웹사이트 구조 개선을 위한 웹페이지 연관 규칙 발견과 웹사이트 성능 평가

김민정[0] 박승수
이화여자대학교 컴퓨터학과
(min, sspark)@ewha.ac.kr

Discovering Web Page Association Rules & Evaluating Web Site Performance To Improve Web Site Structure

Min-Jeong Kim[0] Seung Soo Park
Dept. of Computer Science & Engineering, Ewha Womans University

요 약

현재 수많은 웹사이트들이 웹상에 존재하며 서비스를 하고 있다. 사용자는 여러 웹사이트 중에서 접속하기 편하고 잘 구성된 웹사이트에 접속하기 마련이므로, 잘 구성된 웹사이트 운영은 그 웹사이트의 생존 전략이며 방문자 유지에 필수적이다. 이를 위해 사용자들이 웹사이트에 접속한 기록이 남아 있는 웹 서버 로그데이터(이하 웹로그파일)를 분석하여 사용자들의 브라우징 패턴과 접속 경향, 웹 서버의 에러 발생 정보 등을 파악할 수 있다.

본 논문에서는 Web Usage Mining 과 Web Structure Mining 작업으로 로그파일 분석과 웹사이트 구조 분석을 수행하여 페이지들의 연관 관계와 웹사이트의 구조 정보를 발견해서 웹사이트의 구조를 개선하는 방안을 제안하고자 한다.

1. 서 론

인터넷 이용의 기하급수적인 증가로 수많은 웹사이트들이 사용자에게 다양한 정보를 제공하고 있다. 모든 웹사이트의 일차적인 목적은 많은 사용자가 접속하여 방문을 하도록 하는 것이라고 해도 과언이 아닐 것이며, 특히 상업적인 웹사이트의 경우는 웹사이트의 존재 여부와 직결될 만큼 중요한 문제이다. 사용자는 접속이 잘 되지 않거나 잘 구성되지 않은 사이트는 다시 방문하지 않고 다른 사이트에 방문할 것이다. 웹사이트에 많은 사용자가 접속하게 하기 위해서는 일방적으로 정보를 제공하는 방식에서 탈피하여 대화형 방식을 적용하거나, 사용자의 브라우징 패턴을 고려하여 방문하기 편하도록 설계되어야 한다.

웹 마이닝이 수행되는 웹데이터 중에서는 사용자가 직접 입력한 정보도 있지만 잘못된 데이터를 입력하는 경우도 많으며 계속 갱신이 되지 않는다는 문제점이 있으므로 이것을 분석하여 사용자의 경향을 파악하는 방법 보다는 실시간으로 웹 서버에 사용자의 접속 정보가 기록되는 로그파일을 분석하는 방법이 객관적이고 정확한 분석이라고 할 수 있다.[1]

본 논문에서는 웹 마이닝의 여러 목적들[2]-웹사이트 성능 분석, 웹사이트 디자인 개선, 웹트래픽 이해, 웹사이트 개인화 중에서 웹사이트 디자인 개선에 초점을 두어 웹사이트의 구조를 개선하여 많은 사용자가 웹사이트에 방문하도록 하는 것을 목적으로 한다. 웹로그파일을 분석하여 웹사이트의 페이지들 사이의 연관 규칙을 발견하고, 웹 서버의 트래픽 양과 에러 내용에 관한 분석도 수행한다. 또한 웹마이닝 과정에서 기존의 마이닝 툴에서는 파일 자체만 가지고 분석하는 작업만이 지원되었기 때문에 웹사이트의 구조 정보를 발견할 수 있는 시스템을 구현하여, 발견된 연관 규칙과 페이지 간의 관계를 파악하는데 도움이 되도록 한다.

본 논문의 구성은 2장에서는 웹 마이닝에 관련된 연구와 응용 분야에 대해서 알아보고, 3장에서는 본 논문에서 수행한 웹 마이닝 과정, 웹사이트 구조 분석 과정, 전체적인 시스템 구성을 제시하고, 4장에서는 실제 웹사이트의 웹로그파일에 적용한 실험 결과를 분석하며 마지막으로 5장에서 결론과 향후 과제를 논의한다.

2. 관 련 연 구

2.1. 웹 마이닝

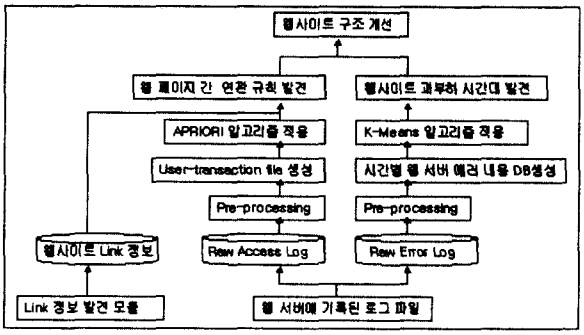
웹 마이닝은 웹 문서와 서비스들로부터 자동적으로 알려지지 않은 유용한 정보를 추출하고 발견하기 위한 과정으로, 데이터 마이닝 기술들을 웹 데이터에 적용한 응용 분야이다. 웹 마이닝은 웹의 리소스를 분석하는 Web Content Mining, 웹사이트의 링크 구조와 계층 관계를 분석하는 Web Structure Mining, 사용자의 접속 패턴을 발견하는 Web Usage Mining 으로 분류된다[3]. 본 논문에서는 Web Usage Mining을 수행하고, 결과 분석을 위해 Web Structure Mining을 수행하였다.

* 이 논문은 BK21(Brain Korea21)사업의 지원으로 연구되었음

2.2. 응용 분야

웹로그파일을 마이닝한 결과를 기반으로 사용자의 다음 방문을 돕는 추천 시스템에 관한 연구[1], 페이지 사이의 연관 규칙을 발견하기 위한 효과적인 알고리즘에 대한 연구[4] 등이 활발히 이루어지고 있으며, 웹 마이닝의 중요성이 강조되면서 상업적인 마이닝 툴에서도 웹 마이닝을 지원하는 템플릿이 개발되고 있다[5][6].

3. 시스템 구성



[그림 1] 시스템 구성도

3.1. 데이터 집합

본 시스템의 데이터 소스는 웹로그파일, 웹페이지의 링크 정보 데이터 베이스이다. 웹서버에 따라 다른 포맷의 로그 데이터가 생성되며, 여러 개의 로그 파일을 생성할 수도 있는데 Access(Transfer), Error, Referrer, Agent 로그 파일의 4종류가 있다. 엑세스(access)로그 파일은 반드시 존재하는 로그 파일로 접속한 사용자에 관한 일반적인 사항이 기록되며 나머지 로그 파일은 선택적으로 기록할 수 있다. 본 논문에서는 엑세스 로그 파일과 에러 로그 파일을 사용했고, 발견된 연관 규칙을 해석하는 과정에서 웹사이트 링크 정보 데이터 베이스를 참고했다.

3.2. 전처리 과정

웹 서버에 기록된 초기 로그 파일은 양이 방대하므로 분석에 필요한 내용만 필터링 해야 한다. 엑세스 로그 파일에서는 여러 필드 중 IP, 접속 시간, 접속 페이지 필드만 남겨 놓고 필터링한 후, 접속 페이지의 확장자가 *.htm, *.html인 데이터만 추출한다. 그리고 사용자 별로 방문을 결정하고, 한번의 방문 동안 이동한 경로를 세션으로 결정한다. 이 때 사용자는 IP 주소로 구별되며, 동일 IP 라도 마지막 접속 후 30분 이후의 접속은 새로운 세션이라고 가정한다.[7][8] 엑세스 로그 파일에 전처리 작업을 수행한 최종 결과는 사용자 트랜잭션 (User-transaction) 파일이다. T={(s_id1,page1,page2 ,…,page_n),(s_id2,page1,page2,…,page_n), (s_id_n,page1,page2,…, page_n)} 의 형태로 저장한다.

에러 로그 파일의 경우는 웹 서버에 에러가 발생한 시간과 내용만 저장되므로 엑세스 로그 파일보다 크기가 훨씬 작다. 통계 작업을 하거나 마이닝 알고리즘을 적용하기에 부적합한 에러 메시지는 제거하고 키워드만 남겨 놓는다. 에러가 발생한 시간대와 내용에 관한 E={Time, Error} 의 형태로 저장된다

3.3. 연관 규칙 발견

엑세스 로그 파일에 전처리 과정을 수행한 사용자 트랜잭션 파일에 마이닝 알고리즘을 적용하여 웹페이지 사이의 연관 규칙을 발견할 수 있다. 발견된 규칙을 분석하는 과정에서 웹사이트 내에서의 사용자들의 움직임을 파악할 수 있고, 웹페이지 사이의 링크 관계를 효율적으로 구성하는 데 도움이 된다.

연관 규칙은 A=>B 로 표현할 수 있으며, A를 포함하고 있는 데이터 베이스 내의 트랜잭션은 B도 함께 포함하고 있다는 의미이다. 본 논문에서 발견한 연관 규칙 집합의 형식은 [표 1] 과 같다.

[표 1]

Rule # for page1 :
If page3 == *.htm
Then -> page1 (instances, confidence)

여기서 page1, page3 은 방문된 임의의 페이지이며, instances 는 연관 규칙을 만족하는 트랜잭션의 수이고, confidence 는 page1을 포함하고 있는 트랜잭션의 수 중에서 page1 과 page3을 모두 포함하는 트랜잭션의 수의 비율이다.

기존의 마이닝 툴을 사용하여 연관 규칙을 발견하는 것은 가능했지만, 규칙을 해석하는 과정에서 필요한 웹사이트의 구조 정보까지는 발견할 수 없었기 때문에 웹사이트의 링크 페이지들의 정보를 발견하는 모듈을 작성하였다. 입력 화면에서 웹사이트의 초기 화면 주소를 입력 받아 링크 정보를 L= { URL, Title, Description, Keyword, Page text, Inlink, Linktext, Outlink, Stamp}의 형태로 데이터 베이스에 저장된다. 각 필드는 링크된 페이지들의 주소, 타이틀, 페이지 설명, 키워드, 페이지 내용, 들어오는 링크, 링크가 걸려 있는 글자, 나가는 링크의 개수, 페이지를 방문한 시간을 의미한다.

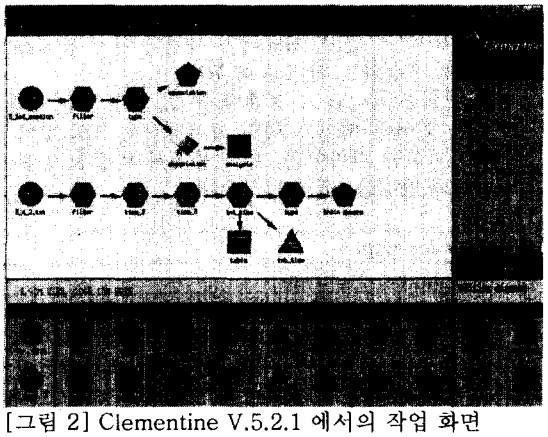
3.4. 웹사이트 과부하 시간대 파악

전처리 과정을 거친 에러 로그 파일을 사용하여 웹사이트에서 에러가 자주 발생하는 시간대와 에러의 원인을 파악할 수 있다. 웹 관리자는 이런 결과를 바탕으로 과부하 시간대 전에 웹 서버를 점검하는 등의 방법으로 웹사이트의 서비스를 향상 시키거나, 웹 서버 구입 등에 대한 정책 결정 시에도 이용할 수 있다.

4. 실 험

본 논문에서는 실험을 위해 2001년 4월 첫째 주 이화여자대학교 홈페이지 로그 파일을 사용했다. 로그 파일의 포맷은 CLF(Common Log file Format)이고, 파일의 크기는 엑세스 로그 파일의 경우는 하루에 약 100MB 정도였고, 에러 로그 파일의 경우는 하루에 약 1MB 정도였다. 시스템 환경은 Windows 2000 Professional 상에서 로그 파일의 전처리과정과 마이닝 작업을 위해 마이닝 툴인 클레멘타인(Clementine) V.5.2.1 [5][6]을 사용했으며[그림 2], [표 2],

[그림 3] 링크 정보를 데이터 베이스에 저장하는 모듈은 JDK 1.3.1 환경에서 자바 서블릿으로 작성했고, 서블릿 엔진으로는 JSDK 2.1를 사용했으며, 데이터 베이스는 MS Access 의 .mdb 파일에 테이블 형태로 저장된다. [그림 4]



[그림 2] Clementine V.5.2.1 에서의 작업 화면

[표 2] 발견된 연관 규칙의 집합의 일부

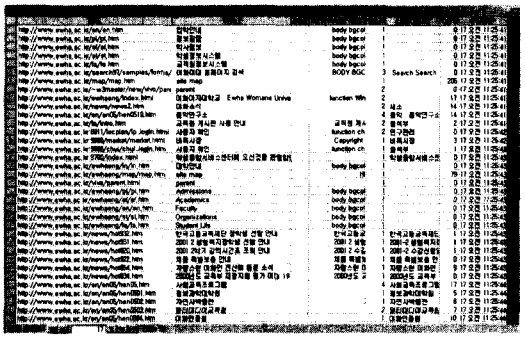
Rules for /:
Rule #1 for /:
if page5 == /st/st.htm
Then -> / (1615, 0.918)
Rule #2 for /:
if page3 == /st/frst.htm
Then -> / (1615, 0.917)
Rule #3 for /:
If page2 == /~hak/sug_login.htm
then -> / (1615, 0.92)

[그림 3] 웹 서버의 시간대 별 에러 내용

발견된 연관 규칙을 기반으로 사용자가 웹사이트를 방문하는 형태를 알 수 있었다. 학교 홈페이지의 경우 특정 페이지의 접속만 활발히 이루어 지는 경향이 있었고(수강 신청 페이지), 몇 개의 페이지가 반복해서 연관 규칙으로 발견되는 결과는 사용자가 웹사이트 내에서 원하는 정보를 잘 찾지 못하는 결과로 해석되므로 웹사이트 링크 정보를 기반으로 사이트 구조를 재구성하는 방법이 제안된다. 또한 웹 서버의 시간대 별 에러 내용 분석 결과는 학생들이 학교에 가장 많이 있는 오전 9시부터 4시 시간대가 가장 많았으며 에러 내용이 failure 보다 warning 이 많으므로 비교적 안정적인 서비스를 하고 있음을 알 수 있다.

5. 결론 및 향후 연구

본 논문에서는 Web Usage Mining 을 수행하여 발견된 페이지 간의 연관 규칙들의 이해를 돕기 위해 Web Structure Mining 을 수행한 결과인 웹사이트 링크 정보를 사용하였다. Web Usage Mining 수행 시 기존의 연구에서



[그림 4] 테이블로 저장된 웹사이트의 링크 정보

는 엑세스 로그 파일에 대한 분석 위주로 이루어 졌는데, 에러 로그 파일에 대한 분석 작업도 수행하여 간단하지만 웹 서버의 서비스 향상에 도움이 될만한 결과를 얻을 수 있었고, 웹사이트 링크 정보를 데이터 베이스에 저장하여 연관 규칙과 결합하여 웹사이트 구조 개선에 이용할 수 있다. 향후 과제로는 본 논문에서 발견한 결과를 바탕으로 웹사이트 구조를 재구성해주는 시스템을 개발할 계획이다.

참고 문헌

[1] B.Mobasher, R.Cooley, and J.Srivastava. Automatic Personlization Based on Web Usage Mining. ACM, Vol.43(8), p142-151, August 2000 .18

[2] O.Zaiane, M.Xin,and J.Han. Discovering Web Access Patterns and Trends by Applying OLAP and Data Mining Technology on Web Logs. In Proc. Advances in Digital Libraries Conf.(ADL'98),p.19-29, 1998

[3] R.Kosala and H.Blockeel. Web Mining Research: A Survey. In ACM SIGKDD Explorations, Vol.2(1): p1-15, July 2000

[4] Y.Ma, B.Liu, and C.K. Wong. Web for Data Mining: organizing and Interpreting the Discovered Rules Using the Web. In ACM SIGKDD Explorations, Vol.2, Issue1, p16-23 June 2000

[5] T.Khabaza, and D.Sigerson. WebCAT : the Clementine Application Template for WebMining and Analytical eCRM, web-mining workshop paper,1st SIAM International Conference on Data Mining, Chicago, April 7, 2001

[6] T.Khabaza, and D.Sigerson. Intelligent Personalization through Business Insight:Web-mining and personalization in the WebCAT. SPSS Advanced Data Mining Group, February 2001

[7] J. Srivastava. R.Cooley, M.Deshpande, and P.N. Tan. Web-Usage Mining: Discovery and Applications of Usage Patterns from Web Data. In ACM SIGKDD Explorations, Vol.1(2), p12-23, Jan 2000

[8] R.Cooley, B.Mobasher, and J.Srivastava. Data Preparation for Mining World Wide Web Browsing Patterns. In Journal of Knowledge & Information Systems, Vol.1, No.1, p.5-32, 1999.