2000 Proceedings of the Autumn Conference, Korean Statistical Society

웹 로그(Web Log) 분석을 통한 정보의 활용

김석기[1], 안정용[2], 한경수[3], 한범수[4]

요 약

인터넷이 데이터 저장 및 서비스를 위한 도구로 폭넓게 활용되고 있으며, 이 과정에서 웹 서버 방문객에 대한 정보인 로그가 발생된다. 이러한 로그는 방문객 주소, 참조 페이지, 방문 시각 등의 정보를 포함하고 있다. 웹 로그에 대하여 패턴분석(pattern analysis), 군집분석(clustering), 판별분석(classification) 등의 통계적 분석을 통하여 방문객이 관심을 가지는 항목이나 항목간의 연관관계 등 새로운 정보를 생성하여 웹 디자인 또는 비즈니스에의 적용에 대한 연구가 활발히 논의되고 있다. 본 연구에서는 웹 로그 분석에 대하여 소개하고 웹 로그 분석을 위한 방안을 제시하고자 한다.

주요용어 : Web log, web log mining, association rule

1. 서론

인터넷이 전 세계 지식 창고의 역할을 훌륭하게 수행함에 따라 이를 통하여 수많은 정보가 교류되고 있으며, 그 부산물인 웹 로그 역시 분석을 필요로 하는 방대한 양의 데이터로 존재하게 되었다.

웹 로그는 웹 서버에 대한 모든 접근을 기록한 데이터로 로그의 형태는 HTTP 프로토콜의 일부로 명시된 *Common Log Format*(A. Luotonen, 1995) 형태나 *Extended Log Format*(P. Hallam-Baker, B. Behlendorf, 1996)을 따라 저장된다. 저장된 로그 파일은 웹 서버에 접속한 사용자의 IP 주소, 접근 시각, 접근 방법, 대상 URL, 전송 프로토콜, 에러 코드, 전송 바이트 수와 같이 사용자를 인식할 수 있는 정보와 사용자에 대한 방문 정보들을 포함하고 있다.

Bamshad 등(1996)이 제시한 바와 같이 웹 로그 분석을 통하여 방문자들에 대한 접속 유휴시간 측정, 상품과 서비스를 교차한 교차 판매(cross marketing) 계획 수립, 캠페인 효과에 대한 평가, 자료 배치를 위한 웹 공간의 효율적인 논리적 구조 등의 결과를 얻을 수 있다. 이러한 분석 결과를 얻기 위해 빈도 수, 평균, 중앙값과 같이 단순 통계량을 비롯한 패턴분석, 군집분석 및 판별분석들이 이루어지고 있다. 특히 Bamshad(1996), Dong-Ha Lee(1998), Jerome(1997), Yongjian(1999) 등에서 볼 수 있듯이 연관규칙(association rule) 탐사와 군집분석에 대한 활발한 연구가 진행되고 있다.

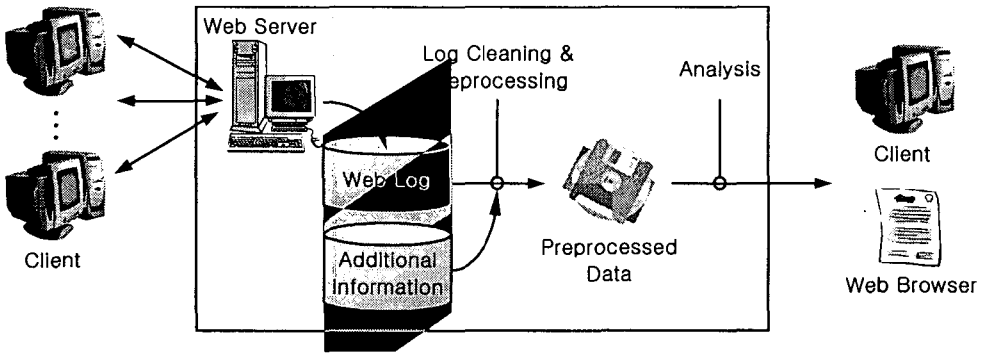
웹 로그를 분석하기 위해서는 기본적인 로그 데이터 이외에 추가적으로 웹 관리자에 대한 정보, 웹 검색 엔진 정보, 웹의 논리적 구조 등과 같은 정보가 필요하다. 웹 로그는 일반적인 방문객들의 정보만을 저장하는 것이 아니며 웹 관리자나 개발자의 작업 내용 및 검색 엔진에 의한 조회 기록까지 저장되기 때문에 분석에 있어서 이러한 자료들을 제거시켜야만 한다. 또한 지역별, 기관별 방문 현황을 분석하기 위해서는 IP 주소를 통하여 얻을 수 있는 IP 주소 할당 정보인 WHOIS 정보가 필요하다.

1) 군산간호대학 간호정보공학센터 연구원, 전북 군산시 개정동 413
2) 서남대학교 컴퓨터정보통신학부 조교수, 전북 남원시 광치동 720
3) 전북대학교 수학 • 통계정보과학부 교수, 전북 전주시 덕진구 덕진동 664-14
4) 전북대학교 전산통계학과 박사과정, 전북 전주시 덕진구 덕진동 664-14

본 연구에서는 웹 로그 분석을 위한 방법과 절차에 대하여 소개하고, 추가 정보의 활용 방안을 제시하고자 한다. 또한 제안된 방법을 통하여 구현된 사례를 소개하고자 한다.

2. 웹 로그 분석 시스템의 구성

일반적인 웹 로그 분석 시스템은 <그림 1>과 같이 웹 서버에서 생성되는 로그와 추가 정보에 대하여 정제(cleaning)와 사전처리(preprocessing) 절차를 거치게 되고, 그 결과 생성된 로그에 대한 분석 결과를 클라이언트 또는 웹 브라우저로 표현하는 구조로 되어있다.



<그림 1> 웹 로그 분석 시스템

분석의 기반이 되는 웹 로그를 수집하는 방법은 Jaideep 등(2000)이 설명한 바와 같이 데이터의 수집 위치에 따라 server level, client level, proxy level로 구분할 수 있다.

수집된 로그는 웹 페이지에 대한 정보 이외에도 웹 페이지에 삽입된 이미지나 오디오와 같은 미디어에 대한 참조 정보 역시 포함되어있다. 따라서 이러한 정보를 제거하기 위한 정제 작업이 필요하며, 이 과정에서 웹 개발자나 관리자에 대한 참조 정보와 웹 검색 엔진인 web robot에 대한 참조 정보 역시 제거해야 한다.

사전처리 절차에서는 웹 로그에 대한 분석을 용이하게 수행하기 위해 정제된 로그를 구조화하는 작업이 수행된다. 이 단계에서 이루어지는 주요 작업 중 하나가 session identification이다. Session identification이란 사용자들의 방문 경로를 체류 시간을 기준으로 구분하는 작업으로 한번의 방문으로 이루어진 경로인지 여러 번의 방문으로 인해 이루어진 경로인지를 구분하게 된다.

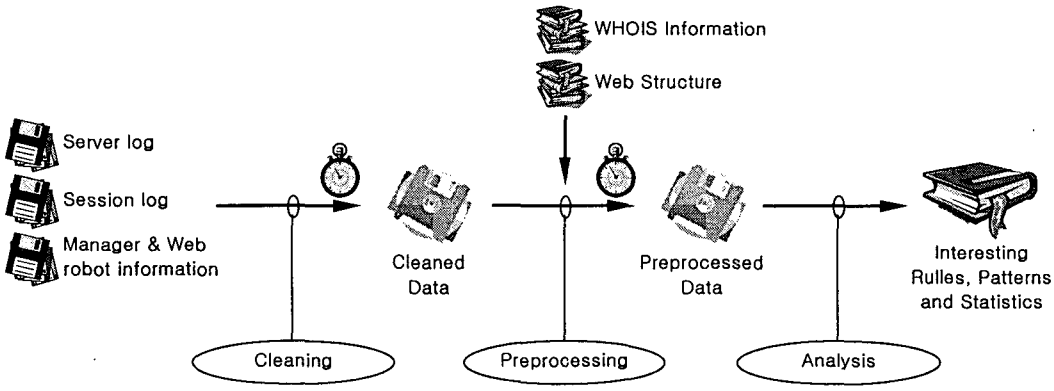
로그 정제와 사전처리 과정을 거쳐 생성된 데이터를 기반으로 조회 수(hit measure), 연관규칙 탐사, 군집분석, 판별분석과 같은 실제적인 로그 분석이 이루어지게 된다. 로그 분석 과정에서 지역별/기관별 방문 현황을 제공하기 위해 WHOIS 서비스를 통해 IP 주소 할당 정보를 사용할 수 있다. 또한 웹 서버에 게시된 항목들과 연관된 분석을 위해서는 웹 구조에 대한 정보의 구축이 필요하며 게시된 항목들의 위치가 구조화되어 있어야 한다.

분석이 완료된 정보들은 클라이언트 어플리케이션을 통해 오프라인(off-line) 형태로 제공될 수 있으며, 웹 어플리케이션을 이용하여 온라인(on-line) 형태로 제공될 수도 있다.

3. 웹 로그 분석 시스템의 구현

본 연구에서 구현한 시스템은 <그림 2>와 같은 절차를 통해 웹 로그 분석이 이루어진다. 웹 로그 분석은 로그 수집, 정제, 사전처리 절차를 거쳐 수행되어지며 웹 서버에서 수집되는 로그 이외에 session log, 웹 관리자와 개발자 정보, 웹 검색 엔진 정보, WHOIS 정보, 웹의 논리적

구조와 같은 추가 정보를 사용한다.



<그림 2> 웹 로그 분석 절차

3.1. 로그 수집(Log Gathering)

본 연구에서 웹 서버로 사용한 Microsoft Internet Information Server에서 server level 로그를 수집하는 방법은 3가지 방법으로 구분할 수 있다. 웹 서버에서 제공하는 방법을 이용하여 로그를 수집하는 방법과 사용자가 웹 서버에 접속되는 시점에서 세션이 생성된다는 점을 이용하여 로그를 수집하는 방법이다. 마지막으로 분석자가 관심을 가지는 특정 페이지에 로그를 기록할 수 있는 스크립트 코드를 삽입하는 방법이다. 이 경우 수집된 로그는 이미 정제된 데이터로 분석이 용이하다는 장점을 가진다.

수집된 모든 로그는 분석 절차를 용이하게 하기 위해 데이터베이스에 저장된다. 본 연구에서는 로그의 저장을 위해 SQL Server 2000을 사용하였다.

3.2. 로그 정제(Log Cleaning)

앞 절에서 언급한 바와 같이 수집된 데이터는 정제 과정을 거쳐야 된다. 정제 과정을 통해 생성된 데이터는 분석의 대상이 되는 문서 자료들로만 구성되어있다. 또한 웹 서버 관리자와 개발자의 로그 정보는 물론 웹 검색 엔진인 web robot에 의해 탐색된 정보 역시 제거된 자료이다. 로그 기록자가 web robot인지의 판단은 기록된 로그 정보 중 방문객의 웹 브라우저 정보를 나타내는 *HTTP_USER_AGENT* 속성을 이용하여 결정한다.

3.3. 사전처리(Preprocessing)

사전처리를 수행하기 위해서는 웹 서버에서 기본적으로 수집되는 로그 이외에 WHOIS 정보와 웹의 논리적 구조에 대한 정보가 필요하다. WHOIS 정보란 각국의 인터넷 정보 센터에서 제공되는 IP 주소 할당 정보로 국내에 배정된 IP 주소에 대한 등록 정보는 한국인터넷정보센터(KRNIC; http://ipwhois.nic.or.kr)에서 조회할 수 있다. 또한 웹의 논리적 구조에 대한 정보는 웹 로그에 대한 분석을 통하여 생성된다.

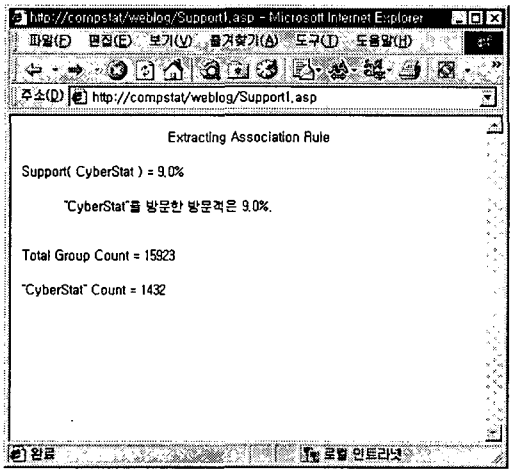
김광용(2000), Dong-Ha Lee(1998), Bamshad(1996) 및 Yongjian(1999)이 제시한 session identification 방법은 방문자 로그의 maximum idle time을 이용한 방법이다. Maximum idle time을 이용한 방법은 초기 idle time을 설정하고 data mining 절차를 통하여 최적의 시간을 산출하여 방문자 로그를 session 별로 구분하는 방법이다. 그러나 이러한 방법은 방문자 로그를

session 별로 구분하는데 있어서 실제 방문 형태에 기반한 것이 아니라 추론에 의해 구분한다는 문제를 가지고 있다. 즉, 실제로 방문자가 웹 방문 중 특정 페이지를 오랜 시간 참조할 경우 새로운 session으로 구분한다는 것이다. 따라서 이러한 문제를 보완하기 위해 본 연구에서는 방문객이 웹 서버에 접속할 때 발생되는 session 정보를 이용하여 방문자 session을 구분하고, session 정보가 존재하지 않는 로그에 대해서는 maximum idle time을 적용하는 방법을 사용하였다.

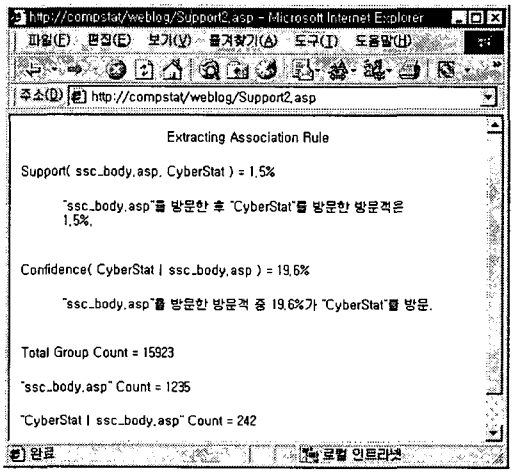
3.4. Summary and Analysis

웹 로그 분석 시스템을 통해 제공되는 로그 분석 결과는 문서 조회 수를 기반으로 한 방문지별 조회 현황과 조회 빈도가 높은 문서들에 대한 분석 자료를 제시하며, 하향탐색(drill-down) 방식으로 분석 결과에 대한 조회가 가능하다. 또한 웹 로그에 기반하여 웹 서비스 상태에 따른 분석 결과와 웹의 논리적 구성에 대한 분석 결과를 제시한다.

연관규칙 탐사와 관련하여 <그림 3>과 <그림 4>에서 볼 수 있듯이 전체 session 중 특정 항목에 대한 참조 비율인 support(A), support(A,B)와 특정 항목 참조 후 다른 항목(A)을 참조한 비율인 confidence(B|A)에 대한 분석 결과를 제시한다.



<그림 3> Support(A)



<그림 4> Support(A,B), Confidence(B|A)

4. 결론

본 연구에서는 웹 로그 분석 시스템에 대하여 소개하고 분석을 위한 절차에 대하여 살펴보았다. 또한 웹 로그 분석을 위해 WHOIS 정보, 웹의 논리적 구조에 대한 정보 및 session 정보 등과 같은 추가 정보의 필요성 및 활용 방안에 대하여 언급하였으며, 이러한 추가 정보를 이용하여 실제 분석 사례를 보였다.

웹 로그 분석은 웹 사용자 개개인에 최적화 된 서비스와 웹에 게시될 정보들의 효율적인 배치 및 마케팅에 직접 활용할 수 있는 정보를 제공한다. 또한 CRM(Customer Relationship Management)을 위한 분석으로 활용되어지고 있다. 따라서 인터넷을 기반으로 한 마케팅 및 정보 서비스에 있어서 중요한 위치를 담당하게 될 것이다.

보다 효율적인 웹 로그 분석을 위해서는 본 연구에서 제시한 추가 정보의 활용 이외에도 웹 사용자의 등록 정보인 user profile과 웹 로그를 통합한 분석이 필요하다. 또한 분석 성능 향상

을 위해 데이터베이스 query를 최적화하는 것은 물론 웹 로그 분석에 적절한 패턴분석, 군집분석 및 판별분석과 같은 분석 알고리즘에 대한 연구가 충분히 이루어져야 할 것이다.

참고문헌

[1] 김광용. (2000). Web Information Center와 Internet Survey. Internet Survey Workshop 논문집. 111-122.

[2] A. Luotonen. (1995). The Common Logfile Format, Technical Reports, World Wide Web Consortium, *http://www.w3.org/Daemon/User/Config/Logging.html*

[3] Bamshad Mobasher, Namit Jain, Eui-Hong (Sam) Han, Jaideep Srivastava. (1996). Web Mining: Pattern Discovery from World Wide Web Transactions. Technical Report 96-050, Department of Computer Science, University of Minnesota, Minneapolis.

[4] Dong-Ha Lee, Dong-Yal Seo, Nam-Ho Kim, Jeon-Young Lee. (1998). Discovery and Application of User Access Patterns in The World Wide Web. *Proceedings of the 4th World Congress on Expert Systems*. 321-327.

[5] Jaideep Srivastava, Robert Colley, Mukund Deshpande, Pang-Ning Tan. (2000). Web Usage Mining: Discovery and Applications of Usage Patterns from Web Data, *Newsletter of the ACM Special Interest Group on Knowledge Discovery and Data Mining*, Volume 1, Issue 2, 12-23.

[6] Jerome Moore, Eui-Hong Han. (1997). Web Page Categorization and Feature Selection Using Association Rule and Principal Component Clustering. *7th Workshop on Information Technologies and Systems*.

[7] Phillip M. Hallam-Baker, Brian Behlendorf. (1996). Extended Log File Format, Technical Reports, World Wide Web Consortium, *http://www.w3.org/TR/WD-logfile*

[8] Yongjian Fu, Kanwalpreet Sandhu, Ming-Yi Shih. (1999). Clustering of Web Users Based on Access Patterns. *WEBKDD'99 Workshop on Web Usage Analysis and User Profilling*.