

## 개인화 된 추천정보 소개를 위한 Web Usage Mining 알고리즘

이은영\*, 곽미라, 염선희, 조동설  
이화여자대학교 컴퓨터학과

### Web Usage Mining Algorithm for Personalized Recommender System

Eunyoung Lee\*, Mira Kwak, Sunhee Youm, Dongsub Cho  
Dept. of Computer Science and Engineering, Ewha Womans University

**Abstract** - 오늘날 인터넷 사용자들은 정보의 흥수 속에 놓여있다. 웹사이트에 들어가면 대부분은 자신과 관련 없는 정보들이 쏟아진다. 따라서 인터넷 사용자들의 관심에 맞는 내용을 제공해주어 시간의 절약과 동시에 사용자에게 가치 있는 정보를 제공할 수 있게 하는 서비스가 필요하다. 이러한 개인화 된 서비스를 제공해주기 위해 사용자에 대한 정확한 분석을 바탕으로 사용자에게 효율적인 서비스를 제공하여야 할 것이다.

따라서 본 논문에서는 사용자 프로파일 및 웹 로그 등을 토대로 각 고객의 성향과 패턴을 정확하게 분석하여, 사용자 각 개인에게 적합하며 효율적인 서비스를 제공해 줄 수 있는 Web Usage Mining 을 통한 사용자 패턴 추출 알고리즘을 개발하고자 한다. 본 논문에서 연구한 Web Usage Mining 알고리즘은 사용자의 웹 사용 습관을 토대로 데이터 마이닝의 과정을 거쳐 사용자의 성향과 관심을 결정하고, 이를 바탕으로 사용자에게 알맞은 내용을 제공할 수 있도록 할 것이다. 이때, 사용자의 정보는 웹 내에서의 행동 중에서 중요하게 사용되는 특정한 페이지를 보는 시간, 웹 서핑 패턴, 전자 상거래 사이트의 경우에는 구매한 상품과 쇼핑 카트에 넣은 상품 등의 관찰된 정보를 기반으로 하며, 개인의 사생활을 침해하지 않는 범위 내에서 이루어지도록 했다.

## 1. 서 론

현재의 인터넷은 무한하고 다양한 자원의 보고이며, 막힘 없이 각지를 연결하는 통로이다. 이제 사용자들은 쏟아지는 인터넷의 정보들 중에서 자신에게 꼭 필요한 정보들을 찾아야한다. 따라서 인터넷 사용자들의 관심에 맞는 내용을 제공해주어 시간의 절약과 동시에 사용자에게 가치 있는 정보를 제공할 수 있게 하는 서비스가 필요하다. 즉, 사용자에 대한 정확한 분석을 바탕으로 사용자에게 효율적인 서비스를 제공해주는 Personalized Recommender System 이 필요하다.

아직까지는 Personalization 방법에 대한 분류 기준조차도 사람에 따라 다르지만 일반적으로 많이 쓰이는 방법들은 규칙기반 필터링(Rules-based filtering), 협업 필터링(Collaborative filtering), 학습 에이전트(Learning agent) 등을 들 수 있다. 각각의 방법들은 실행 방법과 비용에서 차이를 가지고 있다. 또한 일반적으로 웹사이트 개인화에는 한 가지 방법만 사용하는 것이 아니라 위에서 말한 것들 중 두 세 가지 방법을 혼합해서 사용한다. 따라서 각 방법들의 비용과 이로 인해 나타나는 직접/간접적인 효과를 정확하게 판단하고 적절한 방법을 선택하는 것이 매우 중요하다[1].

본 논문에서는 사용자에게 개인화 된 추천정보를 제공해 주기 위해, 우선 사용자 프로파일 및 웹 로그 등을 토대로 각 사용자의 성향과 패턴을 정확하게 분석한다. 그리고 분석된 결과로 사용자 개인에게 적합하며 효율적인 서비스를 제공해줄 수 있는 Web Usage Mining 을

통한 사용자 패턴 추출 알고리즘을 개발하고자 한다. Web Usage Mining 알고리즘은 사용자의 웹 사용 습관을 토대로 데이터 마이닝의 과정을 거쳐 사용자의 성향과 관심을 결정하고, 이를 바탕으로 사용자에게 알맞은 내용을 제공할 수 있도록 할 것이다.

본 논문의 구성은 다음과 같다. 2장에서는 기존의 Personalization 기법들을 설명할 것이며, 3장에서는 본 논문에서 제안하는 Web Usage Mining 알고리즘에 대해 설명하고, 4장에서는 결론과 향후연구과제에 대해 논의 할 것이다.

## 2. 관련 연구

여기서는 다른 개인화 된 정보를 주기 위한 기존의 여러 알고리즘들 중에서 몇 가지 대해 설명하겠다.

### 2.1 Web Usage Mining Algorithm

Bamshad 가 제안한 Web usage를 기반으로 한 개인화 방법은 전체적인 과정이 offline과 online 작업으로 나뉘어진다. offline 과정은 사용자 트랜잭션 파일들을 만드는 데이터 전처리 과정과, 특정 usage mining task 로 구성되어 있다. offline의 첫 번째 단계는 데이터 전처리, 데이터 준비, 포함, 데이터 정제, 필터링, 그리고 트랜잭션 확인의 과정으로 이루어진다. 그리고 두 번째 단계인 usage mining 작업은 association rule과 URL cluster을 이용하여 이루어진다.

마이닝 작업이 완료되고 나면, Frequent ItemSet 과 URL cluster의 결과가 online 과정에서 사용된다. online 과정은 추천 엔진과 HTTP 서버로 구성된다. 웹서버는 사용자의 브라우저가 HTTP 요청을 할 때 사용자 세션을 기억하며 저장하게 된다. 이것은 URL rewrite 또는 웹 서버 접근 로그 파일 등에 의해 얻을 수 있다. 추천 엔진은 추천할 URL을 계산해 내기 위해서 URL cluster 와 발견된 association rule을 결합하게 된다. 그리고 나서 사용자가 요청한 페이지가 클라이언트 브라우저로 전송되기 전에 추천될 URL 집합이 요청된 페이지의 마지막에 덧붙여지게 된다. 이렇게 하여 사용자의 취향에 맞는 개인화 된 추천을 하게된다[2].

### 2.2 Regression-based algorithm

회귀기반 알고리즘은 E-Commerce에서 사용자들 사이의 유사성을 찾는 대신 아이템들의 유사성을 찾아서 추천해 주는 방식으로, 이것은 협업 필터링(Collaborative filtering) 방식의 일종이다. 사용자가 아이템들의 일부에 등급(score)을 매겨주면, 그것과 기준의 사용자들이 등급을 매긴 기록을 기반으로 나머지 아이템들에 대한 등급을 예측하고, 추천해 주는 방식이다. 사용자들이 등급을 매긴 결과로 유사한 성향을 가진 여러 그룹인, expert 들을 미리 모델링 해놓는다. 그 중에서 가장 가능성 있는 최적의 expert 들을 결합하여, 현재 사용자의 성향과

가장 유사한 결과를 찾아 추천하게 된다.

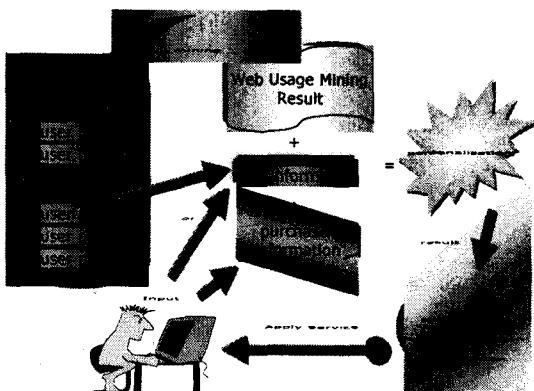
이 방식은 사용자들 사이의 유사성을 찾는 neighbor-based collaborative filtering 방식에 비해 정확도는 현저히 높지만, 속도 면에 있어서는 개선하지 못했다는 단점이 있다[3].

### 3. 본론

2장에서는 기존의 Personalization 기법들을 살펴보았다. 이 장에서는 기존의 Web Usage Mining 방법에서 개선된 Web Usage Mining 알고리즘을 제안하고자 한다.

#### 3.1 시스템 구성

개인화 된 추천정보를 제공해주기 위한 Web Usage Mining의 전체 시스템은 그림1과 같은 구조를 가진다.



[그림 1]. Personalized Recommender System 구성도

개인화 된 추천정보를 사용자에게 제공해 주기 위해서 사이트에서는 기존의 사용자들에 대한 로그 기록들을 잘 활용하여 현재의 사용자가 가장 원하는 것을 추천해 줄 수 있어야 한다. 우선 기존의 사용자들의 기록이 남아있는 웹 로그를 분석하여 전체적인 성향을 알아낸다. 이렇게 나온 결과와 개인의 사용기록 즉, ID 등록한 사용자의 이전 기록을 분석하여 그 사용자의 분석된 성향을 모두 사용하여 추천정보를 주게되면 좀 더 정확한 개인화된 정보가 될 것이다. 또한 필요한 정보를 웹에서 직접 받을 수도 있을 것이다. 쇼핑몰을 예로 들면, 사용자가 현재 어떤 물건을 살펴왔는지, 어떤 용도로 쓸 것인지 등의 상세 정보를 받아서 기존의 결과와 함께 추천정보를 주면 좀 더 정확한 개인화 정보를 생성할 수 있을 것이다. 그러나 ID 등록으로 인한 개인 정보나, 웹에서 직접 받는 정보가 없는 경우도 많이 있을 것이다. 따라서 어떠한 경우에도 최대한 사용자에게 맞는 정보를 제공해 줄 수 있도록 적용적으로 대응 할 수 있는 알고리즘의 구현이 필요하다.

본 논문에서는 전체 Personalized Recommender System의 구성요소들 중에서 Web Usage Mining을 위한 알고리즘을 구현하여 전체 웹 로그를 정확하게 분석할 수 있도록 하는 것이 목적이다.

#### 3.2 Web Usage Mining Algorithm

본 논문에서 제안하는 Web Usage Mining 알고리즘의 기본적인 개념은, ID 등록이나 쿠키를 사용하지 않고 웹 로그에서 사용자의 세션을 정확하게 구분하고 이렇게 구분된 사용자 세션들을 가지고 전체 사용자들의 Frequent web traversal path들, 가장 빈번하게 방문된

페이지들의 그룹들, 가장 빈번하게 나타나는 Itemset들을 찾아내는 것이다. 이렇게 마이닝 하여 나온 결과는 사용자들에게 추천해 주는데 필요하고, 후에 사용자의 개인성향 또는 웹에서 직접 제공받은 정보와 합하여 좀 더 정확한 개인화 된 추천정보를 만들어 낼 수 있을 것이다.

##### 3.2.1 User session identification

웹 로그를 분석하기 위해서는 우선 사용자 세션을 구분하는 것이 가장 우선적인 일이다. 물론 ID등록이나, 쿠키 등을 사용하여 구분할 수 있는 방법도 있지만, 각각의 단점이 있다. 우선 ID 등록으로만 구분하는 경우, 사용자들 중에서 개인의 프라이버시를 침해하고 익명으로 남기를 원해서 개인 정보를 제공하기 원하지 않는 경우가 있기 때문이다. 또한, 쿠키를 사용하는 경우에는 같은 쿠키로 웹서버에 접근할 경우 하나의 세션으로 취급되어 정확하게 세션을 구분할 수 없는 경우가 있다. 그렇다면 웹 로그에 남겨진 IP 주소를 가지고 사용자 세션을 구분해야하는데, 이 때에도 proxy나 방화벽의 사용으로 사용자 세션을 구분하는데 어렵게 하는 문제가 있고, 정보가 cache에 남아있거나, Backward 또는 Forward를 사용하는 경우에 로그에 기록이 남지 않아서 사용자의 기록을 유추해야 하는 경우도 있다[4].

이를 해결하기 위해, 웹 로그에 있는 IP, Time-stamp URL, Referrer 과 Agent로 사용자 세션을 구분하는 알고리즘을 제안한다. 우선 같은 IP를 가진 사용자 Referrer URL과 access URL의 쌍으로 이루어진  $(x \rightarrow y)$  쌍들을 나열해 보자. 만약 다음과 같다면 :  $(- \rightarrow a), (a \rightarrow b), (b \rightarrow c), (c \rightarrow d), (b \rightarrow e), (a \rightarrow f)$

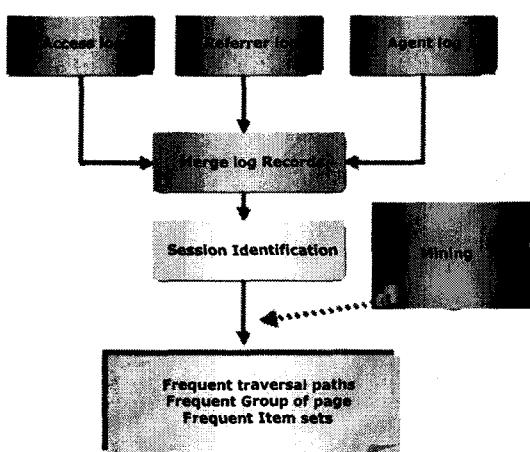
Time-stamp와 Agent 등을 이용해 이 사용자의 Full path를 다음과 같이 유추해 낼 수 있다.

$(- \rightarrow a), (a \rightarrow b), (b \rightarrow c), (c \rightarrow d), (d \rightarrow c), (c \rightarrow b), (b \rightarrow e), (e \rightarrow b), (b \rightarrow a), (a \rightarrow f)$

그러나, backward 또는 forward를 몇 번을 허용할 것인지, 몇 분 정도가 가장 작은 time-stamp의 차이 인지에 대한 표준을 미리 정해두어야 할 것이다.

##### 3.2.2 Discovering Frequent Itemsets

일단, 사용자 세션을 구분하고 나면 이를 바탕으로 Frequent web traversal path들, 가장 빈번하게 방문된 페이지들의 그룹들과 이것을 이용하여 가장 빈번한 Itemset 들을 찾아낸다. 다음의 그림2는 웹 로그에서



[그림 2]. Web Usage Mining 알고리즘의 흐름도

세션을 구분하고, 이를 마이닝하여 결과를 도출해내는 Web Usage Mining 알고리즘의 흐름도를 나타낸다.

### (1) Mining frequent traversal path

가장 빈번한 traversal path 들을 마이닝 하기 위해서 먼저 구분된 사용자의 웹 세션 중에서 가장 maximum forward path 들을 찾아야 한다. 이 때 maximum forward path 는 중간에 끊어지지 않고, 계속 연속된 URL 페이지이어야 한다. maximum forward path를 찾았다면, 그것을 마이닝 하여 가장 빈번하게 나오는 연속적인 subsequence 들을 찾아내는 것이다.

빈번한 연속적인 subsequence 들을 찾아내는 알고리즘은 다음의 표1 과 같다.

```

for each  $F_j$ 
  for each  $(x_1, x_2, \dots, x_m)$  in  $F_j$  {
    if ( $m \geq k$ ) {
      for ( $j=1; j < m-k+1; j++$ ) {
        if ( $(x_j, \dots, x_{j+k-1})$  is already in  $LP_k$ 
          increase its corresponding count;
        else if ((support of  $(x_j, \dots, x_{j+k-2}) \geq s_{k-1}$ )
                  and (support of  $(x_{j+1}, \dots, x_{j+k-1}) \geq s_{k-1}$ )
          insert  $(x_j, \dots, x_{j+k-1})$  into  $LP_k$ 
        )
      }
    }
  }
}

```

[표 1]. Algorithm for finding Large traversal Path set  $LP_k$

(2) Mining groups of pages most frequently visited  
Frequent traversal path 들을 마이닝 하고 나서는, 사용자들이 가장 빈번하게 방문한 페이지들의 그룹을 찾는다. 이 그룹내의 페이지들은 꼭 연결된 페이지이어야 할 필요는 없다. 다음의 표2 는 사용자들이 가장 빈번하게 방문한 페이지들의 그룹을 찾아내는 알고리즘을 나타낸다.

```

Sort the groups in  $LG_{k-1}$  in lexicographical order
for each group  $(x_1, \dots, x_{k-1})$  in  $LG_{k-1}$ 
  for each group  $(y_1, \dots, y_{k-1})$  in  $LG_{k-1}$ 
    such that  $x_2=y_1, \dots, x_{k-1}=y_{k-2}$ 
    construct a new group  $G=(x_1, \dots, x_{k-1}, y_{k-1})$ ;
    test all other combinations of subgroups
    of  $G$  with size( $k-1$ );
    if (all such subgroups are among the
        top  $M$  groups in  $LG_{k-1}$ )
      add  $G$  into  $CG_k$ ;
  }
}

```

[표 2]. Algorithm for generating candidate groups  $CG_k$

표2의 알고리즘을 이용하여 빈번하게 방문하는 페이지를 찾아냄으로써, 웹에 표현되는 web contents 조직이나 linkage를 개선시킬 수 있다.

### (3) Discovering frequent Itemsets

Frequent Itemset 은 많은 트랜잭션들 중에서 함께 가장 빈번하게 일어나는 item들의 그룹을 말한다. 여기서는 (1), (2) 의 알고리즘을 이용하여 Frequent Itemset 을 찾는다.

데이터 마이닝의 가장 중요한 문제중의 하나는 association rule 이다. association rule의 예를 하나 들어보면, “빵과 버터를 사는 사람의 30%는 우유를 산다.

그리고 이 아이템들을 모두 포함하는 사람들은 전체의 2%이다.” 이 사실에서 30%는 rule의 confidence이고, 2%는 rule의 support이다. association rule을 마이닝 하는 데에서 가장 중요한 문제는 minimal support를 가지는 아이템들의 조합이 너무 많이 나올 경우이다. 이러한 아이템의 조합을 large Itemset 이라고 한다[1].

$I = (I_1, I_2, \dots, I_k)$ 는 빈번하게 일어나는 Itemset이고,  $I_i$ 의 support는 다음에 나타난 식과 같다.

$$\sigma(I_i) = \frac{|\{t \in T : I_i \subseteq t\}|}{|T|}$$

보통 support threshold를 마이닝 하기 전에 정하여 알고리즘에 적용한다. 이러한 minimum support 값을 만족하는 Itemset 들을 구한다. 위의 결과로 Association rules 은 여러 아이템들 사이에서 co-occurrence로 일어나는 패턴의 items 사이의 관계를 캡쳐한다. 따라서 사용자들의 traversal 패턴을 기반으로 한 URL reference 사이의 관계를 캡쳐해낸다[2].

$$X \Rightarrow Y (\sigma_r, \alpha_r)$$

association rule  $r$  은  $X \cup Y$ 의 support가  $\sigma_r$ 이고,  $\alpha_r$ 이 주어진 rule의 confidence인 형식의 표현이다.

## 4. 결 론

본 논문에서 제안한 개인화 된 추천정보 소개를 위한 Web Usage Mining 알고리즘을 통하여 좀더 정확하게 빈번한 웹 traversal 패턴을 추출하고, 자주 방문하는 웹 페이지들의 그룹들, 빈번한 Itemset들 등의 사실을 알아내도록 한다.

이러한 알고리즘을 실제 사이트에 적용하여 인터넷 사용자들에게 보다 정확한 개인화 된 추천정보를 제공하여 시간의 절약과 동시에 사용자에게 가치 있는 정보를 실시간으로 제공할 수 있게 하는 것이 최종적인 목표이다. 따라서, Web Usage Mining 알고리즘을 사용하여 웹 로그를 분석하고, 이 결과를 이용하여 실제 사이트에서 실시간으로 정확하게 그리고 빠르게 추천정보를 줄 수 있는 personalization 알고리즘을 개발하는 것이 향후 계획이다.

## 【참 고 문 헌】

- [1] <http://personalization.co.kr/>
- [2] Bamshad Mobasher and Robert Cooley, Jaideep Srivastava "Automatic Personalization Based on Web Usage Mining", Technical Report TR99-010 , 1999.
- [3] Sloban Vucetic and Zoran obradovic, "A Regresion-Based Approach for scaling-up personalized systems in e-commerce", ACM SIGKDD, p13-21, 2000.
- [4] K.-L.Wu, P.S.Yu, and A.Ballman, "SpeedTracer:A Web usage mining and analysis tool", IBM systems Journal, Vol37, No.1, 1997.