

# 사용자 프로파일을 이용한 웹페이지 추천

강귀영<sup>U</sup>                      조동섭  
이화여자대학교 컴퓨터학과  
{Ryderbay, dscho}@ewha.ac.kr

## Web Page Recommendation Using User Profile

Kwi-Young Kang<sup>U</sup>                      Dong-Sub Cho  
Dept. of Computer Science and Engineering, Ewha Womans University

### 요 약

전자상거래 사이트 내에서 제공되는 정보가 많아질수록 사용자는 많은 실패를 거친 후 자신이 원하는 정보에 도달하게 된다. 사용자가 어떤 사이트에 자주 찾아오도록 하기 위해서는 적은 노력으로도 원하는 정보에 도달할 수 있도록 도움을 주는 웹 페이지 추천 기법이 필요하다. 이 기법은 사용자 프로파일의 패턴을 분석한 후 분석된 결과를 바탕으로 사용자에게 현재 있는 페이지와 가장 연관성이 높다고 판단되는 웹 페이지를 연관규칙을 응용한 방법을 이용하여 추천한다.

### 1. 서론

인터넷 상에서 일련의 서비스를 제공하는 사이트는 많은 사용자와 함께 많은 정보가 등록되며 빠르게 성장하고 있다. 각 사이트들은 기본적인 정보만 제공하던 초기 형태에서 벗어나 점차 많은 정보와 다양한 볼거리를 제공하면서 구조가 점차 복잡해지게 되었다. 한 사이트 내에서 제공되는 정보가 많아지면 그만큼 이용할 수 있는 정보도 많아지게 되지만, 또한 단편적인 정보를 효율적으로 찾기도 어려워진다.

사용자가 어떤 사이트에 자주 찾아오도록 하기 위해서는 적은 노력으로도 원하는 정보에 도달할 수 있어야 한다. 사용자가 특정 웹 페이지에서 어디로 가야할지 결정하지 못할 때, 사이트 내부에서 연관성이 있다고 판단되는 웹 페이지를 사용자에게 추천할 수 있다면 연관성 없는 웹 페이지까지 탐색하는 경우를 줄일 수 있을 것이다. 이에 따라 사용자는 탐색 실패 횟수를 줄일 수 있을 것이고, 또한 원하는 정보에 더 빨리 도달할 수 있을 것이다.

본 논문에서는 웹 페이지 추천 기법으로 연관규칙(association rule)을 응용한 방법을 제시할 것

이다. 데이터베이스로는 사용자가 접속하여 종료할 때까지의 시간 순으로 정렬된 페이지 뷰 즉, 사용자들의 경로정보를 이용한다.

### 2. 연관규칙

T를 현재까지 발생한 모든 트랜잭션이라 하고, I를 m개의 상품들 또는 구매자들이라 하면 T와 I는 다음과 같이 표현된다.

$$T = \{t_1, \dots, t_n\}$$

$$I = \{i_1, i_2, \dots, i_m\}$$

어떤 트랜잭션 t에 상품 또는 구매자 w가 포함되어 있을 때, t(w)는 다음과 같이 표현된다.

$$t(w) = 1$$

w는 트랜잭션 t에 존재하는 상품 또는 구매자 중에 하나이다. 또한, W가 상품들의 집합이라고 했을 때, t(W)는 트랜잭션 t에 대해 W에 있는 모든 상품 또는 구매자가 존재함을 말한다.

$$t(W) = 1 : t(w) = 1, \text{ 모든 } w \in W$$

W : I에 속하는 상품들 또는 구매자들의 부분집합

$$(X) = \{i \mid t_i(X)=1\}$$

(X)는 X가 포함하는 모든 상품 또는 구매자들을 포함하고 있는 모든 트랜잭션의 집합이 된다. 여기서, X : |(X)| ≥ σ 어떤 X를 σ-covering이라 부른다. 따라서, 모든 트랜잭션에 대해 다음과 같은 법칙을 발견

할 수 있다.

$W \Rightarrow B : T$ 에 대해  $W \subseteq R$  이고  $B \subseteq R \setminus W$ 를 만족하는 연관 규칙

$W$  : 연관 규칙의 왼쪽 집합

$B$  : 연관 규칙의 오른쪽 집합

$W \Rightarrow B$ 는 지지도(support threshold)와 신뢰도(confidence threshold)에 의해 발견된다. 지지도는 이 법칙이 적용되는 트랜잭션의 수를 나타내고, 신뢰도는 연관 규칙의 왼쪽 집합이 가리키는 전체 트랜잭션 개수 중에서 이 법칙을 만족하고 있는 트랜잭션의 개수를 나타낸다. 따라서, 지지도와 신뢰도에 의해 다음의 연관 규칙이 발견된다.

$T$ 는  $W \Rightarrow B$ 를 만족한다.

단,  $WUB : \sigma$ -covering

(i.e.,  $|(WUB)| \geq \sigma$ )  $\sigma$  : 지지도

$|KWUB|/|W| \geq \gamma$   $\gamma$  : 신뢰도

연관 규칙 발견은 간단히 2가지 알고리즘에 의해 쉽게 발견될 수 있다. 하지만, 이들 알고리즘은 모든 아이템에 대해 가능한 부분집합들을 구해야 하므로 NP-hard로 취급된다. 따라서, 성능을 개선하기 위해 병렬 알고리즘과 같은 다양한 방법에 대한 연구들이 활발히 진행되고 있다[5].

본 논문에서 참고하는 Apriori 알고리즘을 아래의 그림 1에서 설명한다.

우선 각 반복에서 라지 아이템셋(large itemset)을 위한 후보자 집합을 구성한다. 이 후보자 집합에서 각 후보자 아이터셋이 일어난 횟수를 센다. 최소 지지도를 기반으로 라지 아이터셋을 결정한다. 후보자 집합의 원소의 개수가 1보다 작을 때까지 반복한다. 이때, 첫 반복단계의 후보자 집합은 모든 트랜잭션을 스캔하여 구한다. 또한, 후보자 집합을 구할 때 라지 아이터셋의 어떠한 부분집합도 최소 지지도를 가져야 하므로  $C_2$ 를 구할 때  $L_1 * L_1$ (concatenation:  $|L_1|C_2$ 개)를 사용한다[6].

### 3. 연관규칙을 응용한 웹 페이지 추천 기법

이 기법의 기본적인 개념은 현재 사용자가 지나온 경로를 되짚어 가면서, 현재 사용자가 지나온 페이지를 인접한 순서대로 가장 많이 지나온 트랜잭션(transaction)을 찾아 그들 중 Count가 가장 큰 다음 페이지 Next Page를 추천하는 것이다.

그림 2를 보면, 현재 사용자는 C 페이지에 있고 이 페이지에 오기 전에 B와 A를 역순으로 거쳤다. 초기 데이터베이스 DB는 기존 사용자들의 로그인부터 로그아웃까지의 경로정보를 하나의 트랜잭션으로 묶어 모은 것이다. Count는 Next Page가 나온 횟수이다.

초기 데이터베이스 DB에서 현재 페이지 C가 있는 항목들을 선택하여 데이터베이스  $D_0$ 를 구성한다. 동시에 선택된 각 트랜잭션 내에서 C의 Next Page를 찾아 Count를 계산한  $P_0$ 를 만든다. 다음

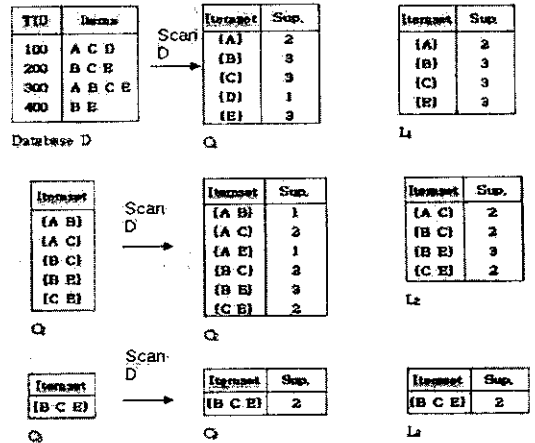


그림 1 Apriori 알고리즘 적용 예

반복에서는,  $D_0$ 를 스캔하여 현재 페이지 C의 바로 이전에 거쳤던 페이지 B가 C의 이전에 있는 트랜잭션들을 선택하여  $D_1$ 를 구성한다. 동시에 선택된 각 트랜잭션 내에서 C의 Next Page를 찾아 Count를 계산한  $P_1$ 를 구성한다. 다음 반복에서도 이전 반복과 마찬가지로,  $D_1$ 를 스캔하여 현재 페이지 C의 두 번째 이전 페이지인 페이지 A가 C의 이전에 있는 트랜잭션들을 선택하여  $D_2$ 를 구성한다. 동시에 선택된 각 트랜잭션 내에서 C의 Next Page를 찾아 Count를 계산한  $P_2$ 를 구성한다. 더 이상 탐색할 페이지가 없으므로 반복을 중지한다. 탐색할 페이지가 더 있더라도 데이터베이스  $D_k$ 가 더 이상 생성되지 않는다면 반복을 중지한다.

k를 현재 페이지와 탐색하는 페이지 사이의 거리라고 볼 때 매 반복은  $D_{k-1}$ 을 스캔하여  $D_k$ 와  $P_k$ 를 만드는 과정으로 볼 수 있다.

이렇게 생성된  $P_k$ 들 중에서 가장 큰 k 값을 가진  $P_{(max k)}$ 에 있는 Next Page를 추천한다. Next Page가 여러 개일 경우에는 Count가 가장 큰 페이지를 추천한다. 만일 추천하고자 하는 페이지가 삭제되었을 경우에는 다음 우선순위를 가진 페이지를 추천한다.

### 4. 결론 및 향후 과제

본 논문에서는 한 사이트 내에서 현재 사용자가 지나온 경로정보에 따라 동적으로 관련된 웹 페이지를 계산하여 추천하는 기법을 설명하였다. 기존의 웹 페이지 추천 시스템에 대한 연구는 한 사이트 내에서의 웹 페이지 추천이 아닌 월드 와이드 웹 상에서의 웹 페이지 추천에 대한 연구가 주류를 이루었다. 기존의 시스템들은 기계 학습 방법을(Machine Learning) 이용하여 나온 결과가 사용자

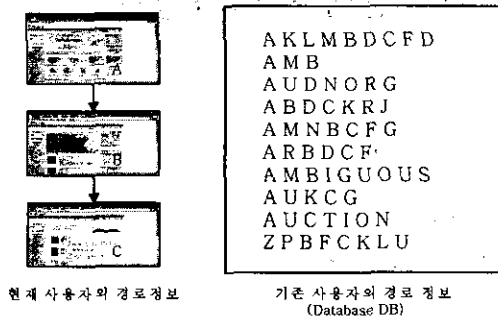


그림 2 사용자 경로정보와 초기 데이터베이스

에게 적합한 것이었는지를 인증 받고 다시 이 결과를 토대로 학습해야 하므로 전자상거래와 같은 한 사이트 내의 웹 페이지에 대해 추천을 할 경우에는 적합하지 않았다. 한 사이트 내의 웹 페이지를 추천할 경우에는 각 사용자가 지나온 경로정보를 기록할 수 있으므로 연관규칙을 응용한 방법을 적용할 수 있었다.

데이터베이스 DB를 전체 사용자의 경로정보로 보지 않고, 사용자를 그룹화 하여 각 그룹의 경로정보로 본다면 현재 사용자에게 좀 더 개인화 된 페이지를 추천할 수 있을 것이다. 또한, 새 페이지, 최근 갱신된 페이지에 대해서도 우선순위를 적용시킬 수 있는 방법을 모색해야 할 것이다.

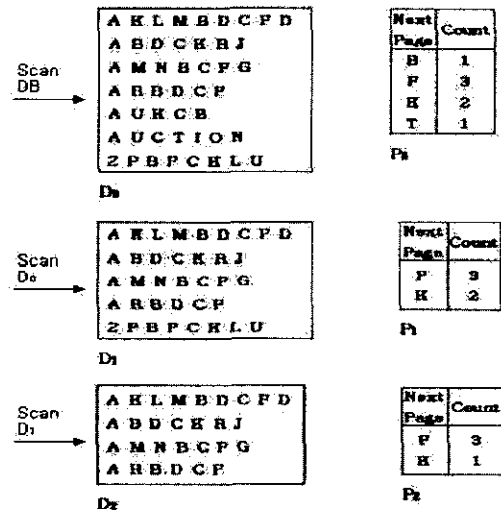


그림 3 추천 페이지 계산 과정

표 1 웹 페이지 추천 알고리즘

```

k=0;
ch=pop(stack); //stack : 현재 사용자의 경로정보 저장
                //ch : 현재 사용자의 현재 페이지
if ch ∈ transaction t of DB then
    { Next Page=ch's next page
      Count 계산 } //Pk 계산
    Save in Dk
end if
while ((Dk)!=0 or stack underflow){
//stack underflow : 현재사용자 경로정보의
// 첫 페이지까지 읽었을 때
    k++;
    ch2=pop(stack);
    if ch2 ∈ transaction t of Dk then
        { Next Page=ch's next page
          Count 계산 } //Pk 계산
        Save in Dk
    end if
end while
    
```

### 5. 참고 문헌

- [1] Gabriela Polcicova, "Recommending HTML-documents using Feaure Guided Automated Collaborative Filtering," ACM SIGIR '99, August 1999.
- [2] Thorsten Joachims, Tom Mitchell, Dayne Freitag, and Robert Armstrong, "WebWatcher: Machine Learning and Hypertext," Fachgruppentreffen Maschinelles Lernen, Dortmund, August 1995.
- [3] Thorsten Joachims, Dayne Freitag, Tom Mitchell, "WebWatcher: A Tour Guide for the World Wide Web," IJCAI97, Aug. 1997.
- [4] Robert Armstrong, Dayne Freitag, Tom Mitchell, and Thorsten Joachims, "Web Watcher: A Learning Apprentice for the World Wide Web," 1995 AAAI Spring Symposium on Information Gathering from Heterogeneous, Distributed Environments, March 1995.
- [5] 문흥기, 이수원, "전자 상거래 에이전트를 위한 연관 규칙 발견 및 확장," 1999년 춘계정보과학회, August 1999.
- [6] Jong Soo Park, Ming Syan Chen and Philip S. Yu, "An Effective Hash-Based Algorithm for Mining Association Rules," 1995 ACM SIGMOD Conference, pages 175--186, San Jose, California, USA, May 1995.
- [7] Sanford Gayle, "The Marriage of Market Basket Analysis to Predictive Modeling," The Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, August 2000.