The Korean Communications
in Statistics Vol. 8, No. 1, 2001
pp. 281-289

A Proposal of Some Analysis Methods for Discovery of User Information from Web Data

Jeong Yong Ahn[1] and Kyung Soo Han[2]

Abstract

The continuous growth in the use of the World Wide Web is creating the data with very large scale and different types. Analyzing such data can help to determine the life time value of users, evaluate the effectiveness of web sites, and design marketing strategies and services. In this paper, we propose some analysis methods for web data and present an example of a prototypical web data analysis.

Keywords : Web Data, Analysis Methods, Knowledge Discovery

1. 서론

World Wide Web(이하 Web)은 이제 우리 생활의 필수 요소로서 자리를 확고히 하고 있다. 정보의 공유는 물론 비즈니스(business)를 수행하기 위한 매체로서 빠른 속도로 확장되고 있으며, 인간의 많은 활동을 가상 공간으로 확장시키고 있다.

이러한 변화는 여러 학문 분야에 많은 도전과 기회를 제공하고 있으며, 특히 데이터의 분석을 통한 정보의 발견에 관심이 있는 통계학 분야에 좋은 연구 환경을 제공하고 있다. Web이 통계학 분야에 제공하는 연구 환경은 여러 가지를 들 수 있겠지만 크게 다음과 같은 두 가지 특징으로 나타낼 수 있을 것 같다.

- Web과 같은 가상 공간을 통하여 대규모의 데이터가 양산되고 있다.
- 기존에 취급하지 않았던 새로운 형태의 데이터가 다양하게 출현하고 있다.

이러한 데이터의 양적 증가와 새로운 형태의 데이터 출현은 데이터 분석 및 활용 방법에 새로

1) Assistant Professor, Division of Computer Science and Information Communications, Seonam University, Chonbuk, 590-170, Korea
E-mail : jyahn@tiger.seonam.ac.kr
2) Professor, Division of Mathematics and Statistical Informatics, Chonbuk National University, Chonbuk, 561-756, Korea
E-mail : kshan@stat.chonbuk.ac.kr

운 기법들을 요구한다. 대규모 데이터에 기존의 분석 방법을 적용하는 것은 의미가 없으며, 새로운 형태의 데이터는 그것을 처리하기 위한 전체적인 방법을 연구해야 되기 때문이다.

Huber(1994)는 대부분의 대규모 데이터는 동질성을 갖지 않는 데이터들로 구성되어 있기 때문에 기존의 통계적 분석 방법으로 취급하는 것은 의미가 없음을 지적하면서, 데이터 시각화 기법의 활용을 강조한다. Friedman(1997) 역시 데이터의 양이 증가되면 그것을 분석하는 방법을 완전히 다시 생각해야 한다고 지적하면서, 데이터에 관련된 여러 기술들과 데이터 마이닝 분야에 통계학자들의 연구와 관심의 필요성을 주장한다. Hand(1998)는 현대의 통계학과 데이터 분석은 통계적 방법들을 설명하기 위한 과거의 전통적인 방법들과는 다르게 생각해야 한다고 말한다. 그는 통계적 모델(model)은 중요한 것이 아니며, 대규모 데이터에서 검정(testing)의 결과가 가지는 '유의성(significance)'은 아무런 의미도 갖지 못함을 언급하면서 데이터 분석의 궁극적 목적은 패턴을 발견하는 것이라 주장한다.

Famili 등(1997)은 현실 세계에서 새롭게 나타나는 데이터가 갖고 있는 문제점들을 지적하고 데이터 분석을 시작하기 전에 해결해야만 하는 과제들을 제시하고 있으며, 적절한 분석 방법의 개발에 대한 필요성을 강조한다.

현대 사회의 정보 기술의 발달은 놀라운 컴퓨팅 능력을 제공함으로서 이러한 요구들을 해결하기 위한 적절한 환경을 조성해 주고 있으며, 데이터를 분석하기 위한 새로운 기법들은 데이터 마이닝(Data Mining)이라는 이름하에 점차로 그 모습을 나타내고 있다. Boosting, Bagging, Scoring 방법, 군집 방법, 연관 규칙(association rules), 순차 패턴(sequential patterns) 등 대규모의 데이터 또는 Web 데이터로부터 정보를 추출할 수 있는 방법 및 알고리즘에 관한 연구가 활발히 진행되고 있으며, 데이터 시각화(Data Visualization), Web 마케팅(Web Marketing), eCRM(electronic Customer Relationship Management) 등 데이터 활용을 위한 많은 방법 또는 기술들이 활발히 연구되어지고 있다.

Web을 통하여 얻어질 수 있는 데이터는 많은 종류(예를 들면, text, content, usage 데이터)가 있다. usage 데이터는 Web 서버에 접속한 방문객의 IP(internet protocol) 주소, 접근 시각, 접근 방법, 대상 URL, 전송 프로토콜, 에러 코드 등의 Server Log 데이터는 물론 이용자의 등록 정보, 마케팅(marketing) 관련 데이터를 포함한다. Web 데이터는 이용 관점에 따라 여러 가지로 분류할 수 있겠으나 본 연구에서 논의의 대상으로 삼는 데이터는 기업의 마케팅 관점에서 많이 이용되는 데이터로 그 범위를 한정한다. 따라서 본 연구에서 언급하는 Web 데이터는 Web 마케팅이나 eCRM 분야에서 이용되는 데이터를 말한다. 이들 데이터를 특성에 따라 분류하고, Web 데이터 마이닝을 통하여 사용자들의 정보를 획득할 수 있는 몇몇 데이터 분석 방법들을 제안하는데 본 연구의 목적이 있다.

본 연구에서 Web 마케팅 또는 eCRM 등에서 다루어지는 데이터를 고려하는 이유는 이 데이터들이 일반적인 통계 데이터의 특징과 위에서 언급한 Web 데이터의 특징(대규모 데이터, 새로운 형태의 데이터)을 포함하고 있기 때문이다.

2장에서는 Web을 통하여 얻을 수 있는 데이터를 크게 3가지 범주로 나누어 그 특징에 대해 살펴본다. 3장에서는 데이터 분석 과정에 대해 언급하고, 분석된 결과가 이용될 수 있는 수준과 데이터 종류별 수준에 따른 몇몇 분석 기법을 제안한다. 4장에서 Web 데이터 분석에 대한 간단한 예제를 제시한다.

2. Web 데이터의 종류 및 특징

인터넷 또는 Web의 기본적인 목적은 학문과 연구를 위한 정보의 공유이다. 그러나 최근의 Web 환경은 정보의 공유 이상으로 비즈니스와 마케팅 분야를 포함한 생활의 전 분야에 그 영향을 미치고 있다. 많은 예측들에 따르면, 10년 이내에 Web을 통하여 인간 활동에 대한 대부분의 정보를 이용 할 수 있을 것으로 보이며(Garofalakis 등, 1999), 이러한 인간 활동의 부산물로서 대규모의 새로운 형태의 데이터가 매일 양산되고 있다.

Web 상에서 발생하는 데이터는 연구자에 따라 여러 가지 방법으로 구분되고 있다. Cooley (2000)는 데이터가 수집되는 위치에 따라 데이터를 (a) 서버(server) 수준, (b) 클라이언트(client) 수준, (c) 프록시(proxy) 수준으로 구분하고 있다. 서버 수준의 데이터는 사이트(sites)를 방문하는 사용자의 정보를 포함하는 server log를 들 수 있다. 클라이언트 수준의 데이터는 클라이언트 브라우저에서 발생하는 사용자의 활동(이 데이터는 적당한 프로그램을 통하여 수집될 수 있다) 정보를 모아둔 것이며, 프록시 수준의 데이터는 프록시 서버에 존재하는 사용자들의 활동에 관한 정보라 할 수 있다.

Spiliopoulou(2000)는 Web에서 획득할 수 있는 데이터를 (a) 사용자가 데이터의 발생을 인식하지 못하는 데이터와 같은 요구 또는 강제성이 없는 데이터(예를 들어, server log), (b) 등록 요구, 질문 등과 같은 요구 또는 강제성이 약간 있는 데이터(예를 들어, 사용자의 신상에 관한 정보)로 구분한다.

본 연구에서는 기업의 마케팅 관점에서 Web으로부터 획득할 수 있는 데이터를 다음과 같이 3가지 범주로 구분하고, 다음 장에서 몇몇 분석 기법을 제안한다.

(i) Server Log 데이터
(ii) User 데이터
(iii) Market 데이터

(i) Server Log 데이터

Server Log 데이터는 Web 서버에 대한 모든 방문객의 접근 및 조회 내역 등에 대한 정보를 기록하고 있는 시스템으로부터 생성되는 데이터이다. 이 데이터는 IP 주소, Log 시간, 이용 브라우저(browser), 전송 프로토콜(protocol) 등의 사용자 인식정보와 개별 페이지 방문에 대한 페이지 방문정보로 구분할 수 있다. 이 데이터의 특징은 첫째, 사용자가 거의 인식하지 못하는 사이에 생성되며 둘째, 사용자들의 수에 의존하기는 하지만 일반적으로 대규모로 발생된다. 셋째, 기존에 취급하지 않았던 새로운 형태의 데이터이며 넷째, 데이터 분석에 불필요한 정보를 많이 포함하고 있기 때문에 분석 과정에서 데이터 정제(cleaning) 과정이 필수적으로 요구된다.

(ii) User 데이터

User 데이터는 Server Log 데이터로부터 추출할 수 없는 정보를 사용자로부터 수집한 데이터이다. 사용자의 성별, 나이와 같은 신상 정보뿐만 아니라 e-mail 주소나 어떤 경로를 통하여 Web 사이트에 접근하였는가 하는 등의 정보를 가질 수 있으며, 이용 목적에 따라 더 세부적인 정보를 수집할 수도 있다. 이 데이터를 통하여 사용자의 잠재적 가치를 평가할 수 있으며, 사용자로부터 직접 입력을 받기 때문에 어떤 특정한 문제를 해결하기 위한 구체적인 정보를 수집할 수 있는 특징이 있다.

(iii) Market 데이터

Market 데이터는 기업에서 취급하는 상품에 대한 정보와 전자 상거래에 대한 모든 정보를 포함한다. 이 데이터의 적절한 활용은 물품 판매 전략, 고객 관리 등의 계획 수립에 매우 중요한 비중을 차지하며, 교차 판매(cross sales) 분석, 장바구니(market basket) 분석, 고객 가치 분석 등에 이용할 수 있다.

이러한 데이터 중 Server Log 데이터의 이용에 관한 연구는 매우 많다. 예를 들면, Borges와 Levene(1999), Buchner 등(1999), Pei 등(2000), Srivastava 등(2000)을 들 수 있으며 SIGKDD, WebKDD 학술회의 등을 통해 많은 연구가 발표되고 있다.

그러나 Server Log 데이터는 User 데이터와 Market 데이터에 비해 상대적으로 정형화되어 있지 않다는 단점을 가지고 있음을 고려해야 한다(특히, 우리 나라와 같은 환경에서는 더더욱 그렇다). 부연하여 설명하면, 일반적으로 Server Log 데이터를 분석할 때 같은 IP 주소를 갖는 사용자를 한 사람으로 취급하여 사용자의 패턴을 발견하고자 한다. 그러나 같은 IP 주소를 사용한다고 해서 같은 사용자라는 보장은 없다. 따라서 IP 주소에 대한 정보가 고려되어야 함은 당연하며, 데이터 활용을 보다 효율적으로 하기 위해서는 Server Log 데이터 뿐만 아니라 위에서 언급한 데이터들을 통합하여 분석하는 것이 필수적이다.

3. Web 데이터를 분석하기 위한 방법들

Web 데이터를 분석하기 위한 대표적인 방법들로 연관 규칙, 순차 패턴, 분류, 군집 분석, 예측 등의 방법이 주로 언급되어지고 있으며, 이러한 방법들에는 일반적인 통계적 방법, 근접 이웃(nearest neighbor), 군집 분석 등의 고전적 기법과 물론 의사결정나무(decision tree), 신경망(Neural Network) 등의 기법들을 포함한다. 이 중 군집 분석은 Web 데이터 분석에 가장 많이 활용되는 분석 방법 중의 하나이다. 군집 분석 알고리즘은 Hartigan과 Wong(1979)의 k-평균 알고리즘과 신경망(Neural Network) 이론의 하나인 Kohonen(1984)의 코호넨 네트웍(Kohonen Network)이 주로 이용되고 있으며 Guha 등(1998), Zhang 등(1996)에서 대규모 데이터를 위한 연구가 진행되었다.

이러한 분석 방법들을 통하여 일반 사용자에게는 개인별 Web 페이지 서비스와 상거래에 있어서 개인별 제안 등의 서비스를 제공할 수 있으며, 기업 관리자에게는 사용자 성향 분석, 구매 패턴 분석, hit 추이 분석 등을 제공할 수 있는 기능을 수행한다.

Web 데이터의 분석은 일반적으로 다음과 같은 과정으로 이루어진다.

- Server Log 데이터, User 데이터, Market 데이터 수집
- 데이터 정제(cleaning)
- 데이터의 재구성(transaction)
- 데이터 통합
- 분석에 필요한 형태로 데이터 변환
- 분석

사용자 정보의 효과적인 추출을 위한 Web 데이터의 분석은 다양한 관점에서 논의될 수 있으며, 분석 기법의 세부적인 방법들에 대한 언급도 필요할 것이다. 예를 들어, k-평균 알고리즘, 코호넨 네트웍(Kohonen Network), BIRCH, CURE 등 대규모 데이터의 군집 분석에 이용할 수 있는 알고

리즘의 종류와 각각의 특성에 대한 검토와, Log 데이터와 같은 새로운 형태의 데이터 분석을 위한 세부 방법들 - 연관 규칙, 순차 패턴 탐색 방법 등 - 에 대한 알고리즘의 분석과 장단점의 비교도 필요할 것이다.

그러나 본 연구의 목적이 Web 데이터 마이닝에서 활용 가능한 여러 가지의 일반적인 데이터 분석 방법들을 제안하는데 있기 때문에 세부적인 방법들에 대해서는 다루지 않기로 한다. 본 연구에서는 대규모 데이터와 새로운 형태의 데이터에 적용할 수 있는 몇몇 분석 기법을 분석 결과가 이용될 수 있는 수준과 2장에서 구분한 데이터 종류별로 제안해 보고자 한다.

3.1 분석결과 이용관점에서의 분석방법

Web 데이터 분석 결과의 이용관점은 여러 가지로 구분할 수 있겠으나 본 연구에서는 Web 사이트 관리 수준, 사용자 관리 수준, 그리고 기업의 전략 수준으로 구분하여 각각의 수준에 대한 데이터 분석방법을 제안하고자 한다.

<표 1> 분석결과 이용수준별 분석방법

분석결과 이용수준	목적	분석내용 또는 방법
Web 사이트 관리 수준	- Web 사이트 관리	Traffic 정보(traffic 분석)
		Web 사이트 효율성(빈도 분석)
		Web 페이지 방문 정보(연관 규칙, 순차 패턴)
		Web 페이지 군집화(군집 분석, Visualization)
사용자 관리 수준	- personalization - 사용자 관리	사용자 정보(기초 통계 분석)
		세분화(군집 분석, 분류)
		이탈 가능성(빈도 분석, 예측)
기업의 전략 수준	- 사용자 서비스 - 이윤의 최대화	연관성 정보(상관 분석, 연관 규칙)
		구매 패턴 탐색(상관 분석, 교차 판매 분석, 장바구니 분석, Branch 분석, Visualization)

Web 사이트 관리 수준에서의 데이터 분석은 트래픽(traffic) 발생의 원인 탐색과 해결, 사이트의 효율성, Web 페이지의 군집화 등을 통해 효율적으로 사이트를 관리하고, 사용자 관리 수준에서는 사용자 개인 또는 그룹별 특성화(personalization)와 효과적인 사용자 관리에 목적이 있으며, 기업의 전략 수준에서 데이터 분석의 목적은 사용자에 대한 서비스와 이윤의 최대화에 있다(<표 1> 참조).

3.2 데이터의 종류에 따른 분석방법

<표 2>는 데이터를 종류별로 구분하여 적용할 수 있는 분석내용과 방법을 정리, 제안한 것이다. Server Log 데이터를 통하여 사용자들의 Web 방문 패턴을 발견할 수 있으며, User 데이터와 Market 데이터 등을 통하여 사용자 세분화, 구매 패턴 등을 파악하고 전체적인 서비스 전략을 수

립할 수 있다.

<표 1>과 <표 2>에서 제안된 분석 방법들의 대부분은 통계학에 관심이 있는 사람들에게는 익숙한 방법들이며, Branch 분석, RFM(Recency-Frequency-Monetary) 분석, 4R(인식, 파악, 전달, 응답) 분석, 교차 판매(cross-selling) 분석, 장바구니(market basket) 분석 등은 비교적 최근에 추가된 새로운 기법들이다. Branch 분석은 분석하고자 하는 속성을 나무의 가지 구조처럼 세부적으로 확장시켜 나가면서 적용하는 기법이다. RFM 분석은 어떤 고객에 대한 최근(recency)의 구매 횟수(frequency)와 구매 금액(monetary)의 데이터를 이용하여 고객의 가치를 점수화(scoring)하고 등급을 평가하는 기법으로, 고객 관리에 있어 가장 많이 활용되고 있는 기법으로 알려져 있다. 4R 분석은 판촉에 대한 반응 정도를 살펴보는 것을 예로 들 수 있다. Web 마케팅 또는 eCRM 등을 수행하기 위해서는 이러한 최근 기법들이 필수적으로 추가되어야 할 것으로 생각된다.

<표 2> 데이터 종류별 분석방법

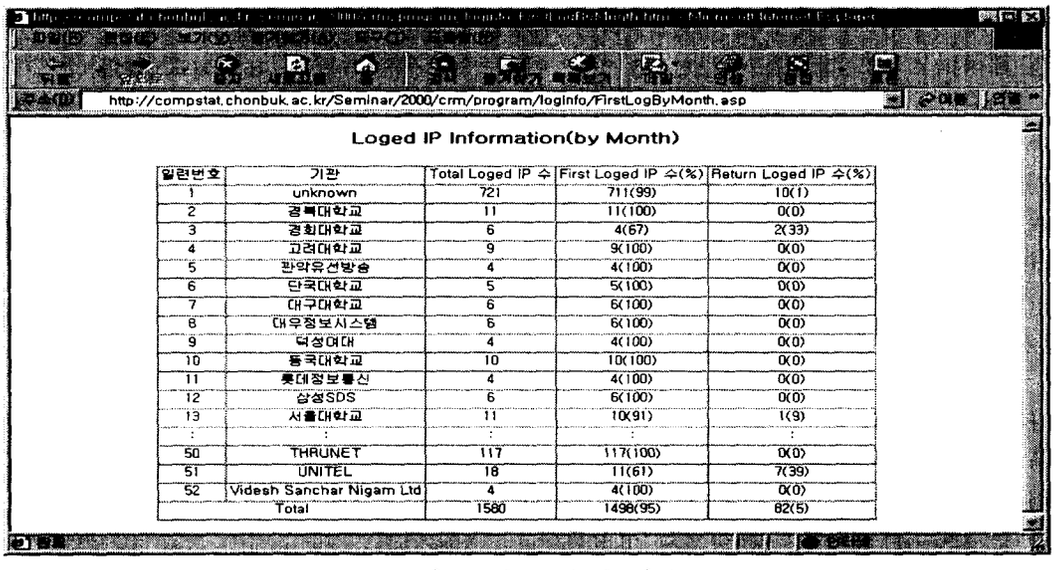
데이터	이용목적	분석내용 또는 방법
Server Log 데이터	Web 방문 패턴 발견	로그 정보(기초 통계 분석, Visualization)
		Traffic 정보(traffic 분석)
		연관성 정보(상관 분석, 연관 규칙)
		Web 페이지 방문 정보(연관 규칙, 순차 패턴)
User 데이터	사용자 세분화	사용자 정보(기초 통계 분석)
		세분화(군집 분석, 분류)
		가치 분석(빈도 분석, 4R 분석)
Market 데이터	구매 패턴 발견	물품 구매 정보(기초 통계 분석, Branch 분석)
		구매 패턴 탐색(상관 분석, 순차 패턴, 교차 판매 분석, 장바구니 분석, Branch 분석)
		고객(포인트)관리(R-F-M 분석)
통합 데이터	서비스 전략 수립	데이터 군집화(군집 분석, 분류, Visualization)
		이탈 가능성(빈도 분석, 예측)
		고객 관리(4R 분석)
		데이터들간의 연관관계(상관 분석, 연관 규칙, Visualization)

4. Web 데이터 분석 예

본 연구에서 제시하는 예제는 Web 데이터 분석에 관한 prototype으로 설계되었으며, 각종 데이터를 통합하여 분석을 실시한다. 분석 내용은 일별, 월별에 따른 로그 정보와 Web 페이지 이용빈도, 연관 규칙, 군집 분석 및 분류, 그룹별 구매 패턴 분석, 이탈 가능성 등을 포함하고 있다

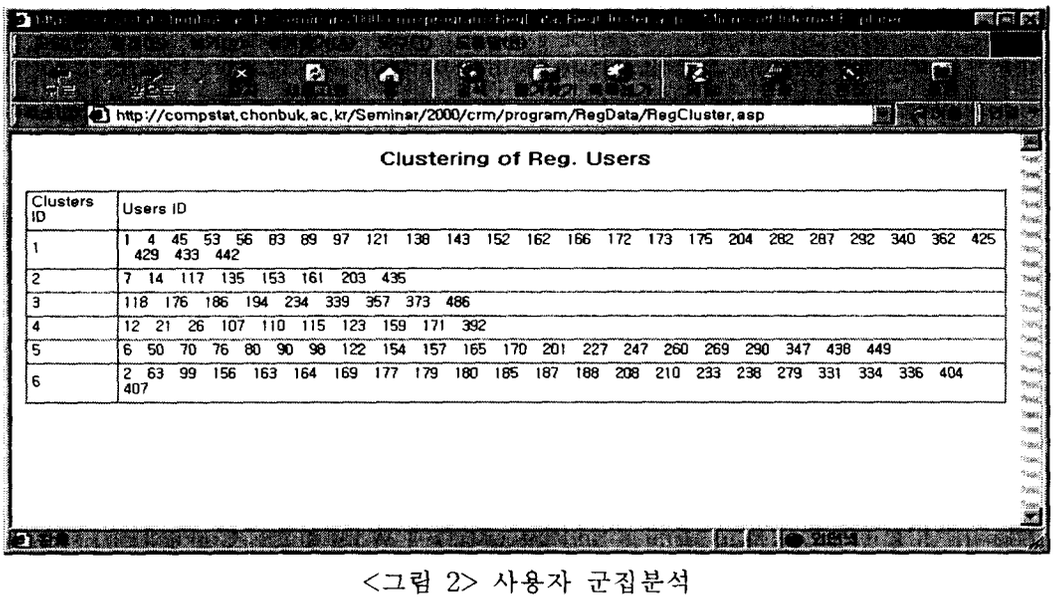
<그림 1>은 Web 사이트에 한달 동안 로그한 IP들에 대한 정보를 보여 주고 있다. 로그한 IP 1580개 중 1498개 즉 95%는 처음으로 로그한 IP임을 보여주고 있으며 <그림 2>는 사용자들의 등

록정보를 이용하여 군집 분석을 실시한 결과이다. 이러한 분석 과정을 거쳐 개인별 혹은 그룹별로 최적의 서비스를 할 수 있고, 어떤 사용자들에게 특별한 관리가 필요하다는 정보와 전략 수립의 토대를 마련할 수 있다.



일련번호	기관	Total Loged IP 수	First Loged IP 수(%)	Return Loged IP 수(%)
1	unknown	721	711(99)	10(1)
2	경북대학교	11	11(100)	0(0)
3	경희대학교	6	4(67)	2(33)
4	고려대학교	9	9(100)	0(0)
5	광역유선방송	4	4(100)	0(0)
6	단국대학교	5	5(100)	0(0)
7	대구대학교	6	6(100)	0(0)
8	대우정보시스템	6	6(100)	0(0)
9	덕성여대	4	4(100)	0(0)
10	동국대학교	10	10(100)	0(0)
11	롯데정보통신	4	4(100)	0(0)
12	삼성SDS	6	6(100)	0(0)
13	서울대학교	11	10(91)	1(9)
:	:	:	:	:
50	THRUNET	117	117(100)	0(0)
51	UNITEL	18	11(61)	7(39)
52	Videsh Sanchar Nigam Ltd	4	4(100)	0(0)
	Total	1580	1498(95)	82(5)

<그림 1> Loged IP 정보



Clusters ID	Users ID
1	1 4 45 53 56 83 89 97 121 138 143 152 162 166 172 173 175 204 282 287 292 340 362 425 429 433 442
2	7 14 117 135 153 161 203 435
3	118 176 186 194 234 339 357 373 486
4	12 21 26 107 110 115 123 159 171 392
5	6 50 70 76 80 90 98 122 154 157 165 170 201 227 247 260 269 290 347 438 449
6	2 63 99 156 163 164 169 177 179 180 185 187 188 208 210 233 238 279 331 334 336 404 407

<그림 2> 사용자 군집분석

제시된 예제는 3장에서 언급한 분석 방법 중 일부의 과정을 보여주기 위한 것이며, 실제로 이용하기 위한 목적은 아님을 밝혀둔다. Web 데이터의 분석 결과를 어떤 포맷으로 보고 또 어떻게 활용할 것인가의 문제는 기업의 환경을 고려해야 하는 다른 문제이기 때문이다.

5. 결론

인터넷은 흔히 '정보의 바다'로 표현되고 있으며, 우리는 정보의 홍수 속에 살고 있다는 말을 자주 들을 수 있다. 정보의 홍수, 즉 정보가 매우 많다는 이야기는 역설적으로 확실한 정보가 없다는 의미로 받아들일 수 있으며, 이는 우리가 이용하고자 하는 데이터에도 그대로 적용된다. 따라서 많은 데이터로부터 정보를 추출할 수 있는 기법들에 관한 연구가 요구되고 있다.

사실 데이터의 활용 및 응용 방법에 관한 내용이 새로운 개념은 아니다. 그러나 실생활에서 데이터가 대량화되어지고 Log 데이터와 같이 기존에 존재하지 않았던 새로운 형태의 데이터들이 출현함에 따라 이 데이터들을 적절하게 이용할 수 있는 방안들이 연구되어질 필요성이 있다.

본 연구에서는 Web 데이터를 획득되는 특성에 따라 구분하고, 데이터로부터 정보를 추출하기 위한 몇 가지 분석 방법들을 제안하였다. 제안된 데이터 분석 방법들은 일반적인 정보 추출에 주안점을 두고 있으며, 이러한 분석 방법들의 적절한 활용은 사용자 만족도 향상은 물론 기업 이윤 추구에 많은 역할을 할 것이다.

앞서 언급한 바와 같이 현재까지는 Server Log 데이터의 이용에 관한 연구가 대부분을 차지하고 있다. 앞으로 대규모 데이터를 분석하기 위한 방법, 새로운 형태를 갖는 데이터를 분석하기 위한 세부적인 방법 등에 관한 연구가 필요할 것으로 생각된다. 또한 Server Log와 같은 (사용자들이 데이터의 발생을 인식하지 못하는)데이터가 이용되기 때문에 사용자들의 프라이버시(privacy)에 관한 문제도 검토되어야 할 것으로 생각된다.

References

[1] Borges, J. and Levene, M. (1999), Data Mining of User Navigation Patterns, *Proceedings of the workshop on Web Usage Analysis and User Profiling (WebKDD'99)*, http://www.acm.org/sigs/sigkdd/proceedings/webkdd99/toconline.htm

[2] Buchner, A.G., Baumgarten, M., Anand, S.S., Mulvenna, M.D. and Hughes, J.G. (1999), Navigation Pattern Discovery from Internet Data, *Proceedings of the workshop on Web Usage Analysis and User Profiling (WebKDD'99)*, http://www.acm.org/sigs/sigkdd/ proceedings/webkdd99/toconline.htm

[3] Cooley, R. W. (2000), Web Usage Mining: Discovery and Application of Interesting Patterns from Web Data, PhD Dissertation, University of Minnesota.

[4] Famili, A., Shen, W. M., Weber, R. and Simoudis, E. (1997), Data Preprocessing and Intelligent Data Analysis, *Intelligent Data Analysis*, Vol. 1, No. 1, http://www-east.elsevier.com/ida/browse/vol1.htm

[5] Friedman, J.H. (1997), Data Mining and Statistics : What's the Connection?, *Proceedings of the International Conference on the Interface : Computing Science and Statistics*, http://www.stat.rice.edu/interface97.html

[6] Garofalakis, M.N., Rastogi, R., Seshadri, S. and Shim, K.S. (1999), Data Mining and the Web: Past, Present and Future, *Proceedings of the workshop on Web Information and Data Management*, http://cs.kaist.ac.kr/~shim/

[7] Guha, S., Rastogi, R. and Shim, K.S. (1998), CURE: An Efficient Clustering Algorithm for Large Databases, *Proceedings of ACM SIGMOD International Conference on Management of Data*, 73-84, http://cs.kaist.ac.kr/~shim/

[8] Hand, D. J. (1998), Intelligent Data Analysis : Issues and Opportunities, *Intelligent Data Analysis*, Vol. 2, No. 2, 1-14.

[9] Hartigan, J.A., Wong, M.A. (1979), Algorithm AS 136: A K-means clustering algorithm, *Applied Statistics*, Vol. 28, 100-108

[10] Huber, P.J. (1994), Huge Data Sets, *COMPSTAT(Proceedings in Computational Statistics)*, 3-13.

[11] Kohonen, T. (1984), *Self-Organization and Associative Memory*, Springer-Verlag

[12] Pei, J., Han, J., Mortazavi-asl, B. and Zhu, H. (2000), Mining Access Patterns Efficiently from Web Logs, *Proceedings of the Pacific-Asia Conference on Knowledge Discovery and Data Mining*, http://db.cs.sfu.ca/sections/publication/kdd

[13] Spiliopoulou, M. (2000), Data Analysis for web marketing and merchandizing applications, *Proceedings of the 4th European Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD'2000)*, http://www.wiwi.hu-berlin.de/~myra/

[14] Srivastava, J., Cooley, R. W., Deshpande, M. and Tan, P.N. (2000), Web Usage Mining: Discovery and Applications of Usage Patterns from Web Data, *SIGKDD Explorations*, Vol. 1, Issue 2, http://www.acm.org/sigs/sigkdd/explorations/

[15] Zhang, T., Ramarkrishnan, R. and Livny, M. (1996), BIRCH: An Efficient Data Clustering Method for Very Large Databases, *Proceedings of ACM SIGMOD International Conference on Data Management*, http://citeseer.nj.nec.com/zhang97birch.html