웹마이닝을 통한 도서관 홈페이지의 사용편의성에 관한 연구

- 육군대학 도서관 홈페이지를 중심으로 -

Study on the Usability Based on Web Mining in Army College Library Homepage

손용배, 이응봉 충남대학교 문헌정보학과

Yong-Bae Son, Eung-Bong Lee

Graduate School of Library & Information Science,
Chungnam National University

본 연구는 육군대학 도서관 홈페이지의 웹서버에 저장되어 있는 로그파일을 실험 데이터로 사용하여, 기존 데이터마이닝(data mining)의 기법들 중에서 연관규칙(association rules) 탐사 기법을 적용함으로써, 사용자들의 웹 항행에 대한 순차패턴을 추출하였다. 이를 분석하여 실제 사용자들이 효과적으로 사용할 수 있는 웹사이트 디자인을 제안하고 나아가 대상 웹사이트의 사용편의성을 평가하였다.

1. 서 론

인터넷은 20세기 후반에 등장하여, 21세기인 현재에 중대한 정보전달의 매체이자 일상 생활에 있어 없어서는 안될 필수품으로 자리 매김하고 있다. 하루에도 수많은 사이트들이 생겨나며, 새로운 삶의 양식과 비즈니스 패러다임을 형성하고 있다. 이러한 가운데, 웹사이트의 평가와 관련하여, 사용편의성(usability)의 평가에 관한 연구가 진행되고 있는데, 이는 웹의 특성상, 해당 사이트에 대해 사용자에게 효과적이고 만족감을 주어야 하는 사용편의성의 요소가 기타 하드웨어나, 소프트웨어의 경우보다 중요하기 때문이다.

웹의 사용편의성을 평가하기 위하여 발견적 평가법(heuristic), 설문지법(questionnaire), 인터뷰(interview), 사용자관찰(user observation), 사용자반응(user feedback)의 반영 등 많은 기법들이 소개되었고 다양한 방법론들이 현재 연구되고 있다. 이러한 모든 방법들의 목적은 개발자의 관점이 아닌, 사용자가 중심이 되는 설계를 위한 것으로서, 사용자의 프로파일을 획득하고 이를 통해 시스템을 평가하고, 개발하는데 있다.

본 연구에서는 웹의 사용편의성을 평가하는 방법의 하나로 로그파일을 이용하여 사용자의 프로파일을 추출하는 웹마이닝을 이용하였다. 즉, 웹서버에 저장되는 로그파일을 데이터베이스화하여 이를 정제한 후, 데이터마이닝 분석툴인 'SAS Enterprise Miner program'을 사용하여 연구를 수행하였다.

대상 웹사이트로는 육군대학 도서관 홈페이지를 이용하였으며, 이를 통해 기존에 방문했던 사용자들의 접속패턴을 추출하였고, 이를 현재 웹페이지의 사용편의성에 대한 문제점으로 제기한 후, 사용편의성 향상을 위해 개선되어야 할 가이드라인을 제시하는 과정을 수행하였다.

2. 연구배경

2.1 로그파일

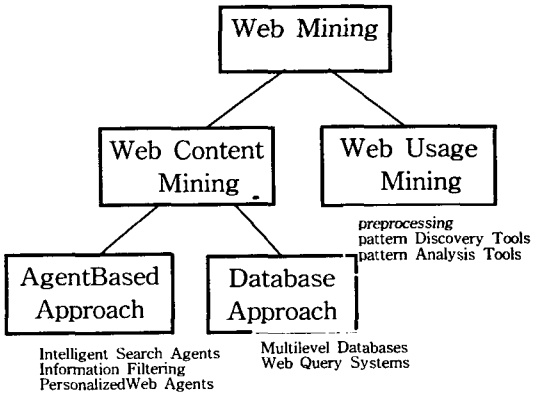
웹서버에 대한 모든 접근 기록을 로그파일이라고 한다. 현재 대부분의 웹서버는 표준 로그파일 형식(common log format)에 따라 로그파일을 생성하고 있다. 로그파일은 웹서버가 지정하는 곳에 위치하며, 웹서버 관리자가 웹서버를 설치할 때 로그파일의 위치와 기록방법 등을 지정하게 되어 있다. 웹서버에 따라서는 한 개가 아닌 여러 개의 로그파일을 생성할 수도 있는데, transfer, access, error, referrer 및 agent 로그파일 등이 있다. 대개 로그파일 분석이라 함은 access 로그파일 분석을 의미하며 본 연구의 실험 데이터도 access 로그파일을 사용하였다.

이와 같이 로그파일이란 웹서버를 통해 이루어지는 모든 작업들에 대한 기록들로 사용자가 웹서버에 접속을 하게 되면 사용자가 요청하는 특정 웹페이지와 관련된 이미지파일, 접속시간, 접속상태와 그 이후의 모든 작업들이 미리 정해놓은 위치에 데이터로 남게 된다. 이러한 데이터들은 사용자들의 실제 웹사이트의 사용에 대한 정보를 담고 있으며. 이를 정제하여 분석하면 사용자들의 프로파일을 추출할 수 있는 유용한 정보가 된다. 프로파일의 분석기법은 클라이언트에서 캐쉬를 사용하는 경우와 브라우저상의 '뒤로(back)', '앞으로(forward)' 버튼을 사용하는 경우, 기록이 남지 않는다는 제약이 있으나, 웹마이닝을 통한 순차패턴의 탐색과 해당 웹의 구조분석을 통해 추정할 수 있다.

2.2 데이터마이닝과 웹마이닝

데이터마이닝이란 방대한 양의 데이터 속에서 직접, 혹은 간접적인 방법으로 의미 있는 패턴이나 규칙을 탐사하고, 분석하는 방법이다. 1990년대 이후 데이터의 양이 많아짐에 따라 유용한 정보를 획득하기 위해 연구되어 개발되었으며, 현재 과학, 의료, 금융 등 여러 방면에서 이에 대한 새로운 기법과 적용사례에 대한 연구가 활발히 진행되고 있다.

또한 이러한 데이터마이닝의 기법들은 웹에서 정보를 탐색하고 구성하는데 이용할 수 있다. 즉, 웹으로부터 유용한 정보를 발견하고 분석하는 것을 총칭하여 웹마이닝이라 하며, 웹마이닝은 크게 사이트에 관련한 데이터만을 분석함으로써 유용한 정보를 얻어내는 컨텐츠마이닝, 웹서버에 접속하는 사람들의 접속 패턴을 발견하고 분석하는 것을 뜻하는 사용마이닝으로 나눠진다. 본 연구에서는 이중에서 사용마이닝의 관점에서 웹마이닝을 수행하였다.



[그림 1] 웹마이닝의 분류도

2.3 웹마이닝과 웹사용성의 관계

웹을 방문한 사용자들이 해당 웹의 가치를 인정하는 데는 내용, 심미적인 페이지 디자인, 그리고 전체적인 사이트 설계라는 세 가지 요소를 들 수 있다. 웹은 단순히 페이지들의 집

합이 아니기 때문에 전체적인 구조가 사용자에게 만족감을 주는 요소로서 중요하며, 사용자들은 자신이 직관적으로 구조를 이해하기 어렵다고 판단되면 해당 웹의 탐색을 쉽게 포기한다. 따라서 웹은 사용자의 의도와 행동에 적합하게 구조되어야 하며, 이를 통해 사용편의성을 향상시켜 만족감을 제공해야 한다.

이러한 사용자의 프로파일을 추출하고 이를 이용해 웹을 재구성하기 위하여 많은 기법들이 소개되어져 왔다. 웹마이닝은 이러한 관점에서 실제로 사용자들이 해당 웹을 방문할 때의 기록, 즉 로그파일을 사용하여 사용자의 실제 웹 항행 패턴을 추출할 수 있고 이를 통해 사용자들이 의도와 행동을 고비용의 실험을 통하지 않고도 획득할 수 있다는 점에서 의의가 있다.

2.4 연관규칙 및 순차패턴

연관규칙[1] 분석이란 활동이 이루어진 항목들의 상호 연관성을 찾아내어 미리 정한 지지도[2]와 신뢰도[3]를 바탕으로 연관 규칙을 찾아내는 것이다. 지지도는 전체 사용자중에서 연관규칙에 관련된 페이지들을 방문하는 정도를 나타내고, 신뢰도는 특정 페이지와 페이지 사이의 연관성을 나타내므로, 먼저 최소한 주어진 지지도 이상의 페이지들을 찾아내고 그 다음으로 페이지들 사이의 신뢰도를 측정해야 된다.

1)연관규칙(association rules)은 A⇒B로 표현되며 이것은 사건 A가 일어났을 때 사건 B가 일어난다는 규칙을 뜻한다. 웹마이닝의 예를 들자면, 어떤 접속자가 A를 방문한 다음 사이트 B를 방문하는 경우가 자주 발생할 수 있는데 이를 연관규칙으로 생각할 수 있다.

2)지지도(support): 규칙A⇒B의 지지도는

$$P(A \cap B) = \frac{\text{품목}A\text{와}B\text{를동시에포함하는거래수}}{\text{전체거래수}}$$

3)신뢰도(Confidnce): 규칙A⇒B의 신뢰도

$$P(B \mid A) = \frac{\text{품목}A\text{와}B\text{를동시에포함하는거래수}}{\text{품목}A\text{를포함하는거래수}}$$

순차패턴, 혹은 순차연관규칙은 이러한 연관규칙에 시간의 개념을 첨가하여 시간의 흐름에 따른 페이지들의 상호연관성을 탐색하는 것이다. 이를 연속(sequence)라 하며, 순차패턴의 탐사는 사용자가 정의한 최소 지지도 이상의 지지도를 갖는 연속인 빈발 연속(large sequence)을 추출하고 이들 가운데 최대 연속(max sequence)을 찾는 것이다.

3. 연구방법

3.1 로그데이터 수집

육군대학 도서관의 웹서버에 기록된 1999년 8월 11일부터 2001년 6월 30일 까지의 로그파일을 사용하였으며, Netscape Enterprise 웹서버에서 지원하는 CLF(Common Log Format) 형식의 텍스트파일을 수집하였다. 파일의 크기는 약 700MB이며, 7,285,100개의 로그가 기록되었다. [그림 2]와 같이 로그파일의 원본데이터에는 사용자의 IP주소, 요청한 날짜와 시간, 요청한 방법, 요청한 URL주소, HTTP 버전, 상태코드, 전송된 바이트 등이 기록되어 있다.

```
26.144.4.20 -- [31/Dec/1999:11:15:11 + 0900]
"GET/opac/body_simple.html HTTP/1.0" 200
7625
26.144.4.21 -- [31/Dec/1999:11:15:39 + 0900]
"GET/opac/img/sagi-2.gif HTTP/1.0" 304 2539
26.144.4.21 -- [31/Dec/1999:11:15:37 + 0900]
"GET/cgi-bin/MaestroCgi?func=book&bib
_no=16001&sid=808&flag=0&rang=30&iSearch=1
&Srch=1HTTP/1.0" 200-
```

[그림 2] 로그파일의 원본

해당 웹사이트는 크게 도서관소개, 이용안내, 자료검색, 공지사항, 게시판 등의 순으로, 각각의 카테고리 내에 다수의 항목으로 구성되어 있다.

3.2 데이터 전처리

수집된 텍스트형태의 로그파일을 데이터베이스화하여, 불필요한 데이터를 삭제하고 분석에 용이한 파일로 전환하는 과정을 수행하였다. 로그파일에는 html, cgi, asp, jsp등 실제 정보가 담겨있는 페이지뿐만 아니라 해당 페이지에 속한 이미지파일, 동영상파일들이 함께 기록된다. 이러한 이미지, 동영상 데이터와 하나의 로그 기록 내에서 요청한 방법, HTTP 버전, 상태코드, 전송된 바이트 등, 분석에 과잉 불필요한 데이터를 삭제하였다. 또한 URL을 아래와 같이 정수형 코드로 매핑(mapping)하여 분석을 용이하게 하였다.

매핑코드	URL	사이트 이름
33	/library/intro/intro.html	도서관소개
41	/library/service/utilize.html	이용안내
6	/opac/menu_keyword.html	키워드검색

[그림 3] URL의 정수형 코드 매핑

다음으로 데이터마이닝 프로그램인 SAS Enterprise-Miner에서 import 가능한 파일형식 중 하나인 CSV(Comma Separated Values)파일로 변환하는 작업을 수행하였다.

사용한 장비는 Microsoft windows 98을 OS로, PantiumIII 500MHz CPU, 128MB RAM을 장착한 PC이며, 프로그램은 Visual Basic 6.0을 사용하여 데이터 전처리와 변환 작업을 수행하였다.

최종적으로 400,406개의 레코드가 정리되었으며, 이는 Num(레코드 일련번호), IP, 일시분, URL 매핑코드의 순으로 [그림 3]과 같이 정렬되었다.

Num	IP	일시분	URL
51	26.144.4.19	11/Aug/1999:14:01:15	33
52	26.144.4.19	11/Aug/1999:14:01:17	41
53	26.144.4.19	11/Aug/1999:14:01:32	6

[그림 4] 전처리 후의 데이터

3.3 패턴 탐사

전처리가 끝난 로그 데이터는 데이터마이닝 분석 툴을 사용하여 일단 접속자를 구별하고 접속 유지시간을 설정한 후 접속자별로 이동경로를 분석해 낸 뒤에 연관성 규칙 분석을 시도하였다. 사용한 분석 프로그램은 SAS 6.12와 데이터마이닝 프로그램인 SAS Enterprise Miner(version 3.0)이며, Enterprise Miner의 자료추출(sampling)노드와 탐색(exploring)노드를 사용하여 연관성 규칙을 조사하였다.

3.4 패턴 분석

SAS Enterprise Miner에서 분석한 결과를 가지고 페이지별 접속회수와, 순차연관규칙을 도출하였다. 이렇게 도출된 규칙들 중 지지도 5% 이상, 신뢰도 10% 이상인 규칙들을 수집하여 분석하였다.

4. 결과 및 토의

4.1 결과분석

페이지별 접속회수를 보면, 도서관 첫 화면이 가장 많은 접속을 나타냈고, 소장자료검색의 단순검색, 자유게시판, 상세검색, 도서관 이용안내, 분류검색, 신착도서열람 등의 순으로 접속회수가 나타났다. 이는 도서관 홈페이지의 특성상 소장도서 검색을 위해 OPAC을 많이 사용함으로서 나타나는 현상으로 해석할 수 있다. 순차패턴에 관한 규칙의 예는 아래 [표 1]과 같다.

지지도	신뢰도	규칙 분석
49.00	60.99	메인화면>단순검색
19.58	24.38	메인화면>자유게시판
16.16	26.99	단순검색>상세검색
13.87	17.26	메인화면>상세검색
11.06	35.07	메인화면>도서관 이용안내
9.31	11.59	메인화면>분류검색
9.17	11.41	메인화면>신착도서열람
6.71	21.26	메인화면>공지사항
6.22	19.71	메인화면>도서관 소개

[표 1] 순차패턴의 예(지지도, 신뢰도의 단위:%)

[표 1]에서, 첫 번째 규칙을 예로 들면, 사용자의 49%가 '메인화면'과 '소장자료검색의 단순검색'을 동시에 이용하고, 60.99%가 '메인화면'에서 '소장자료검색의 단순검색'으로 이동한다. 세 번째 규칙에서는 '단순검색'으로 이동한 사용자들의 26.99%가 '상세검색'으로 이동함을 보여준다. 이것은 소장자료검색 프레임에서 '단순검색'이 제일 상단 왼쪽에 디폴트로 지정됨으로서 다른 검색에 비해 신뢰도가 높게 나타남으로 해석할 수 있다.

[표 1]과 그 외에 도출된 규칙들 중 의미 있는 결과들을 종합하면 다음과 같다.

- 도서관 홈페이지를 방문하는 사람들의 주된 목적은 OPAC을 통해 자료 검색을 하기 위함이다.
- 카테고리별로 분석하면 '소장자료검색', '자유게시판', '이용안내'의 순으로 접속회수가 높다.
- '소장자료검색' 카테고리 내에서는 '단순검색', '상세검색', '분류검색', '신착도서열람'의 순으로 이동한다.
- 접속 회수가 높은 '단순검색' 또는 '자유게시판'등은 1-2회의 링크를 통해야 접속이 가능하다.

4.2 개선방안

도출된 결과들을 해당 홈페이지를 새로 제작하거나 개선함에 있어서 다음과 같은 사항들이 적용되어야 할 것이다.

- 메인화면에 있는 카테고리를 실제 사용자들이 빈번히 접속하는 '소장자료검색', '게시판'의 순으로 재배열함으로써 사용편의성을 향상시킬 수 있을 것이다.
- 현재 '소장자료검색'을 통하여 접속해야 하는 '단순검색'을 메인화면에서 링크하여 다른 검색들과 함께 한번에 접속하게 함으로써 사용편의성을 향상시킬 수 있다.
- '소장자료검색' 카테고리에서 실제 사용빈도가 높은 '단순검색', '상세검색' 외의 나머지 검색방법(일치검색, 인식번호검색)은 거의 사용되지 않고 있다. 따라서 닐슨(Jacob Nielsen)의 제안대로 복잡한 검색을 단순화하는 것이 이용자들의 사용 편의성을 향상시킬 것이다.

5. 결론

본 연구에서는 육군대학 도서관의 홈페이지의 로그파일을 사용하여 웹마이닝 기법을 적용함으로써 해당 사이트를 평가하고 개선하여 사용편의성을 향상시킬 수 있는 방법을 제안하였다.

유용한 연관성 규칙으로 판정된 규칙들은 효율적인 웹사이트의 디자인에 관한 전략을 제안하는데 사용될 수 있으며, 또한 집중적으로 이용되는 사이트와 그와 연관성이 높은 사이트의 효율적인 배치와 관리에 대한 근거를 제공할 수 있다.

웹사이트의 사용편의성을 향상시키기 위해서는 본 연구에서 제시된 방법 이외에 여러 가지 방법들을 함께 사용하여 최적의 사용편의성을 제공할 수 있도록 해야하며, 타 방법들에 비해 저비용으로 웹사이트의 사용편의성을 평가할 수 있다는 점에서 웹마이닝의 의의가 있다고 하겠다.

참고문헌

1. 최종후 외, 1999, SAS Enterprise Miner를 이용한 데이터마이닝, 자유아카데미.
2. 강현철, 박태원, 임난희, 1998, "Data Mining 방법론과 SAS Enterprise Miner", 한국분류학회 발표논문집.
3 한상태, 강창완, 강현철, 1996, SAS 윈도우즈의 길잡이. 자유아카데미. .
4. Jacob Nielsen, 2000, Designing Web Usability, 1st edition, New Riders Publishing.
5. Spiliopoulou, M., 2000, Web Usage Mining

for Web Site Evaluation, Communication of the ACM, Vol. 43, No 8, 127-134.
6. Srivastava J., Cooly, R., Mukund Deshpande, M., Tan, P. N., 2000, Web Usage Mining: Discovery and Applications of Usage Pattern from Web Data, ACM SIGKDD, Vol. 1, No 2, 12-23.
7. Jacob Nielsen,1999, Usability Engineering., 1st edition, Morgan Kaufmann
8. Osterbauer, C., Kohle, M., and Grechening, T., 2000, Web Usability Testing: A case study of usability testing of chosen sites
9. Web Mining : Information and Pattern discovery on the World Wide Web
http://maya.cs.depaul.edu/~mobasher/webminer/survey/survey.html
10. Borges, J. and Levene, M. 1999, "Data Mining of user navigation patterns", WebKDD'99.
11. Spiliopoulou, M., Pohle, C. and Faulstich, L. C. 1999, "Improving the effectiveness of a web site with web usage mining", WebKDD'99.
12. SAS Institute Inc. 1998. Data Mining Primer.
13. SAS Institute Inc. 1998. Enterprise Miner Software: Applying Data Mining Techniques.
14. Jacob Nielsen 2001, Design Guidlines for Search.
http://www.nngroup.com/reports/ecommerce/Search.html