J Korea Industr Inf Syst Res Volume 18 Number 2
http://dx.doi.org/10.9723/jksiis.2013.18.2.047

온라인 뉴스 웹사이트의 로그를 이용한 연관규칙 발견에 관한 연구†

(Mining Association Rules from the Web Access Log of an Online News website)

황 현 석*, 유 기 동**
(Hyunseok Hwang and Keedong Yoo)

요 약 인터넷의 활용으로 기업활동의 많은 영역이 온라인을 통해 이루어지고 있다. 온라인 쇼핑몰에서는 고객이 웹사이트 방문 후에 어떤 활동을 하는지를 파악하고 이를 경영활동의 성과로 연계하기 위해 웹 로그를 분석하고 있다. 온라인 뉴스 사이트에서도 방문자의 활동을 파악하고 어떤 기사에 관심이 많은지, 어떤 분야의 기사를 많이 보는지 등을 파악하여 독자에게 서비스하는 것이 필요하다. 그러나 언론사의 웹사이트 로그를 분석하는 연구는 충분히 이루어지지 않고 있다. 본 연구에서는 온라인 뉴스 웹사이트에서 수집된 로그를 이용하여 방문자의 웹사이트 내에서의 활동을 파악하고 뉴스 기사간 연관규칙을 도출한다. 연구는 크게 방문자의 세션(session)을 파악하는 첫 번째 단계와 방문자가 읽은 뉴스 기사간의 연관규칙을 살펴보는 두 번째 단계로 이루어져 있으며 두 차례에 걸쳐 수집된 웹사이트 로그를 이용하여 분석하였다. 최종적으로 도출된 규칙의 의미와 온라인 뉴스 사이트에서 고려해야 하는 함의를 제시하였다.

핵심주제어 : 웹 로그, 연관규칙, 데이터 마이닝, 뉴스 웹사이트, 산업정보

Abstract Today a lot of functional areas of a firm are operated on the Web. Online shopping malls analyze web log recording customers' activities on the web to connect them to business outcomes. Not only commercial websites, but online news sites also need to collect and analyze web logs to understand their news readers' interest. However, little research has been performed yet. In this research we mined the web access log of an online news website and conduct Market Basket Analysis to uncover the association rules among the categories of news articles. The research is composed of two stages: 1) Identifying the individual session of a visitor; 2) Mining association rule from news articles read by each session. We gather 7-day access logs two times. The results of log mining and meanings of association rules are suggested with managerial implications in conclusion section.

Key Words : Web log, Association rule, Data mining, News website, Industrial Information

1. 서 론

인터넷의 발달은 뉴스매체에도 큰 영향을 끼치고 있다. 지면을 이용한 신문과 전파를 이용한 텔레비전, 라디오가 뉴스 이용의 주된 매체로 사용되었으나 최근에는 인터넷이 그 역할을 대신하고 있다. 인터넷이 가지고 있는 장점 중에 하나인 이용자가 원하는 시간

† 이 논문은 2012년도 한림대학교 교비연구비(HRF-201210-009)에 의하여 연구되었음
* 한림대학교 경영학부, 한림경영연구소, 제1저자
** 단국대학교 경영학부, (kdyoo@dankook.ac.kr)

에 접속하고 지역적인 한계를 넘어 전 세계의 정보를 열람할 수 있다는 점은 뉴스 이용자의 이용패턴을 크게 바꾸고 있다. 단순 정보 조회를 넘어서 본인의 의견을 개진하고 개진된 의견을 신속히 수합하고 공유한다는 점은 일방적 뉴스 전달의 시대에서 벗어나 참여하는 뉴스의 시대가 오고 있음을 의미한다고 하겠다.

인터넷의 활용도가 증가됨에 따라 많은 산업에서 소비자에 대한 이해를 넓이기 위해 고객의 인터넷 이용을 분석하고 활용하려는 시도가 이루어지고 있다. 인터넷 쇼핑몰의 경우 고객의 인터넷상의 활동을 기록하고 있는 웹 로그를 이용하여 고객의 방문 및 활동기록을 저장하고 해석하여 이익을 극대화 하려는 노력을 보이고 있다.. 웹 로그에는 웹 이용자의 접속시간, 브라우저, 접속시간 등의 기록이 남게 되며 기업은 이 기록을 분석하여 한 명의 방문자가 방문한 웹 페이지의 일련순서 파악, 방문 페이지들 간의 관계에 기초한 방문자의 성향 분석, 빈번한 방문 페이지를 다음 번 방문에 추천하기 등에 활용하게 된다.

인터넷의 확산에 따라 뉴스전달 매체로 종이신문이나 TV 보다는 인터넷을 이용하는 경우가 많은데 인터넷 뉴스의 구독자를 인터넷 상 활동을 정확히 파악하는 방법은 구독자가 접근한 기록인 웹 로그를 살펴보는 것이 유일하다.

그러나 인터넷 미디어 분야에서는 아직도 인터넷 방문자의 활동을 분석하고 이를 웹사이트 운영에 반영하는 연구가 많지 않은 것이 사실이다[2]. 이에 따라 본 연구에서는 온라인 뉴스 사이트 방문자의 웹 로그를 이용하여 다음과 같은 내용을 분석하고자 한다.

- 어떤 분야의 뉴스를 많이 읽는가? (뉴스 분야별 선호도 분석)
- 어떤 언어로 작성된 기사를 많이 읽는가? (독자의 국적이나 선호 언어 파악)
- 여러 분야 기사 중 서로 관련 있거나 함께 읽어보는 뉴스 분야는 어떤 것인가? (관련 기사의 배치나 개인화된 서비스 제공)
- 분석 대상 회사가 집중해야 할 고객은 누구인가?

위의 내용에 대한 분석을 위해 본 연구에서는 온라인 뉴스 웹사이트의 로그를 이용하여 방문자의 클릭흐름(Click Stream)을 파악하고 방문자가 읽은 뉴스간의 연관규칙 (Association Rule)을 분석한다. 2장에서는 본 연구와 관련된 기존의 연구를 살펴보고 3장에서는 분석에 사용된 로그와 분석과정을 단계별로 설명한다. 4장에서는 분석결과를 제시하고 분석 결과가 가지는 함의를 5장에서 제공하고자 한다.

분석의 결과를 통해 인터넷 뉴스 구독자의 선호기사와 언어를 파악하고 서로 관련있는 뉴스 분야를 도출함으로써 특정 뉴스기사를 읽는 독자에게 관련 기사와 관련 뉴스 분야를 추천하는 개인화된 서비스를 제공할 수 있을 것으로 기대된다.

2. 배경 지식 및 관련 연구

2.1 웹 마이닝 (Web Mining)

웹 로그 (Web log) 분석은 크게 보면 Common Log Format [5], Extended Log Format [17], LogML [16] 등의 형태로 저장된 웹 로그를 분석하는 웹 마이닝 (Web mining)의 한 분야라고 볼 수 있다. 웹 마이닝은 정보를 발견하고 추출하기 위해서 데이터마이닝 기법을 이용하기 때문에 웹 마이닝을 데이터마이닝과 월드 와이드 웹의 교집합이라 할 수 있다[4]. 웹 마이닝은 아직 분명하게 정의 된 용어는 아니고 여러 가지 연구 분야에서 다양한 의미로 쓰이고 있다. Kosala 등은 웹 마이닝을 웹 문서와 서비스로부터 자동으로 정보를 발견하고 추출하기 위해 데이터마이닝 기법을 이용하는 것으로 웹 데이터로부터 미리 알려지지 않은 유용한 정보나 지식을 발견하는 과정으로 정의하였다 [10].

웹 마이닝은 그 적용대상에 따라 다음과 같은 3 분야로 세분할 수 있다.

- **웹 컨텐츠 마이닝 (Web Content Mining)**

웹 컨텐츠 마이닝 (Web Content Mining)은 웹 상에 존재하는 컨텐츠, 데이터, 문서 등으로부터 유용한 정보를 추출하는 일련의 작업을 일컫는다. 웹 컨텐츠 마이닝을 정보추출(Information Retrieval) 측면에서 본다면 어떤 정보를 필요로 하는 사용자가 자신이 원하는 정보를 쉽게 찾을 수 있게 하거나, 필요 없는 정보를 쉽게 걸러내기 위해 사용하는 마이닝 기법으로 볼 수 있고, 데이터 베이스 측면에서 본다면 일반적인 키워드 검색 외에 좀 더 정교한 질의어를 사용할 수 있도록 웹 상의 데이터를 모델링 하는데 사용되는 마

이닝 기법이라고 할 수 있다.

- **웹 구조 마이닝 (Web Structure Mining)**

웹사이트들은 서로 복잡하게 얽혀 있는데 이러한 링크 구조 내에 잠재하는 모델을 찾으려고 하는 작업이 웹 구조 마이닝 (Web Structure Mining)이다. 즉 웹 구조 마이닝은 하이퍼링크의 토플로지(topology)에 기반한 모델로서 서로 다른 웹사이트 간의 유사성이나 관계를 파악하는데 사용된다.

- **웹 사용 마이닝 (Web Usage Mining)**

웹 사용 마이닝 (Web Usage Mining)은 사용자들이 웹과 상호 작용하는 동안 축적된 정보를 바탕으로 사용자의 행동을 예측하는 기법이라고 할 수 있다. Web 웹 구조 마이닝은 크게 두 가지 방향으로 연구 되었다[4]. 그 중 첫 번째는 일반적인 액세스 패턴을 찾는 작업으로, 최종 분석결과를 활용하여 웹페이지의 디자인에 활용하는 연구이다[3][8][14][15]. 두 번째는 개별 사용자의 사용 패턴을 분석하여 차별화 된 서비스를 제공하려는 노력으로서 Mobasher 등은 개인의 사용 패턴을 학습한 후 각 개인의 선호에 따라 웹 사이트 자체가 적응하는 시스템이 그 예이다[6][7][9][11][12][13][18].

2.2 연관 분석

연관규칙 분석 (Association Rules Mining)은 흔히 장바구니 분석 (Market Basket Analysis)이라고도 불리며 고객이 특정 물건을 구매하면서 함께 구매한 제품을 파악하는 기법이다.

연관규칙 분석에서 거래가 가능한 물품의 집합을 **I** 라고 하면 $I=\{I_1, I_2, \dots, I_n\}$ 로 나타낼 수 있으며 장바구니 분석에 사용된 m개의 거래에 대해 거래 **T**는 $T=\{T_1, T_2, \dots, T_m\}$ 으로, 각 거래는 $T_j=\{P_1, P_2, \dots, P_k | P_k \in I, 1 \le k \le n\}$ 로 나타낼 수 있다. 두 개의 물품 A, B 간의 'A를 구매하면, B도 구매한다' 와 같은 규칙을 A → B와 같이 표현하며 이 규칙의 유효성을 확인하기 위해 다음의 3가지 값이 사용된다 [1].

- A의 지지도(Support)

$$Sup_A = \frac{n(A\text{포함 거래})}{n(\text{전체 거래})}$$

- A→B의 신뢰도(Confidence):

$$Conf_{A\to B} = \frac{Sup_{A\cap B}}{Sup_A}$$

- A→B의 향상도(Lift)

$$Lift_{A\to B} = \frac{Sup_{A\cap B}}{Sup_A \times Sup_B}$$

지지도는 두 물품의 빈번한 거래에 대한 제약을, 신뢰도는 A를 구매한 경우 B도 구매할 확률을, 향상도는 두 물품이 독립인 경우에 비해 얼마나 연관관계가 높은가에 대한 비율을 각각 계산한다.

3. 웹 로그 데이터 분석

3.1 분석 데이터

분석데이터의 수집은 국내의 기간통신사인 Y사를 대상으로 하였다. Y사는 국내외에서 취재된 기사를 타 언론사에 제공하는 소위 '뉴스 도매상'의 역할을 담당하고 있다. 최근에는 외국 통신사와 외국어 뉴스 이용자를 위한 다국어 뉴스 서비스를 제공하고 있다.

본 논문의 연구를 위해 수집된 분석 데이터의 일반적인 사항은 <Table 1>과 같다.

<Table 1> Data gathering information

Category	content
Time period	1st gathering: 2010.06.18~2010.06.24 (Worldcup football game period) 2nd gathering: 2012.12.01~2012.12.07 (Right before Presidential Election)
News Language	1st gathering: Korean, English 2nd gathering: 7 Languages
Log type	Access Log

데이터는 서로 다른 시기에 일주일간 수집되었으며 첫 번째 데이터 수집 후에 뉴스 사이트를 개편하여 영어로만 서비스되던 외국어 뉴스가 6개 의 외국어로 확대되었으며 전반적인 구조의 변경과 보안강화 등이 이루어졌다. 일반적으로 웹 로그 분석의 대상이 되는 로그는 Access Log, Agent Log, Referrer Log를 사용

할 수 있으나 분석 대상 웹 사이트는 Access Log만을 저장하고 있으므로 그 대상을 웹 로그 데이터 중 Access Log로 한정한다.

3.2 분석 도구

분석에는 데이터를 저장하기 위해 Microsoft SQL Server 2000 데이터베이스가 사용되었으며 데이터베이스의 데이터를 가공하고 변환하기 위한 Server Side Script 언어로 Active server page가, 변환된 데이터를 이용하여 분석에는 데이터 마이닝에 많이 사용되는 소프트웨어인 SPSS Clementine이 사용되었다.

3.3 분석 순서

세션(Session)은 익명의 사용자를 서로 구분하는 구분자(Identifier) 개념이며 session의 구분을 어떻게 하는 가에 따라 로그분석의 결과가 상당히 달라지므로 연관규칙 분석의 전처리 단계이나 매우 중요한 단계라고 볼 수 있다.

3.3.1 세션 도출 및 기사 범주화 단계

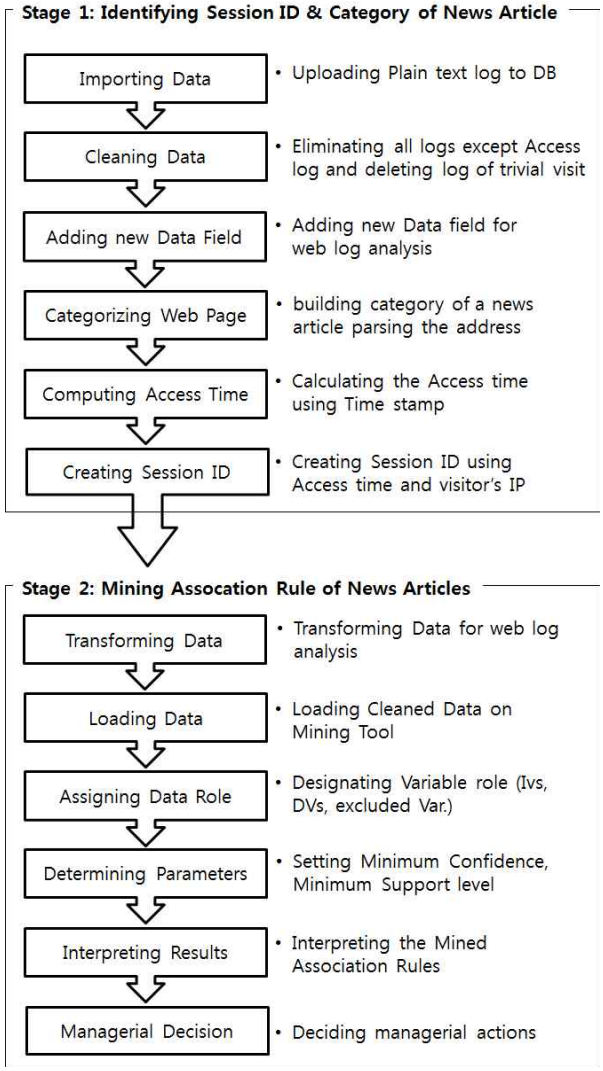
가. 원본 데이터 입수 및 데이터 import

원본 데이터는 단순 텍스트 파일(Plain text file)의 형태를 가지고 있으며 3GByte 용량의 일 단위 로그를 100Mbyte 정도의 용량으로 분할 하여 입수하였다. 데이터베이스의 ETL(Extract, Transform, Load) 기능을 이용하여 원본 데이터를 데이터베이스로 import 하였다. 로그를 구성하는 원본 데이터는 공백을 구분자로 사용하고 있으므로 공백으로 구분되는 내용을 하나의 열로 정하였다.

나. 데이터 정제

데이터 정제는 불필요한 열과 행의 삭제로 이루어진다. Access Log에는 방문자가 요청한 웹페이지 뿐만 아니라 그림파일, 동영상 파일 등 웹페이지를 구성하는 파일들에 대한 요청도 포함한다. DB에 입력된 데이터 가운데 분석에 사용되지 않는 행을 제거하기 위해 jpg, gif, swf, css, js, class 등에 대한 접근 기록은 제거되었다.

또한 분석의 대상을 단위 기사로 한정하였기 때문에 홈페이지 방문시 무조건적으로 방문하게 되는 기본 홈페이지와 인덱스 페이지에 대한 접근기록과 단위 기사를 모아놓은 페이지에 대한 접근기록도 분석에서 제외되었다. 마지막으로 광고용 페이지나 날씨 등 사용자의 방문의도와 무관하게 추가된 페이지도 제거하였다.



<Figure 1> Procedure of Analyzing Web log

다. 추가변수 생성

분석을 위해 필요한 새로운 열을 정의하였다. session_no, access_time, 기사 범주 (대분류, 중분류)와 row index에 관한 변수가 필요하여 새로운 열을 정의하였다.

라. 웹페이지 범주화

로그의 방문자 기록 중 주소을 Parsing하여 뉴스 기

사가 속한 범주를 파악함.

마. 접속시간 계산

원본 데이터에는 접속시간이 'dd/mon/yyyy:hh: mm:ss +mm:00' 형식으로 기록되어 있는데 이를 변형하여 초 단위로 변경하였다. 이를 위해 다음의 T-SQL 문이 사용되었다.

```
update access_log
set time_sec = cast(Col002 as int)*3600+cast(Col003 as int)*60+cast(Col004 as int)
```

바. 세션 생성

세션을 파악하는 일반적인 방법은 동일한 IP에서 20분 (1200초) 이내에 접근된 페이지는 같은 방문자에 의해 접근된 것으로 간주하는 것이며 본 연구에서도 이 방법을 사용하였다.

3.3.2 뉴스기사의 연관규칙 분석 단계

가. 데이터 변환

장바구니 분석을 위해서는 우선 데이터의 형태를 바꾸어야 한다. 장바구니 분석에 적합한 데이터 모양은 다음과 같다.

<Table 2> Typical format of basket data

Tran_ID	Milk	Bread	Beer	…	tofu
1	1	1	1	…	0
2	0	0	1	…	1
…	…	…	…	…	…
n	1	0	0		1

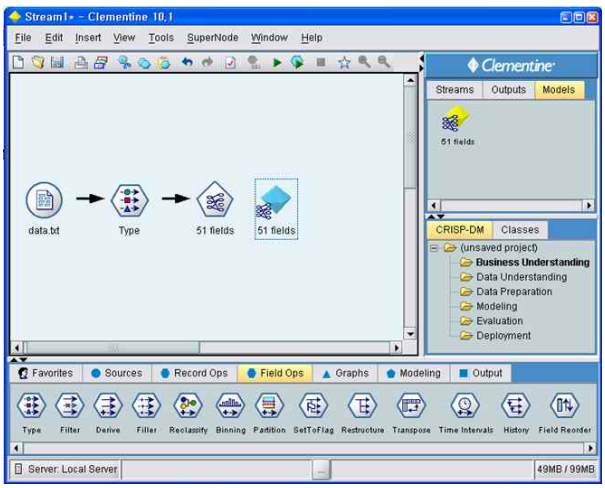
첫 번째 행에 변수가 정의되는데 첫 열에 장바구니 번호가 나오고 두 번째 열부터는 판매된 모든 물건이 나열된다. 그리고 장바구니별로 특정한 물건과 만나는 셀에는 1 (해당 물건 구매하는 경우) 또는 0 (해당 물건을 구매하지 않는 경우)이 기록된다. 데이터의 변환 시에 중요한 문제는 데이터의 상세화 정도 (Data granularity)를 정하는 것이다. 예를 들자면 한 명의 방문자에 의해 구매된 특정 브랜드의 우유를 가장 정확하게 나타내기 위해 '우유 브랜드의 이름'을 변수명으로 사용하여 상세화 정도를 높일 수 있지만 좀 더 일반적인 '우유'라는 더 상위 개념으로 표현할 수도 있고 나아가 '유제품'이나 '식품'과 같은 정도로 표현하여 상세화 정도를 낮출 수도 있다. 데이터가 상세하면 발견된 규칙을 활용하기에는 좋지만 상세화 정도를 낮추어 분석한 경우보다 연관규칙이 발견된 확률이 낮다는 문제가 있다.

본 연구에서는 개별 단위 뉴스기사를 이용하여 장바구니 분석을 할 수도 있으나 뉴스기사의 특성상 날짜가 지난 경우 더 이상 읽지 않는 경우가 많으므로 발견된 규칙을 지속적으로 적용하기 힘든 문제가 있다. 이에 따라 데이터의 상세화 정도를 개별 뉴스기사보다는 뉴스기사가 속한 사회면, 정치면, 경제면 등의 섹션을 분석의 기준으로 정하였다.

데이터의 상세화가 결정되고 상세화 정도에 맞는 형태로 데이터가 변환되었으며 이 데이터는 SPSS사의 Clementine을 사용하여 분석되었다.

나. 데이터 적재 (loading)

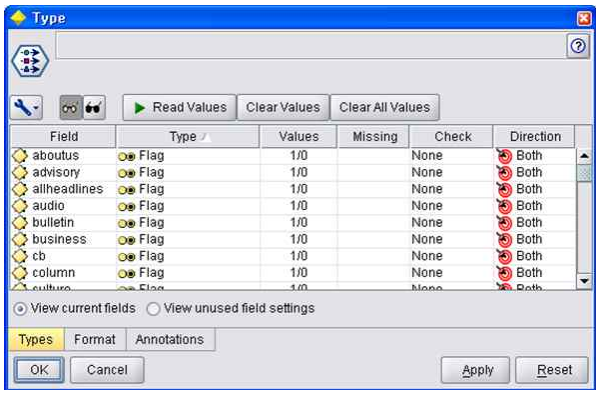
데이터 적재 (loading)는 장바구니 분석에 적합한 형태로 변환된 데이터를 읽어들이는 과정이며 데이터 로딩을 포함한 전체 장바구니 분석 분석과정은 <Figure 2>와 같다.



<Figure 2> Modeling Procedure of Analysis

다. 데이터 역할 설정

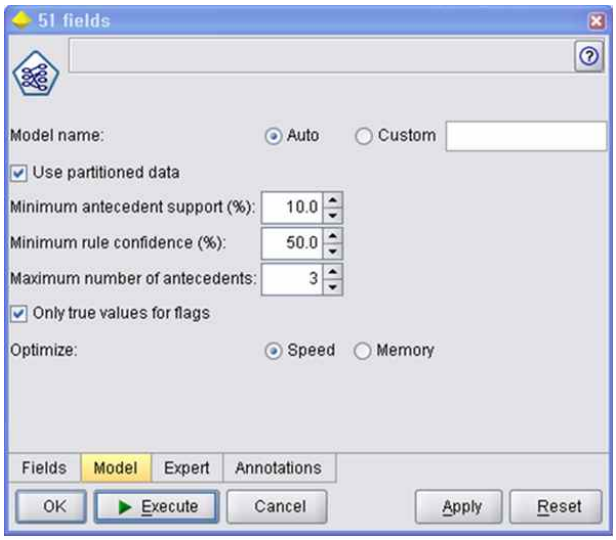
<Figure 2>의 Type 노드를 통해 데이터의 역할을 설정한다. 데이터의 역할은 불필요한 변수는 마이닝 과정에서 제외하고 종속변수와 독립변수를 지정한다. 장 바구니 분석은 일반적인 데이터 마이닝과 달리 독립변수와 종속변수의 구분이 없으므로 <Figure 3>와 같이 Direction 항목을 Both로 설정하였다.



<Figure 3> Designating the Role of Variables

라. 장바구니 분석의 파라메터 설정

Apriori 노드에서는 주요 파라메터를 지정한다. Antecedent support 는 10% 이상 (전체 읽힌 뉴스페이지 가운데 10% 이상을 차지하는 뉴스 분야)로 정하였고, Rule confidence 는 50% 이상 (A→B라는 연관성 규칙이 성립되기 위해서는 A라는 뉴스 페이지를 읽은 이용자의 50%이상이 B라는 뉴스페이지도 읽어야 함)으로 정하였다. 또한 장바구니 규칙의 Antecedents (A→B 중 A에 해당하는 부분)을 최대 3개까지 발견하도록 하였다.



<Figure 4> Configuration of Analysis

4. 분석결과

분석결과는 두 차례의 로그 데이터 분석결과 중 두 번째 로그 분석결과를 위주로 제시되었으며 첫 번째 로그 분석의 결과와 차이가 있는 경우는 추가로 그 차이를 설명하였다.

4.1 전체 통계

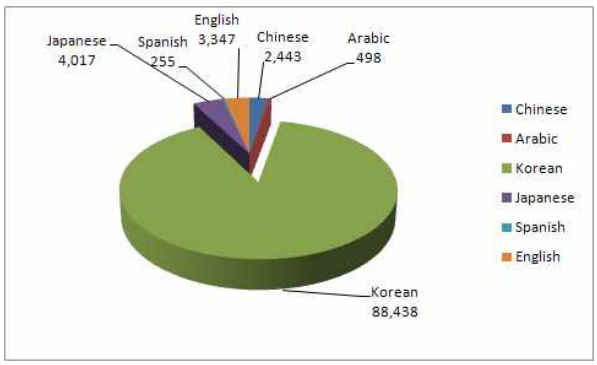
웹 로그 분석의 1단계인 Session 도출 및 기사 범주화 단계 분석을 통해 파악된 주요 통계량은 다음과 같다.

- 일일 단위 뉴스 페이지 뷰 : 98998건
- 일일 단위 뉴스 방문자 수 : 32593 명
- 방문자 당 방문 뉴스 페이지 뷰 : 3.04 페이지/명

4.1.1 뉴스언어 별 접속

1차 수집 데이터 분석에서는 외국어 뉴스 서비스는 영어로만 제공되었으며 영어 페이지 방문 비율이 0.43%에 불과하였으나 웹 사이트 개편 후 수집된 2차 데이터에서는 외국어 뉴스에 대한page view가 전체의 11.6%인로 나타났다.

이는 일반적인 뉴스 사이트에서 나타나는 외국어 page view나 방문자 비율보다 높은 것으로 분석 사이트가 국가기간 뉴스통신사로 국내외에 한국 기사를 조회하는 사람들이 자국어로 된 뉴스페이지에 접근을 많이 하는 특성을 반영한 것으로 분석된다.



<Figure 5> Percentage of News Language

뉴스 언어별 페이지 뷰와 방문자 수를 기록한 <Table 3>에서는 한국어 기사가 다른 언어로 작성된 기사보다 다소 작게 나타났다. 이는 한국 내 뉴스를

외국어로 서비스하는 사이트가 많지 않으므로 외국어 뉴스를 원하는 방문자가 방문하여 한국어 기사보다 많은 페이지를 본다는 사실을 추측할 수 있다.

<Table 3> Page view & number of visits per Language

Language	Page view	# of visitor	Page view per visitor
Chinese	2443	672	3.64
Arabic	498	82	6.07
Korean	88438	29854	2.96
Japanese	4017	788	5.10
Spanish	255	59	4.32
English	3347	1476	2.27
Total/avg.	98998	32931	3.01

이를 좀 더 세분화 하기 위해 뉴스페이지를 세부 카테고리로 나누고 카테고리별 Page view 수와 방문자 수를 살펴보면 다음과 같다. 편의상 페이지 뷰가 1000개를 넘는 카테고리만을 <Table 4>에 제시하였다.

4.1.2 방문자 페이지뷰 (Page view)

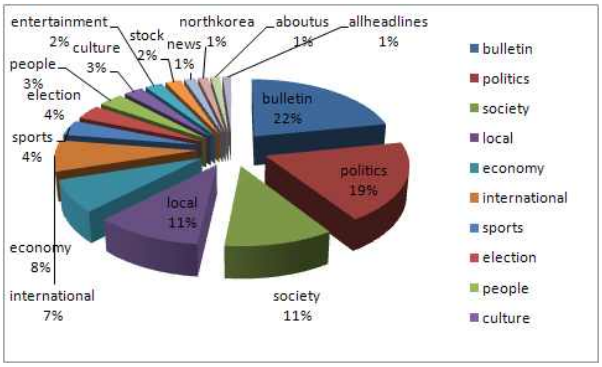
방문자당 페이지 뷰(Page view) 수는 뉴스 카테고리별로 차이가 크다는 사실을 알 수 있다. 북한 관련 기사는 방문자 1인당 평균 5.07개의 page를 본 반면 회사에 대한 소개를 다루고 있는 aboutus는 방문자 1인당 평균 1.18 개의 page를, 정치면은 1.75 개의 page를 본 것으로 나타났다. 이러한 사실은 방문자의 성향을 나타내는 것으로 방문자 가운데 북한관련 기사를 본 사람은 관련 뉴스를 많이 보는 반면에 정치면, 스포츠면, 연예면을 본 사람은 관련 기사를 상대적으로 적게 보는 것을 알 수 있다. <Figure 6>은 주요 뉴스 분야별 방문자 비율을 도식화한 것으로 어떤 분야 기사에 방문자의 방문이 많은지를 나타내고 있다. 대부분의 방문자는 속보, 정치면, 사회면, 지역 등의 순으로 나타났다. 이 결과는 독자가 어떤 분야에 관심이 많은지를 나타내고 있다.

그러나 <Figure 6>에서 비중이 높은 분야가 <Table 4>의 방문자당 Page view(Page view per visitor) 결과와는 일치되지 않고 있는데 이는 <Figure 6>은 대중성을 나타내는 결과라고 본다면 <Table 4>는 기사 몰입도의 개념으로 볼 수 있기 때문이다. 즉

<Table 4> Page view, number of visits per News Categories

News Category	Page view	# of visitor	Page view per visitor
bulletin	19676	7609	2.59
politics	17136	9770	1.75
society	9794	4684	2.09
local	9619	3171	3.03
economy	6930	3407	2.03
international	6679	2679	2.49
sports	3313	1868	1.77
election	3197	1099	2.91
people	2941	1365	2.15
culture	2283	851	2.68
entertainment	2027	1136	1.78
stock	1883	1012	1.86
news	1328	361	3.68
northkorea	1192	235	5.07
aboutus	1138	966	1.18
allheadlines	1018	240	4.24

많은 방문자가 방문하는 분야는 아니지만 한 방문자가 여러 관련 기사를 둘러볼 경우 높은 수치를 나타내기 때문이다.



<Figure 6> Percentage of visiting News Category

위의 결과에서 알 수 있듯이 사이트를 방문하는 대부분의 뉴스이용자들은 속보를 이용하거나 정치 면을 주로 보고 있음을 알 수 있다. 이것은 1차 평가시에 속보, 국제뉴스, 경제, 정치 순이었던 것에 다소 변화가 있는데 이러한 결과는 2차 평가에 활용된 로그가 저장될 당시가 대통령 선거기간이어서 대통령 관련 기사에 상대적인 관심이 증가되었기 때문이라고 풀이

된다.

최종적으로 발견된 연관규칙은 <Table 5>와 같다.

<Table 5> Results of Analysis

Antecedent	Consequent	Support(%)	Confidence(%)
International & Politics	Society	11.8	50.9
Economy & Bulletin	Politics	10.3	60.2
People	Politics	11.9	52.5
Entertainment	Politics	11.0	56.8
People	Bulletin	11.9	53.7
Local & Politics	Society	11.3	57.3
Economy & Society	Politics	10.8	66.5
Society & Bulletin	Politics	13.1	60.2
Sports	Politics	16.1	53.5
Local	Politics	20.6	54.7
International	Politics	20.1	58.8
Economy	Politics	27.2	53.7
Society	Politics	34.2	55.0
Bulletin	Politics	39.8	53.8

총 14개의 규칙이 도출되었으며 그중 첫 번째 규칙, 두 번째 규칙과 마지막 규칙이 의미하는 바는 각각 다음과 같다.

- *Rule 1 International(국제뉴스) & Politics (정치) 기사를 읽은 이용자의 50.9% 정도는 반드시 사회면을 읽는다.*
- *Rule 2 경제면과 속보를 읽은 이용자의 60% 정도는 반드시 정치면을 읽는다.*
- *Rule 14 Bulletin(속보)를 읽은 이용자의 약 53.8%는 반드시 정치면을 읽는다.*

이러한 규칙은 다음과 같이 활용될 수 있다.

- 뉴스 분야별 선호도 파악
- 방문 후 첫 구독 기사 분야 파악 (초기 화면 구성)
- 구독자 특징 파악 (외국인 구독자 비율)
- 연관있는 기사의 하이퍼 링크 추가(Navigation short-cut 제공을 통한 사용자 편의성 증대
- Y 통신사의 산업내 Market Positioning

위의 14가지 규칙에서 알 수 있는 것과 그 의미를 살펴보면 i) Consequent의 약 79%는 정치면이라는 점이다. 이러한 사실을 해석해 보면 '대부분의 이용자들은 뉴스 홈페이지에서 뉴스를 읽을 때 다양한 분야기사와 함께 정치면의 기사를 많이 읽고 간다'라는 것이다.

따라서 대부분의 구독자는 정치면에 관심이 많은 사람이라는 것을 알 수 있고 이 결과를 반영하기 위해 모든 기사를 제공할 때 정치면에 많이 읽혀진 기사의 Link를 함께 제공하여 네비게이션의 편의성을 증대시켜야 한다는 것을 알 수 있다.

그 다음 특징은 ii) Antecedent를 보면 정치면이 포함된 규칙은 2개인데 비해 경제면이나 속보가 포함된 규칙은 각각 3개로 나타나 구독자의 초기 관심은 정치가 아니라 속보나 경제뉴스임을 알 수 있었다. 이러한 결과는 Y 통신사가 긴급한 뉴스를 가장 빠르게 기사화하여 다른 뉴스 통신사에 재판매하는 도매상으로서의 특징을 반영한 것이라고 할 수 있다. 또한 2차 평가의 로그 데이터가 수집된 시기가 대통령선거 유세 시점이라는 점을 고려해 볼 때 뉴스 이용자들이 '**본인이 관심있는 분야 기사를 읽으면서 동시에 대통령과 관련된 정치분야의 기사는 거의 빠지지 않고 읽었다**' 라는 예상을 가능하게 한다.

분석 결과의 또 다른 특징으로 iii) 국제면이 많이 읽히고 있었는데 이 또한 뉴스 구독자의 특징을 잘 반영한 것으로 보인다. 국가 기간 통신사라는 특징상 국제뉴스를 가장 빠르게 전할 수 있고 국제뉴스를 접하기 위해 많은 구독자들이 Y 통신사를 이용한다는 사실이다. 이러한 분석의 결과로 다국어 뉴스 서비스의 필요성이 제기되었고 다국어 뉴스를 제공 후 전체 구독자가 구독한 페이지의 언어별 비율에서 외국어 비율이 11.6%로 나타났다. 이는 한국내 뉴스의 신속한 검색이나 국제 뉴스를 소비하는 구독자는 Y 통신사를 활발히 이용하고 있음을 의미한다.

데이터 수집기간을 고려한 특징으로 iv) 월드컵 기간에 수집된 데이터와 대선기간에 수집된 데이터에서 스포츠면의 Page view가 큰 차이를 보이지 않았는 점이다. 이러한 결과는 해당 뉴스 사이트와 방문자의 특성을 반영한 것으로 국가기간 통신사로서 스포츠보다 사회나 속보, 경제, 정치, 국제뉴스 등의 분야를 신속하게 제공하고 있으며 방문자 또한 이러한 뉴스를 구독하기 위해 찾아오고 있다는 것이다. 월드컵 관련 기사는 모든 뉴스와 신문 등의 매체에서 다루고 있어

서 다른 매체보다 차별화된 특성을 보이지 못하기 때문으로 해석된다.

또 다른 특징으로는 v) 방문자들이 경제와 사회면을 읽은 사람의 비율은 적더라도 (10.8%) 이들의 67% 정도는 동시에 정치 기사를 읽었고 반면 속보는 많은 사람들이 읽으나 (40%) 함께 정치를 읽은 사람이 53.8%로 낮게 나온 결과와는 비교되는 것이다. 이 결과를 볼 때 Y 통신사의 뉴스 구독자는 다양한 관심영역이 갖고 있으며 일부 구독자들은 다수의 구독자에서 보이는 뉴스조회 성향과는 차이가 난다는 것이다. 이는 앞서 언급한 북한관련 기사의 경우 page view수가 다른 기사보다 특별히 높은 예에서도 찾아볼 수 있다.

5. 결 론

5.1 연구요약 및 시사점

본 연구는 인터넷 언론 매체의 Access Log와 연관규칙 분석을 이용하여 방문자의 활동을 분석하였다. 결론적으로 방문자가 주로 읽는 분야가 존재하며 뉴스기사의 분야별로 서로 연관관계가 있음을 알 수 있었다.

본 연구가 갖는 학술적인 측면의 의의는 과거 연구에서 부족한 인터넷 뉴스 구독자를 분석을 시도하였다는 점이다. 기존 온라인 쇼핑몰의 고객을 대상으로 하는 웹 로그 분석은 다수 시도된 바 있으나 인터넷 뉴스 구독자를 대상으로 하는 연구가 시도되지 않았으며 또한 뉴스의 특성상 쇼핑몰의 개별 상품과는 달리 시의성이 있는 서비스라서 개별 뉴스 단위의 분석보다는 단위 뉴스가 속하는 뉴스 분야를 분석의 대상으로 하는 연구를 시도하였다는 점 또한 학술적인 측면의 의의라고 판단된다.

실무적인 측면의 의의는 뉴스를 제공하는 통신사 입장에서 구독자에 대한 선호 기사분야와 선호 언어를 파악하게 하는 방법과 과정을 제시하였으며 서로 관련된 뉴스 분야의 파악을 통해 관련 기사의 배치나 네비게이션의 short-cut을 제공하는 등의 개인화된 서비스를 제공할 수 있는 결과를 도출했다는 점이다.

흥미로운 사실은 방문자가 여러 분야의 뉴스를 보더라도 정치면을 같이 보는 경향이 많았다는 점이며 이는 연구 대상 뉴스 사이트가 가지는 특성에 따라 독자의 주요 관심 기사의의 분야가 정해져 있다고 하겠다. 또한 외국어 뉴스를 읽은 비중이 많은 것을 볼 때 국가기간 뉴스 통신사의 역할을 하고 있음을 알 수 있었다.

발견된 규칙을 이용하여 홈페이지의 재 구성시 방문자가 동시에 관심이 많을 것으로 예상되는 분야들의 주요 기사만을 요약하여 보여준다거나 현재 방문자가 관심있게 본 기사와 장바구니 분석 결과를 결합하여 다른 기사를 추천하는 등에 활용할 수 있을 것으로 기대된다.

또한 Y 통신사의 Market Positioning을 위해서는 속보와 경제면에 풍부한 기사를 제공하고 정치면을 함께 구독하는 다수의 구독자를 위한 편의성을 증대시켜야 하며 다국어 서비스를 통해 외국인도 편하게 구독할 수 있는 서비스를 제공해야 한다는 사실을 도출할 수 있었다. .

5.2 연구의 한계점 및 향후 연구방향

본 연구의 한계점은 통계적 측면의 신뢰성 및 타당성 검증이 부족하다는 점이다. 로그 분석은 그 방법론의 특징상 확률에 의한 계산이 주를 이루고 통계적인 검정이 어렵다. 또한 익명의 방문자를 포함하므로 로그인을 하지 않은 방문자의 인구통계학적인 특징을 찾아내는 것이 힘들다는 한계점이 있다. 또한 데이터가 수집되는 특정 시기에 따라 방문자의 구독 페이지에 영향을 줄 수 있었다는 점도 연구의 한계점이라고 할 수 있다.

향후 연구방향은 첫째, 세밀한 로그 분석을 위해 정확한 세션의 확보가 필수적이며 이를 위해 referrer log와 agent log 를 결합한 분석이 필요하다는 것이다. 예를 들어 뉴스 이용자가 어떤 웹사이트를 통해 자사의 뉴스 웹사이트로 들어오고 어떤 웹사이트로 이동하는지 파악하여 타 뉴스 관련 사이트와의 연관성을 파악하는 것이 필요하고 또한 agent log를 통해 어떤 웹 브라우져를 통해 뉴스를 보는지도 파악하는 것이 좋다. 이용하는 웹 브라우져의 종류 자체는 의미가 많다고 보기 힘들지만 동적으로 할당된 IP에서 접속된 뉴스 이용자를 웹 브라우져의 종류나 다른 정보를 이용하여 한번 더 filtering함으로써 session을 구별하는데 도움을 줄 수 있을 것이다. 두 번째 연구방향은 현재 분석에서는 일주일간의 로그데이터를 사용하였는데 데이터의 범위를 일주일 이상으로 하여 평균적인

구독자의 성향을 분석하는 연구가 필요하다.

참 고 문 헌

[1] Agrawal, R., and Srikant, R., 1994, Fast Algorithms for Mining Association Rules, Proceedings of the 20th VLDB Conference, Santiago, Chile, pp. 487-499.

[2] Batista, P., and Silva, M. J., 2002, Mining Web Access Logs of an On-line Newspaper, Departamento de Informática, Faculdade de Ciencias - Universidade de Lisboa, Portugal, pp. 1-8.

[3] Berendt, B., 2002, Using site semantics to analyze, visualize, and support navigation, Data Mining and Knowledge Discovery, Vol. 6, No. 1, pp. 37 - 59.

[4] Britos, P., Martinelli, D., Merlino, H., and García-Martínez, R., 2007, Web Usage Mining Using Self Organized Maps International Journal of Computer Science and Network Security, Vol. 7, No. 6, pp 45-50.

[5] Configuration file of W3C httpd, http://www.w3.org/Daemon/User/Config/ (1995).

[6] Dai, H., and Mobasher, B., 2002, Using ontologies to discover domain-level web usage profiles, Proceedings of the 2nd Semantic Web Mining Workshop at ECML/PKDD, Helsinki, Finland. pp.1-17.

[7] Fenstermacher, K., and Ginsburg, M., 2002, Mining client-side activity for personalization, Fourth IEEE International Workshop on Advanced Issues of E-Commerce and Web-Based Information Systems, pp. 205 - 212.

[8] Fu, Y., Creado, M., and Ju, C., 2001, Reorganizing web sites based on user access patterns, Proceedings of the Tenth International Conference on Information and Knowledge Management, pp. 583 - 585.

[9] Kim, H., and Chan, P., 2003, Learning implicit user interest hierarchy for context in personalization, Proceedings of the 2003 International Conference on Intelligent User Interfaces, pp. 101 - 108.

[10] Kosala, R., and Blockeel, H., 2000, Web mining research: a survey, ACM SIGKDD Explorations Newsletter, Vol. 2, No. 1, pp. 1-15.

[11] Lin, W., Alvarez, S., and Ruiz, C., 2002, Efficient adaptive - support association rule mining for recommender systems, Data Mining and Knowledge Discovery, Vol. 6, No. 1, pp. 83-105.

[12] Mobasher, B., Dai, H., and Tao, M., 2002, Discovery and evaluation of aggregate usage profiles for web personalization, Data Mining and Knowledge Discovery, Vol. 6, pp. 61 - 82.

[13] Moshaber, B., Cooley, R., and Srivastava, J., 2000, Automatic Personalization Based on Web Usage Mining, Communications of the ACM, Vol. 43, No. 8, pp 142-151.

[14] Spiliopoulou, M., and Pohle, C., 2001, Data mining for measuring and improving the success of web sites, Data Mining and Knowledge Discovery, Vol. 5, No. 1-2, pp. 85-114.

[15] Srikant, R., and Yang, Y., 2001, Mining web logs to improve website organization, World Wide Web, pp. 430 - 437.

[16] Srivastava, J., Cooley, R., Deshpande, M., and Tan, P.-N., 2000, Web usage mining: discovery and applications of usage patterns from web data, SIGKDD Explorations, Vol. 1, No. 2, pp. 12 - 23.

[17] W3C Extended Log File Format, 1996, http://www.w3.org/TR/WD-logfile.html.

[18] Xie, Y., and Phoha, V., 2001, Web user clustering from access log using belief function, Proceedings of the First International Conference on Knowledge Capture (K-CAP 2001), pp. 202 - 208.

황 현 석 (Hyunseok Hwang)

- POSTECH 산업경영공학과 공학사
- POSTECH 산업경영공학과 공학석사
- POSTECH 산업경영공학과 공학박사
- 한림대학교 경영학부 부교수
- 한림대학교 한림경영연구소 연구위원
- 관심분야 : 스마트 비즈니스, 빅데이터, 인텔리전트 시스템

유 기 동 (Keedong Yoo)

- POSTECH 산업경영공학과 공학사
- POSTECH 산업경영공학과 공학석사
- POSTECH 산업경영공학과 공학박사
- 단국대학교 경상대학 경영학부 부교수
- 관심분야 : 지식경영 및 지식관리시스템, 유비쿼터스 컴퓨팅, 차세대형 경영정보시스템, 컨텍스트 기반 자율적 컴퓨팅, 지능적 지식 서비스, 정보전략 기획 및 성과평가

논 문 접 수 일 : 2013년 03월 12일
1차수정완료일 : 2013년 04월 02일
2차수정완료일 : 2013년 04월 19일
게 재 확 정 일 : 2013년 04월 19일