# 페이지 소요 시간을 고려한 웹 액세스 패턴 마이닝

성현정, 용환승
이화여자대학교 컴퓨터 학과
muse@ewha.ac.kr, hsyong@ewha.ac.kr

# Web Access Pattern Mining considering Page Visiting Duration Time

Hyun-Jung Sung, Hwan-Seung Yong
Dept. of Computer Science and Engineering, Ewha Womans University

요 약

웹로그 마이닝은 대용량의 웹로그 데이터로부터 웹액세스 패턴을 추출함으로써 사용자의 행동 패턴을 찾아내는데 이러한 작업은 웹사이트 설계상의 문제점 등을 발견 및 보완하거나 사용자에게 개인화 페이지를 제공하는데 이용될 수 있다. 사용자의 관심도를 반영하는 웹액세스 패턴을 추출할 때 페이지의 액세스 횟수 뿐만 아니라 페이지의 소요 시간까지 고려함으로써 더욱 정확한 액세스 패턴을 추출하는 것이 본 논문의 목적이다.

## 1. 서 론

웹로그 마이닝은 크게 일반적인 데이터마이닝 테크닉을 사용하는 방법과 로그 데이터에 직접적인 프로세싱을 수행하여 의미있는 결과를 얻어내는 방법으로 나뉠 수 있다. 웹로그 데이터로부터 사용자의 웹 항해(액세스) 패턴을 찾아내는 작업은 후자의 영역에 속할 수 있는데 이러한 액세스 패턴으로부터 우리는 사용자의 행동 패턴을 찾아내거나 앞으로 예상되는 사용자의 경로를 예측할 수 있다. 이러한 작업은 웹사이트 전반에 대한 액세스 패턴을 일반화 시킴으로써 웹사이트의 성격을 알 수 있고, 웹사이트 설계 상의 문제점 등을 발견 및 보완하거나 사용자에게 개인화 페이지를 제공하는데 이용될 수 있다.

웹 액세스 패턴 마이닝에 대한 기존의 연구들에서는 사용자가 많이 방문한 페이지 일수록 즉 액세스 횟수가 높은 페이지 일수록 패턴에 포함될 확률이 높다. 본 논문에서는 페이지의 액세스 횟수 뿐만 아니라 웹사이트의 방문자가 현재 페이지를 액세스하는 시각부터 그 페이지를 떠나기 전까지의 시간 즉, 페이지를 얼마나 보고 있었는가 하는 페이지 소요 시간을 함께 고려함으로써 사용자의 관심도를 정확히 반영할 수 있는 웹 액세스 패턴 추출 방법을 제안하고자 한다.

## 2. 관련 연구 및 연구 배경

웹로그 데이터로부터 사용자의 액세스 패턴을 추출하는 기존의 연구들로는 다음과 같은 연구들이 있다. 논문 [1]에서는 패턴 추출시 사용자가 패턴의 개수나 길이를 잘 컨트롤 할 수 있도록 한 점에 초점을 맞추고 있다. 논문[2]에서는 사용자가 MINT라는 쿼리 언어를 통해 룰 생성 시 관심이 있는 룰 만을 볼 수 있도록 하고 있다. 논문 [3]의 경우 각 페이지 마다 헤더 테이블이 있어, 해당 페이지를 suffix로 가지고 있는 액세스 경로들을 쉽게 찾을 수 있도록 한 자료구조를 사용하고 있다.

위의 연구들은 공통적으로 웹 액세스 경로를 페이지 ID와 액세스 횟수 정보를 갖는 노드들로 이루어진 그래프 혹은 트리로 표현한 뒤, 액세스 횟수를 기준으로 그래프(혹은 트리)를 탐색하면서 패턴을 추출하는 방법을 사용하고 있다. 즉, 빈번히 방문되는 페이지일수록 사용자가 관심 있어 하는 페이지라는 것을 가정한 후에 패턴을 추출하게 되는 것이다. 하지만 이러한 방법은 다음과 같은 경우 사용자의 관심도를 정확히 반영한 패턴을 추출하지 못하게 된다.

첫번째는 웹 설계상의 문제점 때문에 페이지 액세스 횟수로는 정확한 사용자의 패턴을 추출하기 어려운 경우가 존재할 수 있다는 점이다. 예를 들어, 사용자는 d 페이지에 관심이 있지만, 웹 설계상 a,b,c 페이지를 반드시 거쳐야 d에 도달할 수 있도록 웹사이트가 설계되어 있는 경우 사용자들은 d 페이지를 찾기 위해 a,b,c 페이지를 방문하게 되고 상대적으로 이들의 액세스 횟수가 d 페이지의 액세스 횟수 보다 높아 d 페이지는 패턴에서 제외될 확률이 높아지는 경우가 그것이다.

둘째, 기존의 방법들이 택하고 있는 페이지당 액세스 횟수라는 패턴 추출의 기준이 사용자의 페이지 이해력 정도나 페이지 액세스 능력, 네트워크 환경의 차이점 등과 같은 요소들을 배제하고 있다는 점이다. 실제로 논문 [4]에서는 페이지 소요 시간이 사용자의 관심도를 반영하는 좋은 측정치라는 것을 보여주고 있다.

따라서, 본 연구에서는 페이지 소요 시간이라는 요소를 패턴 추출 기준으로 추가함으로써 사용자의 관심도를 반영한 보다 정확하게 패턴을 추출할 수 있는 방법을 제시하고자 한

다.

## 3. 페이지 소요 시간을 고려한 웹 액세스 패턴 마이닝

그림1은 페이지 액세스 횟수와 소요 시간에 의해 페이지의 속성을 분류해 놓은 그림이다. X축을 액세스 횟수 Y축을 페이지 소요시간이라고 가정하였을 때, 각 페이지는 다음의 4분면 중 하나에 포함되게 된다. 1에 속하는 페이지들은 소요시간도 높고 액세스 횟수 역시 높기 때문에 사용자가 관심이 있는 페이지이다. 3에 속하는 페이지들은 소요시간, 액세스 횟수 둘 다 낮기 때문에 사용자가 관심을 가지고 있지 않은 페이지로 볼 수 있다. 또, 2에 속하는 페이지들은 기존의 방법을 사용하였을 경우 패턴에서 제외되지만, 새로운 방법에 의해 새로운 패턴으로 포함해야 할 페이지들이다.

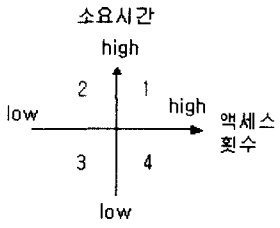


그림 1. 페이지 속성 분류

### 3.1 가정 및 데이터 전처리 과정

다음은 테스트를 위해 가정한 페이지의 속성 및 페이지간 참조 관계이다. 그림2는 각 페이지의 속성에 대한 그림이다. 총 8개의 페이지가 있고, 각 페이지는 앞에서 설명한 사분면의 각각의 위치에 포함된다고 가정한다.

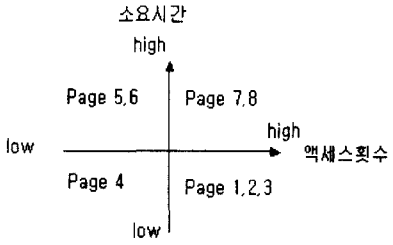


그림 2. 각 페이지의 속성에 대한 가정

그림3은 웹사이트에서 각 페이지가 어떠한 참조 관계를 나타내는 가를 보여주는 그림이다. 페이지0은 루트페이지를 의미하고, 예를 들어 페이지2는 페이지 1을 통해 올 수 있고, 2를 통해 페이지 3혹은 4로 갈 수 있음을 의미한다. 페이지 5,6의 경우 페이지 1,2,3이 사이클 구조의 설계를 가지기 때문에 상대적으로 액세스 횟수가 이들 보다 낮아질 수 있음을 예측할 수 있다.

그림4는 그림3의 페이지간 참조 관계를 가정하고 생성한 로그 데이터이다. 각각의 로그엔트리는 호스트명, 액세스한 시각, 페이지 ID로 이루어져 있고 데이터 전처리 과정을 통해서 각 로그 엔트리간의 시간차를 계산하고 이를 해당 페이지에서 머문 시간 즉, 소요 시간으로 보았다. 그림4에서는 duration이 그것을 의미한다. 또한 세션 ID를 부여하는데 본 논문에서는 IP주소로써 웹사이트 방문자를 구분하였으며, 30분 이상 페이지 요청이 없으면 해당 세션이 끝난 것으로 하였다. 각 페이지별 액세스 횟수와 소요시간, 평균 소요 시간 역시 데이터 전처리 과정에서 측정하게 된다.

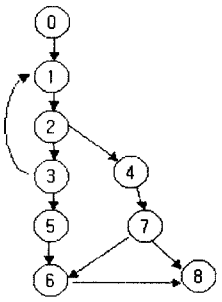


그림 3. 페이지간 참조 관계

| | | | | | |
|---|---|---|---|---|---|
| 1 | 203.255.178.15 | 00-08-14 오후 2:00:11 | 1 | 5 | 1 |
| 2 | 203.255.178.15 | 00-08-14 오후 2:00:16 | 2 | 3 | 1 |
| 3 | 203.255.178.15 | 00-08-14 오후 2:00:19 | 3 | 9 | 1 |
| 4 | 203.255.178.15 | 00-08-14 오후 2:00:28 | 1 | 1 | 1 |
| 5 | 203.255.178.15 | 00-08-14 오후 2:00:29 | 2 | 2 | 1 |
| 6 | 203.255.178.15 | 00-08-14 오후 2:00:31 | 4 | 8 | 1 |
| 7 | 203.255.178.15 | 00-08-14 오후 2:00:39 | 7 | 49 | 1 |
| 8 | 203.255.178.15 | 00-08-14 오후 2:01:28 | 8 | 80 | 1 |
| 9 | 203.255.178.15 | 00-08-14 오후 2:02:48 | 1 | -1 | 1 |
| 10 | 203.255.178.15 | 00-08-15 오후 9:00:00 | 2 | 3 | 2 |
| 11 | 203.255.178.15 | 00-08-15 오후 9:00:03 | 3 | 9 | 2 |
| 12 | 203.255.178.15 | 00-08-15 오후 9:00:12 | 5 | 90 | 2 |
| 13 | 203.255.178.15 | 00-08-15 오후 9:01:42 | 6 | -1 | 2 |

그림 4. 예제 로그 파일

### 3.2 그래프 생성 및 패턴 추출

액세스 경로를 표현한 그래프는 연결리스트를 이용하였고 패턴추출은 그래프의 깊이 우선 탐색 방법을 변형하여 사용하였다. 그래프 탐색 시 각 노드를 방문하였는가의 조건 외에 액세스횟수나 페이지 소요시간과 같은 추가 조건을 가지고 탐색해 나가되 이러한 조건을 만족하지 않으면 더 이상 그 이후로는 탐색해 보지 않는 방법을 사용하였다. 본 논문에서는 액세스 횟수만을 고려하여 패턴을 찾는 기존의 방법과 비교하기 위해서 다음의 두 가지 경우를 모두 고려해본다.

- 액세스 횟수만을 고려
- 액세스 횟수 및 페이지 소요 시간 고려

이때, 후자의 페이지 소요 시간까지 고려하는 방법의 경우는 다시 두 가지로 나뉘는 데, 첫번째는 사용자가 특정 시간을 제시하고 이 시간 이상을 소요한 페이지를 패턴으로 추출하는 방법이다. 두 번째는 각각의 로그엔트리를 스캔하면서 페이지 소요 시간이 해당 페이지의 평균 소요시간 이상을 본 페이지일 경우 이를 패턴으로 취하는 방법이다. 페이지마다 내용, 길이에 따른 난이도가 다르기 때문에 모두 동일한 시간 제약 조건을 적용하는 것이 문제점이 있을 수 있기 때문에 본 논문에서는 후자의 방법으로 테스트를 수행해 보았다.

### 4. 패턴의 가시화

다음은 추출된 패턴을 가시화 시키는 과정이다. 그래프의 노드는 웹사이트의 페이지를 표현하고, 노드의 색은 명도로써 그 값이 크고 작음을 의미하도록 하였다. 각 페이지의 액세스 횟수에 따라 페이지를 분류하여 총 5등급으로 액세스 횟수의 크기 정도를 표현하였다. 그림9는 생성된 그래프의 모든 노드를 탐색했을 경우 즉, 모든 세션을 가시화 한 그림이다.

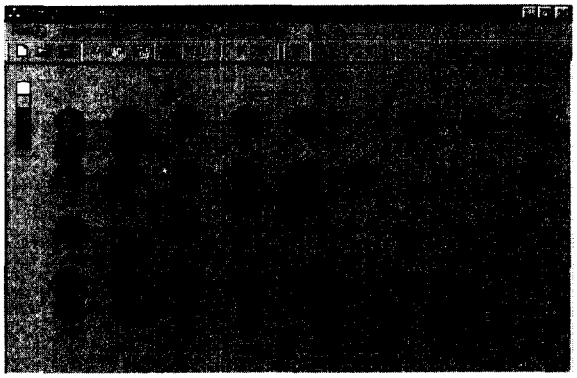

그림 9. 세션 가시화

그림10은 support만을 고려한 즉, 페이지의 액세스 횟수만을 고려하여 패턴을 추출하고 이를 가시화 한 그림이다. 여기서 support란 총 페이지 액세스 횟수에 대하여 각 페이지가 몇 퍼센트의 확률을 가지고 액세스 되는가에 대한 수치이다.

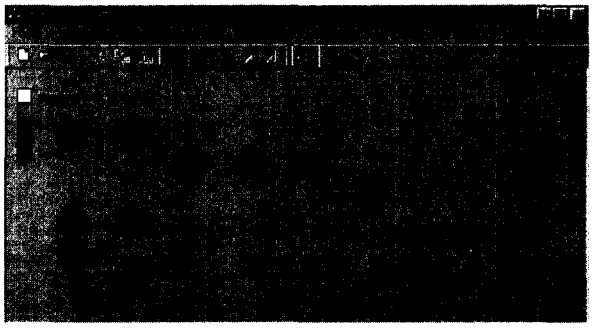

그림 10. 페이지 액세스 횟수만을 고려하여 패턴 추출

그림9에서 어두운 색을 띈 노드들, 즉 액세스 횟수가 상대적으로 높은 페이지 1,2,3,7,8 들만이 패턴으로 추출됨을 볼 수 있다. 그림11은 support 뿐만 아니라 페이지 소요 시간까지 고려하여 패턴을 추출한 뒤 이를 가시화 한 그림이다. 그림10에서 볼 수 없었던 페이지 5,6 등이 패턴으로 포함되었음을 알 수 있다. 즉, 액세스 횟수는 상대적으로 낮지만 페이지를 본 시간이 높기 때문에 사용자가 이 페이지에 관심이 있음을 보여준다.

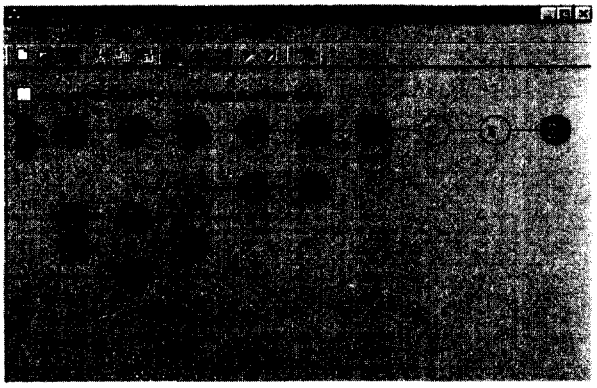

그림 11. 페이지 액세스 횟수
및 소요 시간을 고려하여 패턴 추출

### 5. 결론 및 향후 과제

본 논문에서는 기존의 연구들이 따르고 있는 페이지의 액세스 횟수만을 가지고 패턴을 추출하는 방법이 가지는 문제점을 제시하고 이를 해결할 수 있는 방법으로 페이지의 소요 시간을 고려하는 방법을 제안하였다. 또한, 각 페이지를 나타내는 노드의 색을 달리 표현함으로써 사용자가 각 페이지의 특성을 이해하기 쉽도록 하였다. 위의 경우에는 액세스 횟수를 기준으로 표현하였는데 페이지의 소요시간을 기준으로도 표현 가능하며 이것은 사용자가 관심 있어 하는 즉, 오랜 시간을 두고 본 페이지를 이해하는데 도움을 줄 수 있을 것이다. 본 논문에서는 페이지의 소요 시간만을 추가하였는데 이외에도 사용자가 현재 보고있는 페이지를 스크롤 하는 정도, 윈도우의 크기 조절 여부 등과 같은 요소를 더 추가할 수 있다면 보다 정확히 사용자의 관심도를 반영한 웹 액세스 패턴을 추출할 수 있을 것이다.

### 참고 문헌

[1] J.Borges and M.Levene, "A fine grained heuristic to capture web navigation pattern", SIGKDD , Explorations 2,(2000),40-50

[2] Myra Spiliopoulou, Lukas Faulstich, C., and Karsten Winkler, "A Data Miner analyzing the Navigational Behaviour of Web Users", In Proc. of the Workshop on Machine Learning in User Modelling of the ACAI'99 Int. Conf., Creta, Greece, July 1999

[3] Jian Pei, Jiawei Han, Behzad Mortazavi-asl, and Hua Zhu, "Mining access patterns efficiently from web log", 2000 Pacific-Asia Conf. on Knowledge Discovery and Adata Mining

[4] Mark Claypool, Phong Le, Makoto Wased and David Brown, "Implicit Interest Indicators", IUI'01, January