컨소시엄 기반 전자저널 이용통계 수집 및 분석 개선 방안*

Improving Efficiency of Usage Statistics Collection and Analysis in E-Journal Consortia

정영임(Youngim Jung)**
김정환(Jeonghwan Kim)***

초 록

전자저널의 활용이 급속히 증가하면서 도서관에서는 자관에서 구입되는 전자저널이 얼마나, 어떻게 활용되고 있는지에 대한 관심이 증가하였다. 또한 전자정보 컨소시엄 주관기관에서도 컨소시엄 내에서 유통되는 학술자원의 이용통계에 대한 분석을 통해 국가 차원의 전자학술저널의 유통 현황 파악 및 수요자 중심의 정보수집 정책 개발이 필수적이다. 그러나 기존의 수작업에 의존한 이용통계 수집과 출판사에서 제공하는 저널 이용통계 보고서만으로는 이용에 대한 포괄적이고 심층적인 분석이 불가능하다. 이에 본 연구에서는 대용량 이용통계 수집 및 분석의 기반 마련을 위해 스크린 스크래핑과 SUSHI 프로토콜을 적용한 전자저널 이용통계 자동수집 시스템을 구현하였다. 또 저널 서지정보 및 컨소시엄 계약 데이터베이스를 연동하여 심층적인 이용통계 분석정보를 생성할 수 있는 방안을 제안하였다.

ABSTRACT

The proliferating use of e-journals has led increasing interest in collecting and analyzing usage statistic information. However, the existing manual method and simple journal usage reports provided by publishers hinder the effective collection of large-scale usage statistics and the comprehensive/in-depth analysis on them. Thus we have proposed a hybrid automatic method of collecting e-journal usage statistics based on screen scraping and SUSHI protocol. In addition, the generation method of summary statistics presented in graphs, charts and tables has been suggested in this study. By utilizing the suggested system and analysis data, librarians can compose various reports on budget or operation of the libraries.

키워드: 전자저널 이용통계, 스크린 스크래핑, SUSHI, COUNTER, 통계분석
e-journal usage statistics, screen scraping, SUSHI, COUNTER, statistical analysis

---

* 본 연구는 '핵심 과학기술 전자정보 공동활용 체계 구축(K-12-L01-C01-S01)' 사업의 일환으로 수행되었음.
** 한국과학기술정보연구원 해외정보실 선임연구원(acorn@kisti.re.kr) (제1저자, 교신저자)
*** 한국과학기술정보연구원 해외정보실 책임연구원(kimjh@kisti.re.kr) (공동저자)
■ 논문접수일자 : 2012년 2월 23일 ■ 최초심사일자 : 2012년 2월 28일 ■ 게재확정일자 : 2012년 4월 11일
■ 정보관리학회지, 29(2): 7-25, 2012. 〔http://dx.doi.org/10.3743/KOSIM.2012.29.2.007〕

1. 서 론

학술 정보의 생산 및 유통 구조가 전자환경으로 변화함에 따라 대부분의 학술저널이 전자화되고 컨소시엄을 통해 패키지화되어 유통되고 있다. 또, 전자저널에 대한 활용이 급속히 증가하면서 도서관에서는 자관에서 구입되는 전자저널이 얼마나, 어떻게 활용되고 있는지에 대한 관심이 증가하였다. 전자저널의 이용통계를 저널을 생산하고 공급하는 출판사에서 주로 생성하고 있다. 각 출판사에서 생성된 이용통계는 웹 사이트에서 각 도서관 담당자가 직접 조회를 하거나 혹은 도서관의 요청에 따라 출판사에서 e-mail을 통해 전달하기도 한다. 그러나 각 도서관에서는 다수 출판사의 전자저널를 구독하고 있기 때문에, 담당자가 주기적으로 각 출판사 사이트를 일일이 방문하여 이용통계를 수집하기가 번거롭고, 수집한 이용통계 데이터를 표준화된 포맷으로 재가공하여 일괄 분석하는 데 드는 시간과 비용의 낭비가 적지 않다. 또한 해외 학술저널을 컨소시엄화하여 공급하고 있는 컨소시엄 주관기관에서도 컨소시엄 내에서 유통되는 전자학술자원의 이용통계에 대한 모니터링을 통해 국가 차원의 전자학술자원의 유통 현황 파악 및 수요자 중심의 정보수집 정책 개발이 필수적인 상황이다.

따라서 본 논문에서는 이용통계 데이터 수집 방법의 자동화와 데이터형식의 표준화 등 대용량 이용통계 분석을 위한 기반 마련을 위해 스크린 스크래핑 엔진에 기반한 전자저널 이용통계 자동수집 시스템을 구현하고자 한다. 본 시스템을 통해 개별 도서관에서는 이용통계를 수집하기 위해 여러 출판사 웹 사이트를 일일이 방문하여 정보를 취합하고, 다양한 유형의 이용통계 보고서를 하나의 형태로 통일하는데 드는 시간과 비용을 크게 줄일 수 있고, 컨소시엄 주관기관에서는 컨소시엄에 참여하는 도서관들의 이용통계를 일괄 수집하여 컨소시엄 전체 이용현황을 파악한 결과를 이용해 정보수집 및 유통 정책을 세울 수 있어, 주관기관으로써의 역할을 극대화할 수 있다.

본 논문의 구성은 다음과 같다. 먼저 제2절에서 전자정보 이용통계의 이론적 배경 및 관련 표준 활동에 대해서 소개한다. 제3절에서는 현재 출판사를 통해 제공되고 있는 전자저널 이용통계의 현황 및 그 문제점에 대해 분석한다. 제 4절에서는 전자저널 이용통계 자동수집 시스템의 구현을 통해 대용량 이용통계 수집 개선 방안을 논의하며, 제 5절에서는 자동수집을 통해 구축된 대용량 이용통계 데이터를 다양한 데이터베이스와 연동함으로써 포괄적인 통계 집계 정보와 심층적인 통계분석 자료를 생성할 수 있는 방안에 대해 기술하였다. 마지막으로 제 5절에서는 본 연구의 결론과 향후 과제를 논의하며 끝을 맺는다.

2. 선행 연구

2.1 전자저널 이용통계 연구의 배경

ARL 연구에 따르면 저널 인플레이션은 매년 6~12% 증가하며 도서관의 구매력을 저하시켰다. 또한 컴퓨터 네트워크 기술의 발달로 학술커뮤니케이션 환경이 전자출판 기반으로 변화하면서 도서관은 급속히 증가하는 전자저널에

대한 정보 부족, 담당 인력의 부족, 새로운 패러다임에 접근하는 대응력 부족 등으로 전자저널 구독에 어려움을 겪었다(심원식, 2005).

이에 대한 대응으로 전 세계 도서관은 구매력 및 정보공급사에 대한 협상력을 제고하기 위해 전자정보 공동구매 컨소시엄과 같은 협력체계를 구축해 왔다. 이는 전자저널 구독 시 발생되는 추가 비용을 최소화하고 한정된 예산으로 최적의 전자정보를 구입하고자 하는 노력이다(김성진, 정은경, 한민혜 2008). 도서관 컨소시엄 국제연합기구인 International Coalition of Library Consortia(이하 ICOLC라 칭함)에 등록된 전자정보 공동구매 컨소시엄은 전 세계적으로 330개가 넘는다. 300여 개의 다양한 형태와 규모의 컨소시엄이 전 세계적으로 운영되고 있지만, 컨소시엄을 통한 협력만으로는 현재 학술 커뮤니케이션이 처한 여러 문제를 해결하기에는 역부족이다. 또한 컨소시엄에 참여하는 회원 기관들은 컨소시엄 주관기관이 더 이상 가격협상과 같은 부분적이고 소극적인 대응에 머무를 것이 아니라 변화된 학술환경에 적극적으로 대응하여 컨소시엄에 참여하는 회원기관의 수혜의 범위가 확대될 수 있도록 보다 발전된 컨소시엄 운영모델을 기대하고 있다(김정환, 이응봉 2009).

이에 세계적으로는 ICOLC를 중심으로 컨소시엄 운영자들이 고등교육기관 도서관, 전문도서관 등의 회원 간 공동의 관심사와 공익의 문제에 대한 토론을 활발히 하며, 매년 2회에 걸친 회의를 통해 전자정보 가격 협상 문제 등 다양한 이슈를 알리고 이를 공동으로 해결하고자 노력하고 있다.

국내에서도 다양한 연구를 통해 국내 컨소시엄 운영 모델의 문제점을 밝혀 내고, 이에 대한 대응방안을 제안하고 있다. 김성진, 정은경, 한민혜(2008)는 전자저널 구독과 관련해 국내 컨소시엄 주관기관이 해결해야 할 쟁점 사항을 다음과 같이 정리하였다.

- 빅딜 거래와 가격 모형
- 라이선스 계약 및 표준계약서 개발・적용
- 전자저널 이용통계 제공 및 활용
- 전자저널 원문 장기보존

이상의 쟁점 사항들의 대응방안으로 '이용통계에 기반한 저널 패키지의 재조정(빅딜 거래 대응)', '이용량에 기반한 새로운 가격 모형 수립 및 적용', '저널 구독 및 취소 결정 기준으로써의 이용통계 활용' 등이 관련 연구에서 제시되었다. 이와 같이 전자저널 이용을 계량화한 통계 자료는 다양한 문제를 해결하고 의사 결정을 보조하기 위한 기반 자료로 그 중요성이 날로 더해져 가고 있다(정영임, 김정환, 류범종 2012).

심원식(2005)은 전자정보를 제공하고 지원하는 비용이 도서관 전체 예산의 대부분을 차지함에 따라 도서관 운영의 분석, 보고 및 의사결정을 지원하기 위해 전자정보 이용통계의 활용이 중요함을 지적하였다. 그럼에도 불구하고, 이용통계의 활용에는 많은 한계가 있다. 황옥경(2007)은 국내 48개 대학 도서관을 대상으로 전자저널 이용통계가 제공되는 현황 및 활용 현황을 설문조사를 통해 파악하였다. 주로 도서관 측 요청에 의해 정보공급사로부터 이용통계 데이터가 제공되고 있었고, 제공되는 데이터의 내용 및 형식에 대한 만족도는 반수에 가까운 기관이 매우 불만 혹은 다소 불만(47.8%)을 표하였다. 불만족 이유로는 '출판사마다 제공되는 데이터 형식이 달라 비교가 어려웠다'는 응답

이 61.1%를 차지해 정보공급사별 이용통계 리포트 세부 항목의 차이와 형식의 비표준화가 이용통계 활용에 가장 큰 걸림돌이 됨을 파악할 수 있다(정영임, 김정환 2012).

2.2 이용통계 관련 표준 활동

온라인 정보자원의 활용이 급속히 증가하면서, 정보자원 이용량에 대한 일관적이고 표준화된 통계 구축의 필요성에 대한 인식이 널리 확산되고 있다. 본 절에서는 정보자원의 이용통계 산출 항목, 통계 리포트 형식 및 통계 데이터 전송 프로토콜의 표준화 시도에 대해 기술한다.

2.2.1 이용통계 세부 항목

ICOLC에서는 1998년 정보자원 이용통계 제공 가이드라인을 제시하였고, 이는 2001년과 2006년에 개정되었다. 2006년 개정안에서는 벤더는 과거 데이터를 최소 3년은 유지해야 한다고 하였다. 표준화 항목에 있어서 1998년과 2006년 모두 세션 수(number of sessions: logins), 질의 수(number of queries: search), 메뉴 선택의 수(number of menu selection, number of turnaway)를 필수적 기본요소로 정의하고 있으나, 검토된 자료의 수(number of items examined)의 정의에 있어 2006년 가이드라인에서는 다르게 언급하고 있다(ICOLC, 2006). 접근, 전달 체계, 보고 형식에 있어서는 컨소시엄 멤버 단위, 컨소시엄 내의 저널 단위, 개별 기관별, 컨소시엄 데이터베이스별 보고를 공식적으로 제안한 표준통계 형식을 추가하고 있다.[1)]

2.2.2 리포트 형식의 표준화

각 정보공급사에서 상이한 포맷으로 제공되는 이용통계정보가 일관되고 신뢰할 수 있으며, 서로 비교가능할 수 있도록 국제적인 수준의 이용통계 리포트 형식 표준을 제공함을 목표로 COUNTER(Counting Online Usage of Networked Electronic Resources) 프로젝트가 시작되었다. 2002 12월 시행령 1판의 초안이 마련되었고, 2004년 4월에 발표된 2판의 초안에 대해 의견 조율을 거쳐 2005년 시행령 2판의 최종버전이 발표되었다. 그리고 2006년 1월부터 실제 효력을 발휘하였다. COUNTER 시행령은 측정되어야 하는 데이터 요소, 요소에 대한 정의, 이용통계 보고서의 내용과 형식, 데이터 처리를 위한 요구사항, 회계감사를 위한 요구사항, 중개 게이트웨이 및 DB 생산자가 있을 경우 중복 통계를 위한 지침 등을 포함한다. COUNTER에서 제안하는 이용통계 보고서 형식은 저널 보고서(Journal Report 1, Journal Report 1a, Journal Report 2)와 DB 보고서(Database Report 1, Database Report 2, Database Report 3), 단행본 보고서(Book Report 1, Book Report 2, Book Report 6 등)가 있다.

2.2.3 데이터 전송 방식의 표준화

ICOLC의 이용통계 가이드라인이나 COUNTER 시행령에서 표준화된 이용통계 보고서 항목과 포맷에 대한 기준을 제시하고 있으나, 다수의 출판사로부터 이용통계 보고서를 자동으로 수집 및 처리에 대한 요구도 증가하고 있다.

1) 각 국가별 ICOLC 가이드라인의 채택비율을 살펴보면, 1998년 당시 ICOLC 가이드라인을 채택한 컨소시엄은 총 52개였고, 2006년 가이드라인을 채택한 컨소시엄은 85개로 그 수가 증가하였다.

이에 따라 2005년 여름에 도서관과 도서관 시스템업체 및 COUNTER 제공 출판사가 협력하여 XML형식의 COUNTER 데이터를 시스템 간 자동으로 반출입할 수 있는 규약을 만들기 위한 실무그룹을 만들었다. 생성된 통계자료의 수집과 배포의 용이함을 위해 SOAP을 기반으로 한 The Standardized Usage Statistics Harvesting Initiative(이하 SUSHI)가 나오게 되었다. NISO의 SUSHI는 프로토콜 표준 (ANSI/NISO Z39.93-2007)은 COUNTER 보고서 및 SUSHI Report registry에 등록된 XML 기반의 보고서를 자동으로 전송받을 수 있도록 한 웹 서비스 기반 프로토콜이다. SUSHI 표준을 적용하면 각 벤더의 이용통계 조회 웹사이트에 개별적으로 접속해 통계데이터를 검색, 조회, 다운로드하는 반복적인 수작업이 필요 없고, 표준화된 보고서 포맷인 COUNTER 보고서로 이용통계 데이터를 수집할 수 있다. COUNTER 보고서 포맷이 아니더라도 SUSHI에 등록된 포맷의 보고서이거나 XML로 작성된 이용통계 보고서 역시 SUSHI를 통해 전송받을 수 있다(NISO, 2007).

3. 전자저널 이용통계 제공의 문제점

국내 과학기술의학 분야 컨소시엄을 통해 전자저널을 유통하고 있는 출판사는 2011년 기준 104개사이며, 그 중 31개 출판사의 이용통계 데이터 제공 현황은 〈부록〉과 같다. 국내 다수 기관에서 구독하고 있는 출판사는 대부분 이용통계 보고서를 제공하고 있으나, American Mathematical Society와 일본의 Institute of Pure and Applied Physics 등 7개 학회로 구성된 'Japan Science & Technology Electronic Journal' 컨소시엄의 이용통계는 제공되지 않고 있다.

본 절에서는 출판사에서 제공되는 전자저널 이용통계의 문제점을 유형별로 살펴보고자 한다.

3.1 상이한 이용통계 제공 방식

정영임과 김정환(2012)에서 분석한 바와 같이 각 출판사에서 생성된 이용통계가 제공되는 방식은 (1) 영문으로 된 해외 출판사 웹 사이트 혹은 이용통계 조회용 사이트에서 각 도서관 담당자가 이용통계를 직접 조회/다운로드 하거나, (2) E-mail을 통해 제공받거나, (3) 사서가 요청을 하면 비정기적으로 이용통계 보고서 파일을 전달받거나, (4) 도서관 전산화 솔루션에 구현된 SUSHI 기반 이용통계 수집 기능으로 제공받는 등 이용통계 제공 방식이 제각각이다.

출판사마다 이용통계 제공방식이 다른 것은 이용통계 데이터 생성 및 제공을 위해 출판사별로 다른 소프트웨어를 사용하기 때문이다. 예를 들어, Elsevier, Springer, Jstor, John Wiley & Sons, Inc(이하 Wiley) 등은 자체 소프트웨어를 사용하고 있고, Nature Publishing Group (이하 NPG), IEEE, Institute of Physics Publishing Ltd(이하 IOP), Royal Society of Chemistry(이하 RSC) 등은 'mpsight', American Institute of Physics(이하 AIP), American Physical Society(이하 APS), American Society of Civil Engineers(이하 ASCE), International

Society for Optics and Photonics(이하 SPIE) 등은 'scitation'이라는 전문 소프트웨어를 사용하고 있다.

이처럼 출판사별로 서로 다른 정보 제공 방식과 소프트웨어를 사용하므로, 이용통계를 수집하기가 여간 번거롭지 않다. 대부분의 출판사가 웹을 통해 이용통계 보고서를 제공하기는 하지만, 출판사별로 이용통계를 조회할 수 있는 영문 웹 사이트의 인터페이스도 다르기 때문에 해당 사이트를 처음 방문하게 되는 경우 로그인 단계에서부터 이용통계 보고서를 검색하고 수집하는 데 많은 어려움을 겪는다. 또한 이용통계 보고서 제공일이 일정하지 않아[2] 사서가 월말 혹은 월초에 각 출판사 사이트를 일일이 방문하여 이용통계가 제공되는지를 확인해야 하는데, 이처럼 비정기적으로 제공되는 통계자료를 관리하는 일은 사서에게 큰 부담이 아닐 수 없다.

3.2 자료의 일관성 결여

황옥경(2007)이 밝힌 바와 같이 출판사로부터 구독 중인 전자저널의 이용 데이터를 제공받은 적이 있는 도서관을 대상으로 이용 데이터의 내용 및 형식에 대한 만족도는 '매우 불만(2.3%)'과 '불만(45.5%)'이라는 응답이 '다소 만족(15.9%)' 또는 '매우 만족(9%)'에 압도적이었다. 불만족 이유는 '출판사마다 제공되는 데이터의 형식이 달랐기 때문에 비교가 어려웠다(61.1%)'로 가장 많았다. 〈부록〉에서 보는 바와 같이 출판사에서 제공되는 이용통계 보고서의 파일 유형이 csv, xls, html, pdf, xml 등으로 다양하다. 여러 출판사의 통계 데이터를 비교 분석하려면 이처럼 다양한 파일 유형은 우선 하나의 파일 유형으로 처리하여야 한다. 또, AIP, AAAS와 같은 출판사에서 제공하는 일부 과거 이용통계 데이터는 COUNTER 이용통계 보고서 표준 포맷을 적용하지 않았다.

3.3 분석 데이터 미제공

황옥경(2007)의 연구에 따르면 '전자저널 이용 기관 전체의 평균 이용 데이터'에 대한 요구는 '원문 다운로드 건수', '원문 열람 건수'에 이어 세 번째로 높다. 그러나 현재 이와 관련한 데이터를 제대로 제공받지 못하고 있다. 컨소시엄 관리자 권한으로도 컨소시엄 전체 기관의 평균 이용률을 볼 수 없고, 모든 기관의 데이터를 다운로드 받아서 재합산을 해야 한다는 어려움이 있다.

또한 〈부록〉에서 기술된 바와 같이 대부분의 출판사에서 각 도서관의 이용통계에 대한 보고서는 표 형식으로 제공하고 있으나, 이를 시각화하여 재가공한 자료나 타기관 대비 이용량 및 시기에 따른 전자정보 이용 추이 등의 분석 자료는 전혀 제공하고 있지 않다.

Cambridge University Press(이하 CUP)만 'popular papers'라 하여 지정된 기간 중 가

2) 이용통계 보고서는 월별로 생성되기 때문에 제공되는 시점은 보통 해당월+1개월(+a)가 된다. 그런데 출판사별로 보고서 제공시점이 다르며, 정기적이지가 않다. 예를 들어, American Chemical Society(이하 ACS), APS, Elsevier는 2011년 7월 통계가 9월 1일에 제공된 반면, AIP, IEEE, IOP, NPG, American Association for the Advancement of Science(이하 AAAS라 칭함), Springer, Wiley 등은 제공되지 않고 있다. 기존 방식으로는 담당자가 매일, 매주 확인을 해야 하는 상황이라 낭비되는 시간이 많다.

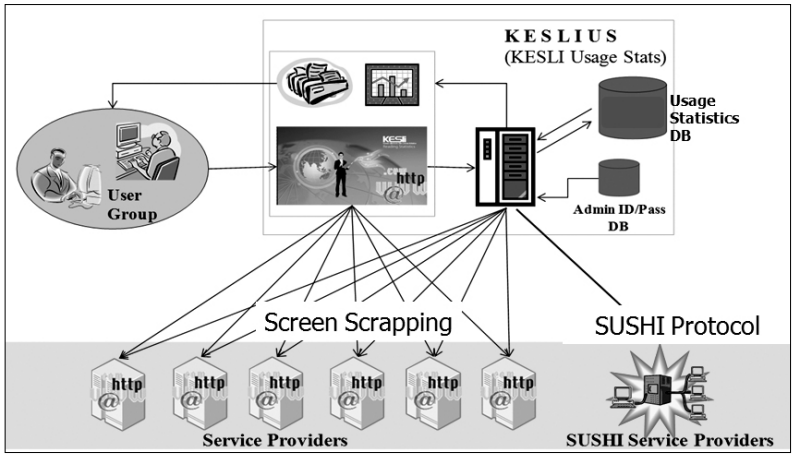
장 많이 이용된(다운로드된) 논문을 저널별/전체저널에서 산출하여 'top-10/25/50'으로 제공하고 있고, 나머지 정보공급사에서는 이러한 이용통계 분석 자료를 제공하지 않는다.

4. 전자저널 이용통계 수집 개선 방안

3절에서 살펴본 바와 같이 각 기관의 담당자가 다수의 출판사 웹 사이트를 일일이 방문하여 비정기적으로 제공되는 이용통계 데이터를 수집하는 것은 많은 비용과 시간이 소요된다. 따라서 기존에 수작업으로 이루어지고 있는 전자저널 이용통계 수집의 비용을 절감하고 효율을 높이기 위해 본 연구에서는 각 출판사의 전자저널 이용통계를 자동으로 수집할 수 있는 시스템을 개발하였다. 개발한 시스템의 개념도는 〈그림 1〉과 같다.

4.1 수집 대상 정보자원의 선정

본 시스템을 통해 수집하고자 하는 이용통계 데이터는 국내 과학·기술·의학 분야 전자저널 컨소시엄을 통해 가장 많이 구독되는 출판사 순으로 하되[3] 웹 사이트나 이메일 혹은 SUSHI 프로토콜을 통해 이용통계를 제공하는 출판사의 이용통계를 대상으로 하였다. 선정된 출판사는 Elsevier, Wiley, Springer, AAAS, IEEE 등을 포함한 31개 출판사이다(부록 참조). 이용통계 데이터는 과거년도의 이용통계를 소급하여 최신 데이터까지 각 출판사에서 제공하는 기간의 통계 데이터는 모두 수집한다. 대부분의 출판사에서 2008년도부터의 이용통계 데이터를 제공하고 있으며 이들 데이터는 세계적 표준 양식인 COUNTER JR1을 준용한 이용통계 보고서(Journal Usage Report)로 생성되고 있다. 출판사별로 JR1a, JR4, CR1 등을 추가로 제공하기도 하지만, 본 시스템에서는 31개 출판사



〈그림 1〉 전자저널 이용통계 자동수집 시스템의 개념도

3) 〈http://www.kesli.or.kr/kesliindex.html〉.

가 공통적으로 제공하는 JR1 포맷으로 제공되는 이용통계 보고서만 수집 대상으로 한다. 그리고 3.2.2절에서 밝힌 바와 같이 AIP, AAAS, Wiley 등 일부 출판사에서는 특정년도 이전의 과거 이용통계 데이터를 COUNTER 표준 포맷이 아닌 출판사 자체 양식에 따라 생성하여 구축하였다. 이러한 비표준 이용통계 데이터도 수집하여 세부 항목을 비교 분석한 후 본 시스템에서 COUNTER JR1 포맷으로 변형하여 저장한다.

4.2 복합적 자동수집 방식

전자저널 이용통계를 자동으로 수집하기 위해 본 연구에서는 (1) 출판사 웹 사이트에 접속하고 이용통계 조회화면 혹은 E-mail에서 이용통계 보고서 데이터를 추출하는 스크린 스크래핑(Screen scraping) 기술을 이용한 방식과 (2) 이용통계 데이터 전송 프로토콜인 SUSHI 표준을 적용한 방식을 복합적으로 적용하여 전자저널 이용통계 자동수집 시스템을 개발하였다. 각 방식에 대해서 4.2.1절과 4.2.2절에서 자세히 기술한다.

4.2.1 스크린 스크래핑 기반 이용통계 자동수집

스크린 스크래핑은 자동으로 시스템에 접속해 데이터를 화면에 나타나게 한 후 필요한 자료만을 추출해 가져오는 기술이다. 웹사이트에 있는 정보를 끄집어내 다른 사이트나 데이터베이스에 저장하기 때문에 웹 스크래핑(Web Scraping)이라고도 한다. 인터넷 뱅킹에 필수적인 프로그램으로 각 금융기관에서 활발하게 운영 중이며, 호텔과 항공사・렌터카・주유소 등의 마일리지와 같은 보상프로그램, 전자우편 통합 조회, 뉴스・채팅・날씨 등 사용자가 클릭해 정보를 얻을 수 있는 프로그램 등에 널리 사용되고 있는 기술이다.[4)]

본 시스템에서 스크린 스크래핑 방식으로 이용통계를 수집하는 출판사는 Springer, Wiley, Elsevier 등 대형 출판사를 포함한 20개사이다(부록 참조). 스크래핑 엔진에서는 확보된 이용통계 조회 계정 정보를 이용하여 각 정보공급사 통계조회 사이트에 자동 로그인하고, 자동수집 일정에 따라 스크래핑 엔진을 실행하여 전자정보 이용통계 데이터를 추출한다(정영임 외, 2011).

AIP&APS는 통합된 출판사이나 이용통계 제공 사이트는 독립적으로 운영 중이므로 본 통계수집 시스템에서도 분리하여 관리한다. SPIE는 Springer에서도 SPIE 출판사 발간 저널에 대한 이용통계를 제공하고 있고 SPIE 자체 웹 사이트에서도 이용통계를 제공하고 있어 두 출판사에서 제공되는 이용통계를 모두 수집한다. 대부분의 출판사는 웹 사이트에서 이용통계 보고서를 제공하고 있으나, ACS, Sage Publications(이하 Sage)의 경우 e-mail로 이용통계를 제공하고 있어, e-mail 스크래핑 모듈도 구현하였다. 이용통계 보고서의 파일 포맷은 HTML, CSV, PDF, XML, XLS 등 다양하며, 출판사별로 이용통계 보고서 양식이 달라 각 이용통계 보고서 DB를 분석하여 이용통계를 수집한

4) 〈http://100.naver.com/100.nhn?docid=769989〉.

다. 스크래핑 엔진은 스케쥴러에 기록된 자동 수집 일정(월 1회)에 따라 자동 실행되며 자동 수집일에 수집이 실패하면 다음 정해진 2차 일정(주 1회), 3차 일정(일 1회)대로 수집이 성공할 때까지 수집을 시도한다. 그리고 실패 원인을 기록하여 시스템 관리자가 실패 원인을 파악하고 문제를 해결할 수 있도록 하였다. 네트워크 오류 및 출판사 서버 등의 일시적인 오류에 의한 수집 실패는 일정 시간이 지난 후 대부분 복구되기 때문에 2, 3차 시도에서 수집이 성공적으로 이루어지며, 출판사 통계 시스템 변경, 통계조회 계정 오류 등에 의한 수집 실패는 시스템 관리자가 해당 변경 사항을 적용하거나 오류를 복구한 후 수집을 실행할 수 있도록 비정기 수집 기능도 지원한다.

스크린 스크래핑에 기반한 자동 수집 방식은 근본적으로 다음과 같은 한계를 가진다. 출판사별로 이용통계 전송방식을 확인하고 이용통계 데이터를 스크래핑할 수 있는지에 대한 가능 여부를 분석하는 데 시스템 개발 초기에 많은 시간이 소요된다. 본 연구에서는 한 개 출판사당 최소 일주일 정도의 시간이 소요되었으며, 동일한 상용 이용통계 제공 소프트웨어를 쓰더라도 출판사별로 이용통계 데이터가 저장된 디렉토리 등 세부 정보는 다르기 때문에 이에 대한 분석에도 2-3일이 소요되었다.

또 출판사에서 이용통계를 제공하는 웹 사이트 및 시스템을 변경하거나 도서관 사서들이 통계 조회용 계정 정보를 변경하면 스크래핑에 의한 통계 수집이 이루어지지 않기 때문에 시스템 관리자는 지속적인 모니터링을 통해 자동 수집 실패 원인에 따른 조치를 취해야 한다. e-mail 스크래핑은 출판사에서 e-mail 발송을 지연하면 본 스크래핑 엔진에서 e-mail 수신 대기에 상당한 시간을 사용하므로 전체 시스템 부하가 커지기도 한다.

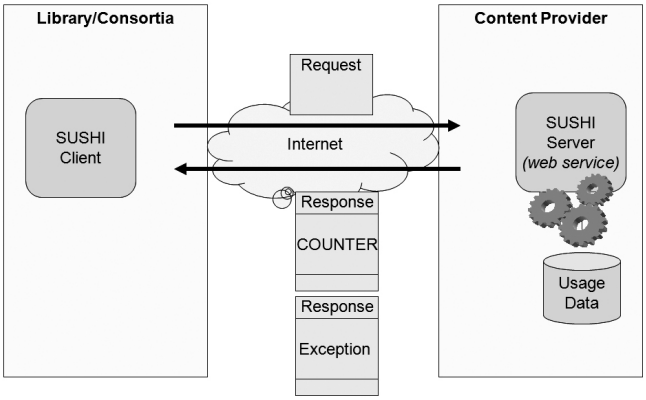
마지막으로 스크린 스크래핑 방식의 특성상 간혹 출판사 서버에서 스크래핑 엔진을 무허가 웹 로봇이나 해커로 인식하여 서버 IP 이동 및 스크래핑 엔진 차단 등의 방어를 하는 경우도 있으므로 스크래핑에 의한 이용통계 자동수집은 출판사와의 사전 협의 및 기술진과의 의사소통을 통해 이루어져야 한다.

4.2.2 SUSHI 기반 이용통계 자동수집

스크린 스크래핑 방식이 가지는 근본적인 한계를 보완하고 자동수집 시스템 개발 및 운영의 효율성을 높이기 위해 본 연구에서는 SUSHI 표준을 추가로 적용하였다. 〈그림 2〉와 같이 SUSHI 프로토콜 표준을 채택한 출판사 서버에 클라이언트의 이용통계 요청 메시지(Request Report)가 접수되면 서버에서 이용권한[5]을 확인한 후 결과값으로 요청한 이용통계 데이터를 담아 응답 메시지(ResponseReport)를 전송한다.

본 연구에서는 Annual Reviews, Cambridge University Press 등 11개 출판사의 이용통계 수집을 위한 11개 SUSHI-client를 개발하였다

5) 해당 출판사의 전자저널을 구독하는 도서관은 이용통계 접근권한을 가지며, 권한이 없는 클라이언트로부터 요청 메시지가 접수되면 출판사 SUSHI 서버에서는 결과값으로 '서비스 접근권한이 없음(Requestor Not Authorized to Access Service)' 혹은 '해당 기관 이용통계 접근권한이 없음(Requestor is Not Authorized to Access Usage for Institution)'이라는 오류 메시지(Exception)를 반환한다.



〈그림 2〉 SUSHI 프로토콜을 통한 이용통계 전송

(부록 참조). 스크린 스크래핑 엔진과 마찬가지로 SUSHI client 역시 스케줄러에 정해진 일정에 따라 각 출판사의 SUSHI 서버에 이용통계 요청 메시지를 전송하고 이용통계 보고서를 전송 받는다.

SUSHI 서버가 안정적으로 구축되고 운영된다는 전제 하에서 SUSHI에 기반한 이용통계 자동수집은 아래와 같은 장점이 있다. 우선 이용통계 데이터의 전송에 대해서는 SUSHI 표준에 대한 분석만 하면 되기 때문에 시스템 개발 시간이 단축된다. 시스템 운영에 있어서도 XML에 기반한 SUSHI 프로토콜의 특성상 내부 파라미터를 변경하더라도 client 프로그램의 변경이 거의 필요하지 않기 때문에 운영·유지 비용이 절감된다. 둘째로 SUSHI를 통한 데이터 전송은 수밀리초에서 수초 이내에 이루어지기 때문에 스크린 스크래핑 방식에 비해 통계 수집에 소요되는 시간이 현저히 줄어든다. 셋째, 출판사 이용통계 웹 사이트에 로긴하기 위한 계정 정보와는 달리 시스템 간 이용통계를 주고 받기 위해 필요한 SUSHI ReferenceID는 거의 변경되지 않기 때문에 계정 정보 변경으로 인해 이용통계를 수집하지 못하는 경우는 발생하지 않는다. 마지막으로 SUSHI client에서 보낸 이용통계 요청 메시지가 접수되는 순간 각 출판사 SUSHI server에서는 해당 요청자(requestor)의 권한을 확인하여 이용통계 보고서 혹은 오류 메시지를 응답으로 보내주므로 원인불명의 서비스 중단은 일어나지 않는다.

그러나 SUSHI 프로토콜을 통해 수집된 결과물이 XML 파일이기 때문에 사람이 보기에 편한 엑셀 파일 등으로 변환해야 한다. 또, 표준이라고는 하지만 여전히 이용자 인증 방식, WSDL 지원 여부, 오류 처리 방식이 출판사별로 다르다는 문제를 안고 있다. 예를 들어, ProQuest는 웹 서비스 보안을 적용하여 ID/PW 및 SUSHI ReferenceID, CustomerID의 4가지 값을 입력해야 하고, OSA는 IP 인증 방식을 취하고 있다. 대부분의 SUSHI 채택 출판사에서는 WSDL을 지원하나 Jstor, ACS, Oxford University Press (이하 OUP) 등은 WSDL을 지원하지 않아 프로그램 유지가 어렵다.[6)]

무엇보다 Elsevier, Wiley 등 대형 출판사를 포함한 대부분의 출판사에서 SUSHI를 적용한

이용통계 서비스를 제공하지 않고 있고, SUSHI를 지원한다는 출판사 중에서도 AAAS, Brill, Emerald 등과 같이 SUSHI를 통한 이용통계 서비스가 실제로는 제공되지 않는 경우도 있다.

따라서 본 연구에서는 스크린 스크래핑 방식과 SUSHI 프로토콜을 적용한 방식을 복합적으로 활용하여 각 출판사의 전자저널 이용통계를 자동으로 수집하고 있다.

4.3 이용통계 수집 현황

자동수집을 통해 구축된 359개 도서관의 이용통계 데이터의 규모는 〈표 1〉과 같으며, 2012년 2월을 기준으로 총 7,539,083건이 구축되었

〈표 1〉 구축된 전자저널 이용통계 데이터 규모

정보공급사	이용 기관수	제공 연도												
		2012	2011	2010	2009	2008	2007	2006	2005	2004	2003	2002	2001	2000
AAAS	95	9	95	95	95	95	95	95	95	95	95	95	95	95
ACS	105	5,200	5,200	5,040	5,040	5,040	0	0	0	0	0	0	0	0
AIP&APS	95	1,863	2,423	2,101	2,173	1,910	1,970	2,398	387	0	0	0	0	0
APS	88	684	692	700	677	677	676	0	0	0	0	0	0	0
AR[7]	28	41	1,148	1,148	1,148	1,148	1,148	1,148	1,148	1,148	1,148	0	0	0
ASCE	45	402	2,686	2,867	2,509	1,254	1,246	1,145	0	0	0	0	0	0
ASME	22	529	597	650	432	352	350	240	220	0	0	0	0	0
Berkely	10	456	570	570	570	570	570	570	570	570	570	570	570	570
BioOne	9	1,512	1,512	1,509	1,508	0	0	0	0	0	0	0	0	0
BMJ	32	855	1,035	1,059	1,059	1,021	1,037	1,039	1,059	1,059	1,059	0	0	0
Brill	14	0	1,514	1,416	0	0	0	0	0	0	0	0	0	0
CUP	50	17,618	17,980	19,234	13,426	12,381	12,732	9,163	13,061	13,410	17,752	0	0	0
Elsevier	250	0	562,889	589,297	583,500	585,778	578,138	0	0	0	0	0	0	0
IEEE	75	0	32,018	31,125	26,936	22,866	18,358	0	0	0	0	0	0	0
IOP	73	6,649	7,957	7,957	7,661	0	0	0	0	0	0	0	0	0
Wiley	185	395,556	395,370	383,420	365,190	312,582	91,808	83,520	71,544	57,715	46,864	0	0	0
JSTOR	73	0	246	25,728	26,349	23,478	22,858	24,858	24,583	25,144	0	0	0	0
Karger	23	950	2,172	2,162	1,034	0	0	0	0	0	0	0	0	0
Mary Ann Liebert,Inc	13	840	1,086	1,072	1,072	1,072	1,072	0	0	0	0	0	0	0
NAS	34	34	34	34	34	34	34	34	34	34	34	34	2	2
NPG	110	0	1,452	1,401	1,795	1,884	1,621	718	0	0	0	0	0	0
OSA	24	483	483	437	0	0	0	0	0	0	0	0	0	0
OUP	61	0	17,141	16,607	16,637	5,303	4,878	4,711	3,365	2,680	0	0	0	0
Pion	8	0	34	34	0	0	0	0	0	0	0	0	0	0
Project MUSE	1	502	496	490	490	0	0	0	0	0	0	0	0	0
RSC	68	0	2,730	2,713	2,211	1,860	1,485	1,296	0	0	0	0	0	0
SAGE	11	0	5,169	4,926	4,885	4,922	0	0	0	0	0	0	0	0
SPIE	11	8	115	182	77	36	34	20	15	0	0	0	0	0
Springer	142	0	373,561	383,620	334,772	316,206	316,206	0	0	0	0	0	0	0
Thieme	13	0	313	301	0	0	0	0	0	0	0	0	0	0
Walter de Gruyter	6	0	721	846	846	846	0	0	0	0	0	0	0	0

6) WSDL이 지원되지 않을 경우 통신에 이용되는 메시지 전체 Packet을 XML로 작성하여야 한다. 그리고, 출판사 SUSHI 서버에서 전송 관련 세부 정보가 변경될 때마다 SUSHI client도 수정해야 하므로 안정적으로 유지하기 힘들다. WSDL이 지원되면 서비스 업데이트 기능만으로 변경 정보를 반영할 수 있어 프로그램 수정이 필요 없다.

7) 〈표 1〉의 출판사 약어는 〈부록〉에 기술하였다. 본 표는 정영임과 김정환(2012)에 제시된 〈표 6〉을 확장 보완하였다.

다. 이용통계 데이터 한 건은 한 개 참여기관에서 한 개 저널의 이용 수치가 한 건이고, 각 출판사에서 제공되는 데이터를 기초로 한다. 〈표 1〉에서 수집된 이용통계 건수가 '0'으로 표시된 것 중 2012년도 데이터는 출판사에서 2012년 1월의 이용통계 데이터를 산출하여 이용통계 보고서를 생성하는 중이라 제공이 되지 않았고(2012년 2월 기준), 과거 데이터에 '0'으로 표시된 경우는 해당 연도의 이용통계를 제공하지 않은 것이다. Brill Academic Publishers(이하 Brill), Optical Society of America(이하 OSA), Pion Ltd, Thieme Publishing Group(이하 Thieme)을 제외한 27개 출판사들이 ICOLC의 2006년 이용통계 가이드라인에 따라 과거 이용통계 데이터를 3년 이상 보유하고 있으며, 이들 과거 데이터를 제공하고 있다.

5. 이용통계 분석 개선 방안

현재 각 출판사에서 제공되는 이용통계 보고서는 〈그림 3〉과 같다.

〈그림 3〉에서 보는 바와 같이 출판사에서 제공하는 저널 이용통계 보고서는 저널에 대한 월별 이용통계만 포함되어 있다. Elsevier의 Science Direct 패키지에 포함된 저널이 2011년도 기준 3,529종이고 Springer는 2,813종, Jstor는 1,906종으로 한 도서관에서 이상의 3개 패키지를 구독한다면 사서가 분석해야하는 총 저널 수는 8,248종이 된다. 그런데 단순히 출판사명, 저널명, ISSN만으로는 해당 저널이 어떤 저널인지 파악하기 힘들어 다시 저널의 서지정보를 검색해야 한다. 또 주제 분류, Impact Factor 등에 대한 정보도 결여되어 있기 때문에 출판사에서 제공하는 원자료만으로는 저널의 이용에 대한 심층 분석을 할 수 없다(정영임, 김정환 2012).

그리고 출판사의 저널 이용통계 보고서는 한 개 기관이 구독하는 저널의 1년간 월별 통계만을 포함하고 있기 때문에 특정 컨소시엄 패키지 혹은 저널의 연도별 이용량을 살펴보기 위해서는 연도별로 생성된 이용통계 보고서를 일일이 확인해야 하고, 다년간 이용량을 구하고자 할 때는 매년 총 이용량을 합산해야 한다.

따라서 본 연구에서는 수집된 이용통계 데이터를 저널의 서지정보 및 컨소시엄 계약정보 데이터베이스와 연동하여 다양한 측면에서 이용통계 분석 정보를 생성할 수 있는 방안을 제안한다. 아울러 표, 그래프, 차트 등 시각화된 요약자료(summary)로 통계 분석 정보를 산출하여 각 도서관의 담당자가 자관에서 구독 중인 컨소

Total Full	Total PDF	Total HTML	1월	2월	3월	4월	5월	6월	7월	8월	9월	10월	11월	12월
78,840	48,609	30,231	6,765	6,464	6,702	4,979	5,409	6,155	7,138	7,557	12,322	5,803	5,237	4,309

NO	정보 공급사	Title	PISSN	OISSN	Total Full	Total PDF	Total HTML	1월	2월	3월	4월	5월	6월	7월	8월	9월	10월	11월	12월
1	Elsevier	Parasitology International	1383-5769	NA	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2	Elsevier	Information & Management	0378-7206	NA	54	27	27	15	1	9	0	2	6	0	0	11	10	0	0
3	Elsevier	Applied Soft Computing	1568-4946	NA	35	23	12	2	1	3	5	4	3	5	0	0	3	7	2
4	Elsevier	Discrete Applied Mathematics	0166-218X	NA	13	4	9	0	0	0	0	0	0	0	0	13	0	0	0
5	Elsevier	The Journal of Strategic Information Systems	0963-8687	NA	8	5	3	0	0	6	0	1	0	0	1	0	0	0	0
6	Elsevier	Journal of Hepatology	0168-8278	NA	56	38	18	6	4	2	2	10	2	20	0	4	5	0	1
7	Elsevier	Nuclear Physics A	0375-9474	NA	17	17	0	0	2	0	1	0	2	0	0	11	1	0	0
8	Elsevier	Science & Sports	0765-1597	NA	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
9	Elsevier	Annals of Tourism Research	0160-7383	NA	21	13	8	6	2	1	0	0	2	1	0	0	4	0	5
10	Elsevier	Children and Youth Services Review	0190-7409	NA	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
11	Elsevier	Food Quality and Preference	0950-3293	NA	50	26	24	0	0	2	0	2	7	22	6	2	1	3	5

〈그림 3〉 출판사 제공 이용통계 보고서(COUNTER JR1 포맷)

시엄의 전자정보 활용도가 어떤지 한눈에 직관적으로 파악할 수 있는 방안을 제안한다.

5.1 데이터베이스 연동을 통한 심층 이용통계 분석

5.1.1 저널 단위 이용통계 분석

본 시스템에서 31개 출판사에서 자동수집한 저널의 연간 이용통계 보고서를 데이터베이스로 구축하였기 때문에 SQL 쿼리문을 활용하여 다음과 같은 다양한 통계 분석 자료를 생성할 수 있다.

- 연도별 저널 이용추이: 출판사에서 매년 분리된 파일로 제공하는 저널의 이용량을 연도별로 집계하여 저널의 이용추이를 분석할 수 있다. 또 각 기관의 해당 저널 이용추이 외에도 컨소시엄에 참가하는 전체 기관의 해당 저널의 연도별 이용량을 산출하여 컨소시엄 내 평균 이용추이 대비 한 기관의 저널 이용추이를 비교할 수 있다.
- 저널 이용량 합산: 저널의 구독 시작년도부터 현재까지의 이용량을 합산한 값을 산출할 수 있다. 또 여러 출판사의 컨소시엄 패키지에 포함된 한 개 저널의 이용량을 합산하여 산출할 수 있다.
- 기관 내/컨소시엄 내 저널 이용 순위: 한 기관에서 구독하는 모든 저널을 이용량을 기준으로 순위를 낼 수 있고, 또 컨소시엄 내 전체 기관에서 구독하는 모든 저널의 이용 순위를 매길 수 있다.

본 이용통계 보고서 데이터베이스와 KISTI NDSL 시스템에 구축되어 있는 해외 전자저널 서지정보 데이터베이스를 연동하여 심층 분석자료를 생성할 수 있는데, 이를 위해서는 출판사명, 저널명, P-ISSN, O-ISSN을 매칭 키로 활용한다.

- 특정 그룹 저널의 이용량: NDSL 시스템에 구축된 저널의 DC 정보를 연계하여 주제별 저널 그룹의 이용량을 산출하여 비교하거나, SCI 등재 저널 그룹 대비 비등재 저널 그룹의 이용량 비교 등의 분석을 할 수 있다.
- 특정 그룹 내 저널 이용순위: 주제별 저널 그룹 내 저널 이용순위, SCI 등재/비등재 저널의 이용순위를 산출할 수 있으며, 저널의 Impact Factor별 순위와 이용순위를 비교분석할 수 있다.[8)]

5.1.2 컨소시엄 단위 이용통계 분석

해외 출판사의 컨소시엄 패키지는 국가별로 다르게 구성되기 때문에 출판사에서 제공하는 이용통계 보고서에는 컨소시엄에 대한 정보가 없다. 한 개 출판사에서 2개 이상의 컨소시엄 패키지를 제공하는 경우 각 패키지별 이용량을 알 수 없다. 따라서 본 이용통계 보고서 데이터베이스와 KESLI 컨소시엄의 구독 계약 데이터베이스를 연동하여 아래와 같이 컨소시엄 단위 이용통계 분석이 가능하다.

- 연도별 컨소시엄 패키지 이용추이: 컨소시엄 구독 계약 데이터베이스에서 제공하는 컨소시엄 패키지별 저널리스트를 이용하여 한 개 패키지에 포함된 저널의 이용

8) McDonald(2006)의 연구에 따르면 저널의 이용량은 해당 저널의 인용지수의 선행 지수로 적용할 수 있는 사례가 제시되었는데, 더욱 포괄적인 연구를 통해 두 지수 간의 선행 관계를 연구해 볼 수 있을 것이다.

량을 합산하고 연도별로 집계하여 컨소시엄 패키지의 이용추이를 분석할 수 있다. 또 각 기관의 해당 컨소시엄 패키지 이용추이 외에도 컨소시엄에 참가하는 전체 기관의 연도별 이용량을 산출하여 컨소시엄 내 평균 이용추이 대비 한 기관의 컨소시엄 패키지 이용추이를 비교할 수 있다.

- 컨소시엄 패키지 이용량 합산: 패키지의 구독 시작년도부터 현재까지의 이용량을 합산한 값을 산출할 수 있다.
- 컨소시엄 패키지 이용순위: 한 기관에서 구독하는 모든 컨소시엄 패키지를 이용량을 기준으로 순위를 낼 수 있고, 또 컨소시엄 내 전체 컨소시엄 패키지의 이용 순위를 매길 수 있다.

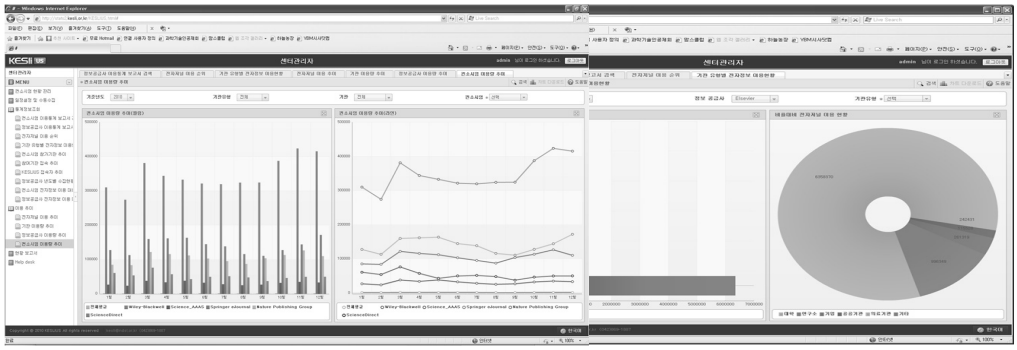
인쇄저널과 달리 전자저널은 개별 구독이 아니라 컨소시엄 패키지로 유통이 되고 있다. 컨소시엄 패키지의 구독비용 정보, 저널의 논문 수를 활용하여 컨소시엄 패키지의 이용 대비 비용 및 이용률에 대한 데이터를 산출할 수 있다. Hahn과 Faulkner(2002)의 전자저널 평가 방법을 다음과 같이 변형하여 해당 지수를 산출한다.

- 접근당 평균 비용(average cost per access) = 구독비용/원문이용 건수
- 논문당 평균 비용(average cost per article) = 구독비용/온라인 논문 수
- 수록 논문 수를 감안한 이용률(content adjusted usage) = 원문 이용 건수/온라인 논문 수

5.2 시각화 자료를 통한 직관적 분석

5.1절에서 제시한 다양한 분석 자료를 각종 수치의 나열로 이루어진 표로만 보게 되면 오히려 직관적으로 이해하기가 어려워 제대로 된 분석을 하기 힘들다.

따라서 본 시스템에서는 저널/컨소시엄 패키지의 연도별 이용추이, 저널/컨소시엄 패키지의 전체 평균 대비 기관 이용량 등의 통계 요약 자료를 〈그림 4〉와 같이 표, 그래프, 차트 등의 시각화된 자료로 표현하였다. 또 연도별 이용추이에서는 1개 저널/컨소시엄 패키지만 표현하는 것이 아니라 최대 10개 그래프를 겹쳐 표현함으로써 다수의 저널/컨소시엄 패키지의 이용추이를 한눈에 비교할 수 있도록 하였다.



〈그림 4〉 시각화된 이용통계 요약자료

6. 결론 및 제언

출판사, 도서관, 컨소시엄 주관기관 외에도 전자정보 이용통계를 필요로 하는 주체는 생각보다 다양해서(심원식, 2005), Project COUNTER, SUSHI와 같은 이용통계 관련 표준 프로젝트 담당자와 Usage Factor Project 관련자도 이용통계에 많은 관심을 가지고 있다. 드믈기는 하지만 일부 학술지는 저자들에게 저자 본인의 논문이 얼마나 사용되었는지에 대한 이용통계를 제공하는 등 학자를 포함한 다양한 그룹에서 이용통계를 활용하고 있다(Morrison, 2005).

이에 본 연구에서는 기존에 수작업을 통해 이루어지던 전자저널 이용통계 데이터의 수집 효율성을 높이고, 다양하고 심층적으로 이용통계를 분석할 수 있도록 국내 과학기술의학 분야의 컨소시엄 내에서 유통되는 전자저널의 이용통계 자동수집 및 분석자료 생성 시스템을 구현하였다. 본 시스템을 통해 도서관 담당자와 컨소시엄 주관기관의 운영자는 주요 출판사의 대용량 이용통계를 한 개 사이트에서 편리하게 표준 포맷으로 검색할 수 있으며, 다양한 통계 분석 정보를 시각화된 자료를 통해 직관적으로 확인할 수 있다. 다수 출판사의 이용통계를 통합된 시스템에서 조회/저장함으로써 도서관 사서들은 전자저널 이용통계 기초 데이터 수집부터 시간과 비용 절감이 가능해 업무 효율성을 높일 수 있다. 또한 대학 당국 및 경영진에 제출하는 보고서에 본 시스템에서 제공하는 고급 분석 정보를 포함함으로써 보고서 내용의 객관성 및 가시성을 강화시켜 사서가 경영진에 이해시키고 설득하고자 하는 내용을 효율적으로 전달할 수 있다. 컨소시엄 주관기관의 운영자는 대형 출판사를 포함한 31개 출판사에서 제공하는 전자저널에 대한 대다수 참가기관의 다년간에 걸친 이용을 종합적이고 포괄적으로 파악할 수 있다. 이를 통해 이용량이 적은 컨소시엄 패키지를 재편성할 수도 있고, 컨소시엄 패키지별 이용 대비 비용을 비교 분석하여 새로운 전자저널 가격 모형을 제안할 수도 있다. 또, 본 시스템에서 제공되는 통계 분석 자료를 컨소시엄 패키지의 구독 및 취소 결정 등 다양한 의사 결정의 보조자료 및 저널의 가치 평가를 위한 근거자료로 활용할 수 있을 것이다.

또한 도서관 담당자와 컨소시엄 주관자들의 이용통계 자동수집 및 분석 방안에 대한 관심이 높아지고 있어[9] 본 연구에서 상세하게 기술한 자동수집 방식과 분석자료 생성 방안을 참고할 수 있을 것이다.

아울러 본 시스템을 통해 생성된 다양한 통계 정보를 활용하여 연구자들이 국내 도서관들의 전자저널 구독에 대해 이용 대비 비용 측면에서의 평가 연구를 진행할 수도 있고 각주 9에서 기술한 것처럼 계량 정보학 분야의 연구자들이 이용지수와 인용지수 간의 선행 관계를 연구해 볼 수 있을 것이다.

한편 4.2절에서 살펴본 것처럼, 스크린 스크래핑 기술이 가지는 한계 때문에, 출판사 통계

9) 현재 영국의 JISC에서는 100여 개 도서관의 13개 출판사와 3개 에그리게이터의 이용통계를 SUSHI 프로토콜을 통해 수집하고 있고(MacIntyre, 2011) 독일의 Max Planck 디지털 도서관에서는 수동으로 수집한 이용통계를 SPSS와 같은 소용량 통계처리 도구를 통해 일부 분석하고 있다. Jung, Kim, You(2011)가 ICOLC 13th Europe Meeting에서 본 시스템에 대한 발표 후 Elsevier, Wiley 등의 출판사 이용통계를 포함한 보다 포괄적인 이용통계 분석 방안에 대한 질의가 많았으며, 시스템 구현 방법에 대한 관심도 높았다.

웹 사이트 조회 계정정보가 수정된다거나 출판사 통계 시스템에 수정이 있을 때마다 스크래핑 엔진도 수정을 해야 한다. 이에 본 연구진은 SUSHI 프로토콜을 적용한 이용통계 수집 모듈을 추가적으로 개발하여 두 가지 수집방식을 복합적으로 적용하고 있으나 현재로서는 SUSHI 프로토콜을 통해 이용통계 보고서를 안정적으로 제공하는 출판사가 10여 개에 불과하다. 특히 전 세계 주요 학술저널의 출판을 담당하고 있는 Elsevier, Wiley와 같은 대형 출판사가 SUSHI 기반 이용통계 서비스를 지원하지 않아 도서관계와 컨소시엄 그리고 관련 연구자들이 출판사에 표준 적용 이용통계 서비스를 제공할 것을 촉구해야 할 것이다.

참 고 문 헌

김성진, 정은경, 한민혜 (2008). 전자저널 컨소시엄을 둘러싼 학술커뮤니케이션의 쟁점과 대응동향. 정보관리연구, 39(1), 27-52.

김정환, 이응봉 (2009). KESLI 컨소시엄의 주요 이슈 분석에 관한 연구. 정보관리연구, 40(3), 99-123.

심원식 (2005). 전자정보 이용통계 활용 전략. 정보관리학회지, 22(2), 5-21.

정영임, 김정환 (2012). 전자저널 이용통계 자동수집 시스템을 이용한 컨소시엄 서비스 강화. 디지틀도서관, 65, 38-55.

정영임, 김정환, 류범종 (2011.4). 스크린 스크래핑 기반 전자저널 이용 통계 자동 수집 시스템 개발. 한국정보과학회 2011 컴퓨터종합학술대회 발표자료, 경주.

황옥경 (2007). 대학도서관에서의 전자저널 이용 통계 제공 및 활용 현황. 정보관리연구, 38(4), 68-87.

Association of Research Libraries. Annual ARL supplementary statistics. Washington. DC: Association of Research Libraries.

COUNTER Code of Practice (2008). Release 3 of the COUNTER code of practice for journals and databases. Retrieved from http://www.projectcounter.org/r3/r3_intro.pdf

Hahn, K. L., & Faulkner, L. A. (2002). Evaluative usage-based metrics for the selection of e-journals. College & Research Libraries, 63(3), 215-227.

ICOLC (2006). Revised guidelines for statistical measures of usage of web-based information resources. Retrieved from http://www.library.yale.edu/consortia/webstats06.htm

Jung, Youngim, Kim, Jeonghwan, & You, Beom-Jong (2011.9). KESLI automatic collecting system of e-journal usage statistics. Paper presented at ICOLC 13th Europe Meeting, Istanbul, Turkey.

MacIntyre, R. (2011.9). The journal usage statistics portal (JUSP). Paper presented at ICOLC

13th Europe Meeting, Istanbul, Turkey.

McDonald, J. D. (2006). Understanding online journal usage: A statistical analysis of citation and use. Journal of the American Society for Information Science and Technology, 57(13), 39-50. doi: 10.1002/asi.20420

Morrison, H. (2005). The implications of usage statistics as an economic factor in scholarly communications. Retrieved from http://hdl.handle.net/10760/6816

NISO (2007). NISO standardized usage statistics harvesting initiative (SUSHI): Z39.93. Retrieved from http://www.niso.org/apps/group_public/download.php

• 국문 참고문헌에 대한 영문 표기
(English translation of references written in Korean)

Hwang, Ok-Kyung (2007). The current status of the electronic journal usage statistics at the academic library. Journal of Information Management, 38(4), 68-87.

Jung, Youngim, & Kim, Jeonghwan (2012). Enhancement of Consortia Service by Utilizing E-Journal Usage Statistics Collection System, Digital Library, 65, 61-78.

Jung, Youngim, Kim, Jeonghwan, & You, Beom-Jong (2011.4) Development of automatic collecting system of e-journal usage statistics based on screen scrapping. Paper presented at 2011 Korea Computer Congress, Kyeongju, Republic of Korea.

Kim, Jeong-Hwan, & Lee, Eung-Bong (2009). A study on main issue analysis of the KESLI consortium. Journal of Information Management, 40(3), 99-123.

Kim, Sung-Jin, Jung, Eun-Kyung, & Han, Min-Hye (2008). Challenges and recent movements in scholarly communication concerning electronic journal licensing consortia. Journal of Information Management, 39(1), 27-52.

Shim, Won-Sik (2005). Strategies for leveraging usage statistics of electronic resources. Journal of the Korean Society for information Management, 22(2), 5-21.

〈부록〉* 31개 출판사 이용통계 서비스 제공 현황 및 본 시스템의 출판사별 자동수집 방식

출판사	Site URL	구독 기관수	저널 종수	이용통계 보고서 포맷					출판사 제공 방식	자동 수집 방식	보고서 표준 지원								분석 자료 제공	생성 주기	통계 제공일	제공 시작 년도	표준 지원 년도
				html	csv	xlm	pdf	xls			JR1	JR1a	JR1b	JR2	JR3	JR3c	JR4	JR5					
American Association for the Advancement of Science(AAAS)	http://www.sciencemag.org/cgi/instusage	95	4	○		○			웹조회 다운로드 SUSHI	Scrapping	○				○	○	○		×	월별	비정기	2000	2003
American Chemical Society(ACS)	http://pubs.acs.org/	105	39	○		○			E-mail SUSHI	SUSHI	○	○	○	○	○				×	월별	비정기	2008	2008
American Institute of Physics(AIP)	http://www.scitationreports.org	95	41	○	○	○	○	○	웹조회 다운로드 SUSHI	Scrapping	○	○			○		○		×	월별	비정기	2002	2006
American Physical Society(APS)	http://publish.aps.org	95	41	○	○	○	○	○	웹조회 다운로드 SUSHI	Scrapping	○	○			○				×	월별	비정기	2002	2007
American Society of Civil Engineers(ASCE)	http://www.scitationreports.org	45	35	○	○	○	○	○	웹조회 다운로드	Scrapping	○	○			○		○		×	월별	비정기	2006	2006
American Society of Mechanical Engineers(ASME)	http://www.scitationreports.org	22	25	○	○	○	○	○	웹조회 다운로드	Scrapping	○	○			○		○		×	월별	비정기	2005	2005
Annual Reviews (AR)	http://arjournals.annualreviews.org	28	48	○	○				E-mail SUSHI	SUSHI	○	○		○	○				×	월별	비정기	2003	2003
Berkely	http://www.bepress.com/journals	10	70		○	○		○	다운로드	Scrapping	○								×	월별	비정기	2000	2000
BioOne	http://www.bioone.org	9	162	○	○				E-mail SUSHI	SUSHI	○			○	○				×	월별	비정기	2009	2009
Brill Academic Publishers(Brill)	http://www.ingentaconnect.com/content/brill	14	109			○			E-mail SUSHI	Scrapping	○								×	월별	비정기	2010	2010
British Medical Journal(BMJ)	http://journals.bmj.com	32	26		○	○			E-mail	Scrapping	○	○			○		○		×	월별	비정기	2003	2003
Cambridge University Press(CUP)	http://journals.cambridge.org	50	242					○	다운로드 SUSHI	SUSHI	○	○							○	월별	비정기	2003	2003
Elsevier	http://usagereports.elsevier.com/asp/Main.aspx?evt=3019&src=Main.aspx.3019	250	3,529	○	○	○			웹조회 다운로드	Scrapping	○	○							×	월별	비정기	2007	2007
IEEE	https://www.mpsinsight.com/ieee	75	415	○		○			웹조회 다운로드 SUSHI	Scrapping	○			○	○		○	○	×	월별	비정기	2007	2007
Institute of Physics Publishing Ltd(IOP)	https://www.mpsinsight.com/iop	73	83	○		○			웹조회 다운로드 SUSHI	SUSHI	○				○			○	×	월별	비정기	2009	2009

*본 〈부록〉은 정영임 외(2012)의 〈표 2〉와 〈표 3〉을 통합하고 수정 및 확장하였음.

출판사	Site URL	구독 기관수	저널 종수	이용통계 보고서 포맷					출판사 제공 방식	자동 수집 방식	보고서 표준 지원							분석 자료 제공	생성 주기	통계 제공일	제공 시작 년도	표준 지원 년도
				html	csv	xlm	pdf	xls			JR1	JR1a	JR1b	JR2	JR3	JR3c	JR4	JR5				
International Society for Optics and Photonics(SPIE)	http://www.scitationreports.org/	11	7	○	○	○	○	○	웹조회 다운로드 SUSHI	Scrapping	○	○			○		○	×	월별	비정기	2005	2005
John Wiley & Sons, Inc(Wiley)	http://onlinelibrary.wiley.com/	185	1,223	○		○			다운로드	Scrapping	○	○	○						×	월별	비정기	2003	2006
Jstor	http://www.jstor.org	73	1,906	○		○			웹조회 다운로드 SUSHI	Scrapping	○								×	월별	비정기	2004	2004
Karger	http://www.karger.com/	23	84	○					다운로드 SUSHI	SUSHI	○	○							×	월별	비정기	2009	2009
Mary Ann Libert (MAL)	http://online.liebertpub.com/action/institution UsageReport	13	89	○	○				E-mail SUSHI	SUSHI	○	○		○	○				×	월별	비정기	2007	2007
National Academic of Science(NAS)	http://www.pnas.org/cgi/instusage	34	1	○					웹조회	Scrapping	○								×	월별	비정기	2000	2002
Nature Publishing Group(NPG)	https://www.mpsinsight.com/npg	112	92	○		○			웹조회 다운로드 SUSHI	Scrapping	○	○			○		○	○	×	월별	비정기	2006	2006
Optical Society of America(OSA)	http://www.opticsinfobase.org/	24	20	○		○	○	○	웹조회 다운로드 SUSHI	Scrapping	○								×	월별	비정기	2010	2010
Oxford University Press(OUP)	http://oxfordjournals.org/	61	234	○					웹조회 다운로드	Scrapping	○								×	월별	비정기	2004	2004
Pion Ltd (Pion)	http://www.pion.co.uk/lib_login.cgi	8	5			○			다운로드	Scrapping	○								×	월별	비정기	2010	2010
Project MUSE	http://stats.muse.jhu.edu/	1	196		○			○	다운로드 SUSHI	SUSHI	○				○		○		×	월별	비정기	2009	2009
Royal Society of Chemistry(RSC)	https://www.mpsinsight.com/rsc	68	39		○	○			웹조회 다운로드 SUSHI	SUSHI	○	○			○		○		×	월별	비정기	2006	2006
Sage Publications (Sage)	http://online.sagepub.com/site/subscriptions/	9	479		○	○			E-mail	Scrapping	○	○			○		○		×	월별	비정기	2008	2008
Springer	http://www.springerlink.com/	143	2,813	○	○	○			웹조회 다운로드 SUSHI	Scrapping	○							○	×	월별	비정기	2007	2007
Thieme Publishing Group(Thieme)	https://www.thieme-connect.com/ejournals	13	38		○			○	다운로드 SUSHI	SUSHI	○								×	월별	비정기	2010	2010
Walter de Gruyter	http://www.degruyter.com/	6	85	○	○				E-mail SUSHI	SUSHI	○	○		○	○				×	월별	비정기	2008	2008