인페이지 → 공지사항, 메인페이지 → 공지사항 → 취업공지 등의 세 개의 주요패턴들은 공지사항 확인형의 브라우징 유형으로 구분하였다. 이와 같은 방식으로 6가지의 유형, 공지사항 확인형, 학사자료 열람형, 커뮤니티 확인형, 단대별대학 방문형, 지원페이지 방문형, 연락처 확인형 등으로 구분할 수 있었다.

18가지 주요패턴에 대응되는 로그들을 상위 빈도순으로 정렬을 하여 정리하였으며, 공지사항확인형이 가장 높은 빈도를 가지며, 나머지는 <표 4>의 유형 순으로 빈도가 높은 것을 알 수 있었다. 주요패턴 중에서 특정 하나의 페이지만 클릭해 들어가는 경우가 많은데 메인페이지를 거치지 않고 들어가는 경우는 학교의 모바일용 홈페이지로 접속하는 경우로 분석되었으며, 패턴의 빈도를 살펴보았을 때 기초분석의 결과와 일치함을 알 수 있다.

사용자그룹별로 브라우징 패턴 유형들에 대한 빈도수 분석 결과(<표 5> 참조), 교외, 교내일반, 교내본부 사용자 그룹 모두 공지사항 확인형이 가장 많았으나, 교외의 경우 취업을 위해, 교내의 경우 일반 공지 확인을 위해 접속하는 것으로 나타났다.

교외 사용자그룹은 취업을 위한 공지사항 확인형이 가장 많았으며, 다음으로 학사자료 열람, 구인구직을 위한 커뮤니티 확인이의 순으로 많은 것으로 보아서 교외 사용자 그룹은 학생들

〈표 5〉 사용자그룹별 패턴유형 빈도

No.	교외		교내일반		교내본부	
	유형	비율	유형	비율	유형	비율
1	공지사항확인(취업)	35%	공지사항확인(일반)	50%	공지사항확인(일반)	56%
2	학사자료열람	25%	학사자료 열람형	24%	지원페이지 방문형(연락처)	32%
3	커뮤니티확인(구인/구직)	18%	커뮤니티 확인형	10%	학사자료 열람형	7%
4	단대별대학방문	10%	지원페이지 방문형	10%	연락처 확인형	5%
5	기타	12%	기타	6%	-	-

이 학교 외부에서 접속하는 경향이 많은 것을 알 수 있었다. 교내일반 사용자그룹도 교외 사용자그룹과 유사한 패턴을 보인 것으로 분석되었는데, 이는 교내일반 사용자 그룹은 주로 학생들에 의해 접속되었다는 것을 알 수 있다. 교내본부 사용자 그룹은 공지사항 확인 이외에 연락처 열람을 위한 지원페이지 방문과 연락처 확인이 많은 것으로 분석되었는데, 이는 교내본부 사용자 그룹은 주로 교내 직원들에 의해 접속되었다는 것을 알 수 있다.

브라우징 패턴분석 결과 시사점

패턴분석의 관점에서 분석결과를 몇가지로 요약해보면 다음과 같다. 첫째, 예비대학생들의 대학에 대한 정보 열람이 많다. 메인페이지 → 학교소개 → 단대별대학을 방문하는 패턴으로 보아 교외접속자들의 접속 패턴 중 학교소개 메뉴와 단대별 대학페이지를 거쳐가는 패턴이 많은 이유는 데이터 수집기간이 방학 중이었고, 수시 모집을 앞둔 시점이었다는 점에서 예비대학생들이 학교에 대한 정보를 열람하는 것으로 예상 할 수 있다.

둘째, 교내 접속자들은 학교홈페이지를 업무용으로 사용하는 비율이 높은 것으로 나타났다. 메인페이지 → 조직도, 메인페이지 → 단대별대학 → 지원페이지를 방문하는 패턴들을 보면 본부접속자들이 학교소개 메뉴와 단대별 대학을 접속하여 업무상 조직도나 각 단과대 행정실에 대학 연락처를 확인하는 것으로 예상해 볼 수 있다.

셋째, 교외접속자들의 경우에 모바일용 학교 홈페이지를 많이 이용한다. 공지사항 페이지만 열람하거나, 공지사항 → 구인/구직을 거치는 교외접속자들의 패턴을 살펴보면 메인홈페이지를 거치지 않고 바로 특정메뉴로 들어가는 경우들이 많이 보이는데, 이는 학교 홈페이지를 스마트폰으로 접속을 하는 경우가 많을 것이라 볼 수 있다.

넷째, 학교주변 건물주들의 홍보용 방문이 있다. 메인 → 거주정보(하숙) → 글쓰기로 가는 패턴은 커뮤니티 카테고리 중에 거주정보(하숙)이라는 메뉴로 학교주변에 방을 구하기 위한 정보를 열람하는 곳인데 패턴에 대한 빈도는 274로 그 중 62개의 패턴은 글쓰기 메뉴를 거치는 패턴을 보였다. 실제 해당메뉴는 방을 구하는 입장에서는 거의 글을 남기지 않고 방을 내놓는 사람들이 주로 홍보용으로 글을 남기는데 사용되는 것으로 분석되었다.

4.4 홈페이지 활용도 향상을 위한 방안 고찰

종합분석 결과, 홈페이지 활용도 향상을 위해서 다음의 몇 가지 대안이 고려될 수 있다. 첫째, 공지사항과 커뮤니티 카테고리에 대한 메뉴의 순서가 바뀔 필요가 있다. 공지사항과 커뮤니티의 카테고리는 사람들이 방문하는 빈도가 제일 높은 페이지 들이다. 각 메뉴를 클릭했을 때 보여지는 페이지는 각각 대학뉴스와 학교생활 메

뉴인데 이 메뉴들은 패턴분석에서 보았을 때 대다수가 다른 세부 메뉴들로 이동하는 패턴을 보였다. 이는 방문자들이 해당 페이지들보다는 카테고리에 다른 세부 페이지에 대한 정보에 대해 관심이 더 높음을 알 수 있다. 그러므로 불필요한 트래픽의 발생을 줄일 수 있도록 세부메뉴들 중 방문 빈도가 가장 높은 페이지들을 클릭할 때 첫 페이지로 나타나도록 변경할 필요가 있다.

둘째, 학사정보에 대한 메뉴 통합이 필요하다. 학사정보들의 경우, 최상위 대분류 메뉴인 학사행정 카테고리 외에 공지사항과 커뮤니티에도 메뉴 성격에 따라 각각 메뉴가 하나씩 할당이 되어있다. 학사정보에 관심이 많은 방문자들의 비율이 교외와 교내 일반의 경우 25% 정도로 많은 비중을 보이는데 학사자료에 대한 정보를 열람을 위해 상위메뉴 세 개의 페이지를 번갈아가며 들어가야 하며, 이 또한 불필요한 브라우징 동선의 낭비를 유발하므로 타 카테고리에 있는 학사정보 관련메뉴들을 학사행정메뉴로 통합을 하는 방안을 제안한다.

셋째, 교직원 메뉴에 연락처 검색 추가를 제안한다. 학교 홈페이지를 업무의 용도로 사용하는 방문자들의 경우 각 행정부서의 홈페이지나 조직도, 단대별대학 페이지를 많이 거치는 패턴들을 보인다. 이는 연락처의 확인을 하는 용도로 각 페이지들을 방문하는 것으로 생각해 볼 수 있다. 따라서 메인페이지의 교직원 메뉴에 각 부서별 담당자들의 연락처를 검색할 수 있는 메뉴를 생성한다면 연락처 검색에 대한 동선이 간소화 될 것으로 예상한다.

5. 향후 연구 방향

본 연구에서는 홈페이지 활용도 향상을 위해 로그 파일에 대한 기초 통계 분석 및 프로세스 마이닝 분석을 통해 메뉴의 개선방안 및 기능의 향상 방안을 제안하였다. 구체적으로 사용자의 효율적 브라우징을 위한 메뉴의 개선 방안을 제시하였으며, 구성원 검색기능의 추가를 제안하였다. 프로세스 마이닝은 작업 단위에 대한 분류와 해당 작업들이 누구에 의해서 이루어 졌는지에 대한 Case ID와 작업이 이루어진 시간인 Time Stamp가 명확해야 한다. 웹 로그라는 데이터의 특성상 ID부분에서 특정 사용자를 개별로 명확히 구분하는 것이 매우 힘들었다. IP의 경우 일반적인 사용자들은 고정IP보다 유동IP들을 많이 사용하며, 학교라는 특성상 개인의 PC가 지정이 되어있는 것보다 공용적으로 사용하기 때문에 IP라는 정보만 가지고서 사용자를 구분하기가 힘들었다. 이로 인해서 외부 사용자와 내부 사용자의 패턴을 구분하여 분석하는데 한계가 있다. 본 연구에서는 최소한의 구분을 위해 IP와 Time Stamp를 활용하여 ID라는 새로운 변수를 생성을 하는 방법을 사용하였다. 향후 연구에서는 사용자들에 대한 트랜잭션 구분에 대한 방법이 연구 되어야 한다. 각 Case에 대한 ID를 명확히 구분을 하고 ID마다의 트랜잭션을 식별하는 방법에 대한 연구가 필요할 것으로 사료된다. 또한, 이 사례는 특정 대학의 홈페이지 접속 로그를 활용한 결과이므로 타 대학의 사례에 직접 활용하기 어렵지만 웹 로그 분석과정과 분석결과 활용방안에는 기여할 수 있을 것으로 기대된다.

참 고 문 헌

[1] 고석하, 김주성, “김영기, 웹사이트의 구조와 정보량 및 사용자 과업 복잡도가 사용성에 미치는 영향”, *Journal of Information Technology Applications and Management*, 제12권 제2호, 2005, pp. 145-161.

[2] 김광용, 고급분석과 eCRM, 시대의 창, 2002.

[3] 김광훈, “워크플로우 소셜 네트워크 인텔리전스 발견 알고리즘”, *인터넷정보학회 논문지*, 제13권 제2호, 2012, pp. 73-86.

[4] 김동곤, “빅데이터 로그분석 활용사례에 관한 연구”, 전남대학교 대학원, 석사학위논문, 2013.

[5] 김민정, “사례 중심의 웹 로그 마이닝 활용에 관한 연구”, 이화여자대학교 대학원, 석사학위논문, 2002.

[6] 김재훈, “웹 로그 분석을 이용한 실시간 온라인 마케팅 시스템 설계 및 개발에 관한 연구”, *한국전자거래학회지*, 제16권 제3호, 2011, pp. 249-261.

[7] 김정현, “학습자 시간관리 전략과 학업성과 간 관계분석 : 학습분석학적 접근”, 이화여자대학교 대학원, 석사학위논문, 2013.

[8] 남궁영, “웹 로그 분석을 통한 이러닝 효율화 방안에 관한 연구”, 단국대학교 대학원 석사학위논문, 2005.

[9] 문경선, “효과적인 사이트 구현을 위한 로그 분석에 대한 연구”, 세종대학교 대학원, 석사학위논문, 2002.

[10] 문현주, “웹 사용성 관점에서 공공기관 웹사이트의 링크 유효성 분석 및 개선 과제”, *재활복지*, 제17권 제4호, 2013, pp. 291-309.

[11] 송민석, “프로세스 마이닝을 활용한 생산 공정 데이터 분석”, 2012 대한산업공학회 춘계학술대회논문집.

[12] 전재환, “웹 로그 파일을 이용한 사용자 특성 분석에 관한 연구”, 서울산업대 산업대학원, 석사학위논문, 2002.

[13] 전종근, 박 철, “웹 로그 데이터를 이용한 온라인 소비자의 가격민감도 영향 요인에 관한 연구”, *Journal of Information Technology Applications and Management*, 제13권 제1호, 2006, pp. 1-16.

[14] 정민수, “중소기업을 위한 BPM의 도입 및 적용방안”, 홍익대학교 대학원 석사학위논문, 2010.

[15] 하현식, “프로세스 마이닝 기법을 이용한 프로세스 개선 방안에 관한 연구”, 부산대학교 대학원 석사학위논문, 2012.

[16] Christian, W. Gunther, Fuzzy Mining Adaptive Process Simplication Based on Multi-Perspective Metrics, International Conference on Business Process Management (BPM 2007), Lecture Notes on Computer Science, Vol. 4714, 2007, pp. 328-343.

[17] Cooley, R., Mobasher, B., and Srivastava, J., “Data preparation for mining world wide web browsing patterns”, *Knowledge and Information System*, Vol. 1, No. 1, 1999, pp. 123-132.

[18] Garter, Enterprises Focus on Business Process Management, Yefim Natis, 2003.

[19] Smith, H. and Fingar, P., Business Process Management : The Third Wave, Meghan-Kiffer Press, 2003.

[20] Srivastava, J., Cooley, R., Deshpande, M., and Tan, P., “Web Usage Mining : Discovery and Applications of Usage Patterns from Web Data”, *ACM SIGKDD*, Vol. 1, No. 2, 2000, pp. 1-12.

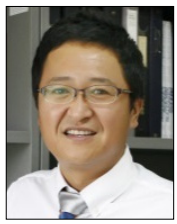
[21] van der Aalst W. M. P., “Workflow Mining : Discovering Process Models from Event Logs”, *IEEE Transactions on Knowledge and Data Engineering*, Vol. 16, No. 9, 2004, pp. 1128-1142.

[22] van der Aalst W. M. P., Process Mining : Discovery, Conformance and Enhancement of Business Processes, Springer Heidelberg Dordrecht, New York, NY, 2011.

저자소개

이 용 욱

충북대학교 정보통계학과 학사, 비즈니스데이터융합학과에서 석사 학위를 취득하였으며, 사이버다임에 재직 중이다. 관심분야는 빅데이터 분석, 데이터마이닝, 프로세스 마이닝 등이다.



최 상 현

한양대학교 공학사, 한국과학기술원 산업공학과 석사, 경영정보공학 박사를 취득하였고, 아리조나주립대에서 박사 후 연구과정을 수행하였다. 현재 충북대학교 경영정보학과 교수로 재직 중이며, 한국빅데이터서비스 학회 부회장, 한국CRM협회 기술위원으로 활동 중이다. LG CNS 엔트루컨설팅에서 CRM 전략 컨설팅 및 시스템 구축, 정보화 전략 계획 수립, ERP 시스템 구축 등의 IT 컨설팅을 수행하였으며, 주요 관심분야는 빅데이터, 데이터마이닝, 전략적 의사결정 시스템 등이다.