# 웹 마이닝 시스템 설계 및 유용한 접근 패턴 정의

김종달*, 김성민*, 남도원*, 이동하**, 이전영*

*포항공과대학교 전자컴퓨터공학부

** 포항공과대학교 정보통신연구소

(lark, smkim, irene, dongha, jeon)@postech.ac.kr

# Design of the web data mining system and definition of useful access patterns

Jong-dal Kim*, Sung-Min Kim*, Do-Won Nam*, Dong-Ha Lee**, Jeon-Young Lee*

* Division of Electrical and Computer Engineering, POSTECH

** POSTECH Information Research Laboratories, POSTECH

요약

인터넷 서비스 제공자들이 관심을 가지고 있는 것 중 하나는 인터넷 사용자들의 서비스 이용 패턴과 경향을 분석하는 것이다. 이를 통해 매출 증대와 실제 경영에 도움이 되는 사용자의 특성을 이해할 수 있기 때문이다. 이와 관련된 기본적인 접근방법은 사용자가 웹 서버에 접근했을 때 서버에 남는 웹 로그를 분석하여 사용자 패턴을 분석하는 것이다. 웹 로그 분석에 전형적인 통계기법이 사용되고 있다. 그러나 단순 통계 기법 만으로는 알려지지 않는 데이터들 사이에 숨겨진 유용한 정보를 찾는 데에는 한계가 있다. 최근에는 이러한 한계를 극복하기 위해 데이터 마이닝 기술을 이용한 새로운 접근 방법이 시도되고 있다. 그러나 실제로 웹 로그에서부터 데이터 마이닝 기술을 이용하는 데에는 전처리 과정의 어려움과 실제 유용한 패턴을 어떻게 정의하는 가가 어려운 문제이다. 본 연구에서는 로(raw) 데이터인 웹 로그에서 유용한 패턴을 찾기 위한 전처리 과정을 알아보고, 웹 마이닝 시스템에 적합한 트랜잭션의 데이터 구조를 제시한다. 그리고 정의된 데이터 구조를 통한 패턴 발견 과정인 웹 사이트의 개념계층을 이용한 통계 기법과 연관규칙(Association Rules) 탐사에 대해 알아본다. 마지막으로 정의된 데이터 구조를 통한 새로운 유용한 패턴을 정의한다.

## 1. 서론

인터넷의 보급이 급속히 증가함으로써 기존의 판매, 유통, 금융 등이 인터넷으로 옮겨 가고 있다. 대표적인 예로 서점과 음반 판매의 경우 인터넷 구입이 증가하고 있으며, 금융의 경우 인터넷 뱅킹의 사용이 증가하고 있는 실정이다. 이와 함께 인터넷 서비스 제공자들이 관심을 가지고 있는 것 중 하나는 인터넷 사용자들의 서비스 이용 패턴과 경향을 분석하는 것이다. 이를 위해 사용되는 정보에는 사용자가 웹 서버에 접근했을 때 남기게 되는 웹 로그가 있다. 웹 로그는 웹 서버에 대한 모든 요구를 웹 서버에 의해 기록된 파일로써 대부분의 웹서버는 CDRN과 NCSA에 의해 HTTP 프로토콜의 일부로 명시된 '*Common Log Format*'을 따른다.[11] 웹 로그를 분석하는 방법에는 기존의 통계기법이 사용되고 있다. 이와 관련된 상용제품에는 weblog[1], webtrends[2], accrue[3] 등이 있다. 그러나 단순한 통계기법으로 사용자의 유용한 패턴을 분석하는 데에는 한계가 있다. 최근에는 데이터 마이닝 기술을

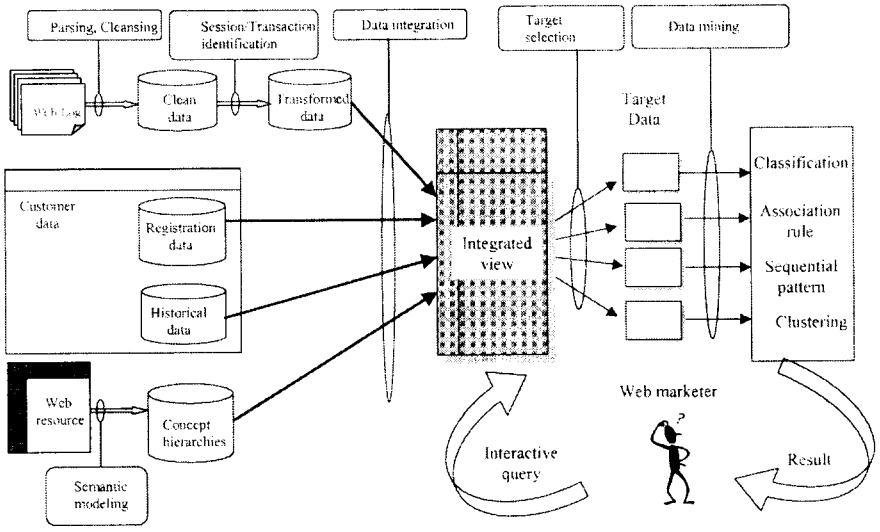



그림 1 : 웹 마이닝 시스템 구조

도입하여 유용한 새로운 패턴을 찾을려는 시도가 이루어 지고 있다.[4] 그러나 실제로 웹 로그에서부터 데이터 마이닝 기술을 이용하는 데에는 전처리 과정의 어려움과 실제 유용한 패턴을 어떻게 정의하는 가가 어려운 문제다. 본 연구[1]에서는 로(raw) 데이터인 웹 로그에서 유용한 패턴을 찾기 위한 전처리 과정을 알아본 후 웹 마이닝 시스템에 적합한 트랜잭션의 데이터 구조를 정의한다. 이를 통해 웹 사이트의 개념계층을 이용한 통계 기법과 연관규칙(Association Rules) 탐사가 가능하다. 그리고 시간 구간을 적용한 새로운 패턴, 사용자가 사이트를 떠나게 되는 패턴을 정의해 보고, 실제 응용방법도 알아 본다.

이 논문은 다음과 같은 구조를 가진다. 2절에서는 웹 데이터를 기반으로 한 웹 마이닝에 관한 관련 연구를 알아본다. 3절에서는 [4][5]에 기반한 웹 마이닝의 전처리 과정을 알아 보고, 웹 마이닝 시스템에 적합한 트랜잭션 테이터 구조를 제시한다. 4절에서는 3절에서 제시한 트랜잭션의 데이터 구조에서 얻을 수 있는 새로운 유형의 패턴들에 대해서 알아본다. 끝으로 5절에서는 결론을 제시한다.

## 2. 관련연구

기존의 데이터 마이닝은 대용량의 데이터 베이스로부터 기존의 알려 지지 않은 단순한 질의어로 추출할 수 없는 형태의 유용한 정보를 찾아내고 이를 바탕으로 데이터에 대한 통찰(insight)를 얻는 것으로 정의할 수 있다.[6] 반면 웹 마이닝은 대상되는 데이터가 데이터 베이스가 아닌 WWW의 데이터에서 유용한 정보의 발견과 분석으로 정의한다. 현재 웹 데이터 마이닝은 대상이 되는 웹 데이터(구조,내용,사용)에 따라 다음과 같은 3가지의 분야로 나눌 수 있다.[7]

- 웹 구조 마이닝(web structure mining)
- 웹 내용 마이닝(web content mining)
- 웹 사용 마이닝(web usage mining)

웹 구조 마이닝은 웹 사이트와 웹 페이지의 구조적 요약 정보를 얻는 것을 목표로 한다. 웹 내용 마이닝은 실제 웹 사이트를 구성한 의미적인 내용에 대한 정보 추출에 관심이 있다. 끝으로 웹 사용

[1] 본 연구는 BK21 사업을 통하여 포항공과대학교 전자.컴퓨터공학부에 주어진 교육부의 재정지원을 통해 이루어진 것입니다.

마이닝은 사용자의 웹 사용 패턴을 분석하여 사용자에게 더욱더 친숙하게 페이지를 재구성하거나, 웹 서버 로드 밸런스, 사용자별로 차별화된 웹 페이지 구성 등에 이용되고 있다.

이 논문의 웹 마이닝 시스템은 웹 사용 마이닝과 관련이 있다.

## 3. 웹 마이닝 시스템

웹 데이터 마이닝은 로(raw) 데이터로부터 새로운 패턴을 발견하기 위한 전처리 과정(preprocessing), 전처리 과정에서 얻은 데이터에서 유용한 정보를 얻기 위한 패턴 발견 과정(Pattern Discovery) 그리고 마지막으로 생성된 규칙과 패턴을 분석(Pattern Analysis)하는 과정으로 나눌 수 있다. 이는 생성된 규칙과 패턴을 질의의 형태로 분석 가능하게 한다.[4]

이 논문의 웹 마이닝 시스템은 그림1과 같은 형태의 구조를 가진다.

### 3-1 전처리 과정

웹 로그(web log)에서 패턴 발견을 위해 데이터를 추상화 하는 일련의 과정을 전처리 과정이라 한다. 전처리 과정은 데이터 정제(Data Cleaning), 유일한 사용자 구분(User Identification), 세션 구분(Session Identification), 세션 보정(Path Completion)으로 나눌 수 있다.[5]

#### 3-1-1 데이터 정제(Data Cleaning)

웹 로그는 웹 서버가 클라이언트의 요청을 처리한 내용을 기록한다. 예를 들어 A라는 페이지에 10개의 그림(gif or jpg 파일 포맷)을 포함하고 있다고 했을 경우, 클라이언트가 A 페이지를 요청을 했을 때, 11개의 요청의 처리 결과가 웹 로그에 기록된다. 그러나 우리가 원하는 웹 로그의 정보는 사용자의 페이지 탐색의 리스트이다. 따라서 웹 로그에서 필요 없는 요청(그림 파일 등)을 제외한 페이지만의 리스트가 필요하다.

#### 3-1-2 유일한 사용자 구분(User Identification)

웹 사이트에 접근한 사용자를 구별하기 위해서는 기본적으로 웹 로그의 IP를 통해 구별한다. 또한 서버에서 제공하는 쿠키(cookie)정보를 이용하여 좀 더 정확하게 사용자를 구별할 수 있다. 그러나 다음과 같은 문제점을 포함한다.

1) 인터넷 서비스 제공자(ISP)는 프럭시 서버(Proxy Server)를 이용하여 가입자에게 제공하는 경우가 많다. 이 경우 하나의 IP를 여러 사용자가 공유할 수 있다.
2) 인터넷 서비스 제공자(ISP) 또는 응용프로그램으로 사용자의 IP를 가변적으로 할당할 수 있다. 이 경우 여러 IP를 통해서 들어오는 사람에도 불구하고 한 명의 사용자로부터 발생한 요청이다.

위에서 제시한 문제점을 해결하기 위해서는 웹 로그 뿐만 아니라 다른 메커니즘이 요구된다.

#### 3-1-3 세션구분(Session Identification)

세션(Session)은 한 사용자가 주어진 사이트에 요청(request)을 처음 시도해서 사이트(site)를 떠날 때까지 발생한 일련의 과정이다. 따라서 실제 사용자의 웹 사용 패턴을 발견하기 위해 세션 구분(Session Identification)은 데이터 마이닝을 위한 입력 데이터로써 매우 중요한 의미를 가진다. 그러나 실제 웹 로그만으로 세션을 구별하기에는 문제가 있다. 가장 어려운 문제가 사용자 구별이 쉽지 않다는 것이다. 여기에서는 IP에 따라 사용자를 구별한다고 가정한다. 앞의 가정에 의해 같은 IP에 의해 사용자를 구별하더라도, 다른 시간에 접속하는 경우 이 세션(session)은 다른 세션으로 보아야 한다. 따라서 사용자가 사이트를 떠나는 시점을 파악해야 한다. 이를 위해 사용자의 요청이 있은 후 일정한 시간동안 요청이 발생하지 않은 경우, 사이트를 떠

난 것으로 가정하고 세션을 종료시키고, 새로운 세션을 시작한다.

### 3-1-4 세션 보정(Path Completion)

웹 로그의 데이터를 가지고 세션을 구한 경우 실제 웹 사이트의 구조에서 생성될 수 없는 세션이 발생할 수 있다. 이런 경우 웹 로그에서 이미 만들어 놓은 사이트 구조를 통해서 세션을 보정할 수 있다. 예를 들어 B페이지에서 D를 페이지를 가지 위해 반드시 C페이지를 거쳐야 하는 사이트 구조를 가정하자. 이 때 세션을 구했을 때 B→D가 발생한 경우 C페이지를 추가하여 B→C→D로 구성할 수 있다.

## 3-2 트랜잭션 구분

트랜잭션(Transaction)은 세션(session)에 의미를 부여한 것으로 마이닝을 위한 기본 단위가 된다. 이는 패턴 발견 과정에서 사용될 연관 규칙(Association Rules), 순차 패턴(sequential patterns), 타임 세그먼트를 이용한 사용자 패턴(the user pattern using time segment), 떠남 패턴(exit pattern)을 발견하는데 사용 되는 기본적인 단위이다.

세션(session)에서 트랜잭션(Transaction)을 구분하는 방법에는 세션과 트랜잭션을 동일하게 보는 기본적인 방법, 페이지를 참조한 길이(reference length)를 기반으로 한 방법, [8]에서 가장 처음 시도된 방법으로 사용자가 브라우저(browser)의 백 버튼을 클릭하기 전 까지를 하나의 트랜잭션으로 보는 Maximal Forward Reference 방법, 그리고 일정한 시간 윈도우를 적용한 Time Window 방법이 있다. [5]

트랜잭션에 의미를 부여하는 또 다른 방법은 추가 페이지(Auxiliary Page)를 제외한 내용 페이지(Content Page)만으로 트랜잭션을 다시 만드는 방법이 있다. 내용/추가 페이지를 함께 고려한 트랜잭션(transaction)에 기반한 패턴 발견은 의미 없는 많은 패턴을 발생하며, 패턴 발견에 많은 오버헤드가 발생한다. 따라서 내용 페이지만을 고려한 트랜잭션 생성은 의미를 가진다. 그림2,3 는 하나의 세션을 트랜잭션으로 구분할 때 내용/추가 페이지를 함께 고려한 경우와 내용 페이지 만을 고려한 경우의 예를 보여준다.[5]

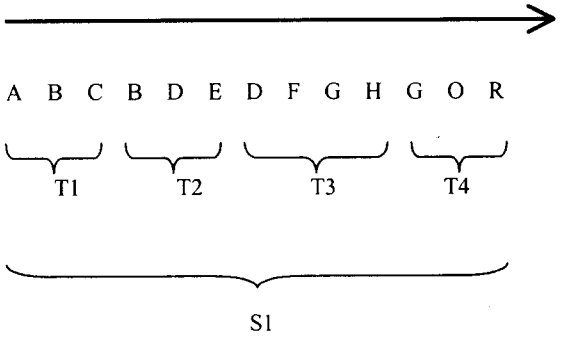



**그림2:내용/추가 페이지를 함께 고려한 트랜잭션**

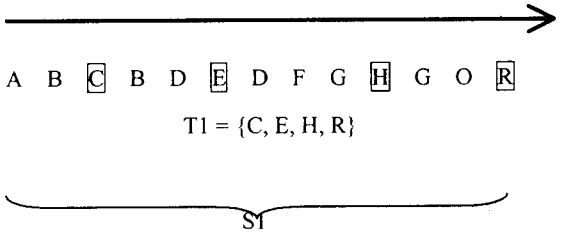



**그림3 : 내용 페이지만을 고려한 트랜잭션**

**정의 1 : 내용 페이지(Content Page)**

트랜잭션(transactioin)에 포함된 페이지 중에서 사용자가 정보를 얻을 수 있는 페이지

**정의 2 : 추가 페이지(Auxiliary Page)**

트랜잭션(transaction)에 포함된 페이지 중 내용 페이지(content page)를 찾기 위해 필요한 페이지

## 3-3 트랜잭션의 데이터 구조

**정의 3 : 세션(Session) $S_i$**

$S_i = <IP_i, \{R_1=(R_1.url, R_1.time), \ldots, R_{n-k}=(R_{n-k}.url, R_{n-k}.time)\}>$

$IP_i$는 세션에 해당하는 IP 주소, 사용자의 요구(request) $R_i$는 사용자가 요구한 페이지(url)와 페이지를 요구한 시간(time)으로 구성된다. k는 [0, n-1]. 즉,

세션을 구성하는 요구(request)의 수는 최소 1부터 최대 n까지 이다. 그리고 i<j인 경우 $R_i$.url < $R_j$.url 인 조건을 만족시킨다. 세션 구분은 앞에서 제시한 방법을 따른다.

**정의 4 : 트랜잭션(Transaction) $T_i$**

$T_i = <UID_i, \{R_1 = (R_1.url, R_1.time, R_1.segment), \ldots, R_{n-k} = (R_{n-k}.url, R_{n-k}.time, R_{n-k}.segment)\}>$

1) $UID_i$는 세션(session)의 $IP_i$, 사용자 데이터베이스, 사용자 히스토리(history) 정보[2]에 의해 생성된 사용자 고유 번호이다. 만약 사용자 데이터베이스와 사용자 히스토리(history)가 존재하지 않은 경우, 세션(Session) $S_i$의 $IP_i$와 같은 것으로 간주한다.
2) 사용자의 요구(request) $R_i$는 사용자가 요구한 페이지(url), 페이지를 요구한 시간(time)과 페이지에 머문 시간 세그먼트(segment)로 이루어진다. 세그먼트(segment)를 구하는 방법은 $|R_{n+1}.time - R_n.time|$으로 구한다. 마지막 요구(request)의 세그먼드(segment)는MLE(Maximal Likelihood Estimate)을 이용하여 값을 예측한다.
3) $T_i$가 가지는 요구 R의 최대 개수 $|T_i|$는 n, 최소 개수는 1이다.

## 3-4 패턴 발견 과정

### 3-4-1 개념 계층을 이용한 통계 기법

웹 로그에서 얻을 수 있는 기본 정보는 다음과 같은 부분을 포함한다.

1) IP – 웹 서버에 접근한 사용자의 IP 주소를 알 수 있다.
2) Time – 웹 서버에 사용자(클라이언트)가 접근한 시간을 알 수 있다.
3) Resource Page – 사용자가 웹 서버의 어떤 페이지를 요청했는지 알 수 있다.

[2] 사용자 데이터베이스는 사용자 나이/주소/직업 등의 신상정보를 포함하고 있으며, 히스토리(history) 정보에서 등록 날짜, 마지막 로긴한 시간, 마지막 접속한 IP 주소 등을 알수 있다.

웹 로그를 통하여 앞의 정보에 대한 통계를 구할 수 있다[1][2][3]. 본 논문에서는 그림 4와 같이 3-스타 스키마(3-star schema)에 기반하여, 웹 로그에서 얻을 수 있는 통계정보를 구체화 하였다.

3-스타 스키마는 1, 2, 3 차원(dimension)에 대한 분석 기능을 제공한다. 예를 들어, 시간에 따른 A 페이지의 경향을 알고 싶을 경우 시간과 resource/ip 에 의한 분석이 가능하다.

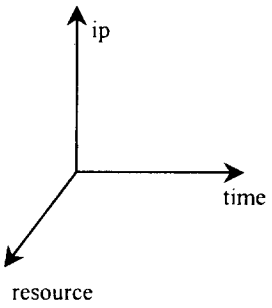



**그림 4 : 3 star 스키마에 기반한 통계 정보**

또한 개념 계층의 개념을 도입하여 하위 개념을 통하여 상위 개념의 통계값을 구할 수 있다. 그림 5는 ip dimension을 개념 계층을 사용하여 나타낸 것이다. 예를 들어, 한국 교육기관에서 접속된 사용자들이 가장 선호하는 페이지는 접속한 사용자의 개념 계층에서 도메인 명이 ac.kr의 사용자만의 통계값을 구한다.

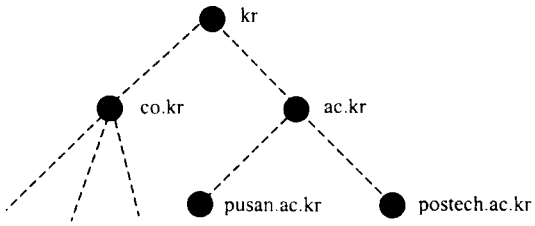



**그림 5 : 개념 계층을 이용한 IP dimension 트리**

1차원 도메인에서 얻을 수 있는 정보는 ip, time, resource에 해당하는 정보이다.

1) ip – 가장 많은 서비스를 요청한 ip 주소 or 가장 적은 서비스를 요청한 ip 주소
2) time – 가장 많은 서비스를 요청한 시간대 or 가장 적은 서비스를 요청한 시간대

3) resource – 가장 많이 요구된 페이지 or 가장 적게 요청된 페이지

2차원 도메인에서 얻을 수 있는 정보는 다음과 같다.

1) ip vs time – 시간에 따른 ip 접근 경향
2) ip vs resource – ip별로 자주 요구하는 페이지
3) time vs resource - 시간에 따른 페이지 접근 경향

앞에 설명한 1,2차원 도메인에서 얻을 수 있는 정보는 각 도메인에 관한 배경지식을 반영한 개념 계층을 이용하여 모두 확장 가능하다. 3-스타 스키마과 개념 계층에 기반한 통계 정보는 추상적인 통계정보를 가시적으로 나타냄으로써 중요한 의미를 가진다.

### 3-4-2 연관규칙 탐사

연관 규칙 탐사는 93년 R.agrawal에 의해 처음 소개 되었다.[7] 이는 두 항목집합 사이의 관련성을 나타낸다. 연관 규칙은 A→B와 같은 형태로 나타내며, A의 행위를 했을 때 B의 행위를 함을 나타낸다. 예를 들어 슈퍼마켓 데이터베이스의 경우 {맥주} → {땅콩}의 연관규칙은 맥주를 사는 사람들은 땅콩을 사는 경향이 있다라고 해석 가능하다. 여기에서 사용되는 중요한 척도 2개는 최소 지지도(support)와 최소 만족도(confidence)이다.

연관 규칙은 기본적으로 바스켓 데이터를 기본으로 한다. 따라서 웹 로그를 이용한 연관 규칙 탐사를 위한 새로운 바스켓 데이터가 필요하다. 이를 위해 앞 절에서 정의한 트랜잭션을 사용한다.

바스켓 데이터가 정의되면, 연관 규칙을 찾는 방법은 [9][10]과 동일하다. 적용 가능한 기본적인 알고리즘으로는 [9]의 apriori 알고리즘, [10]의 DHP 알고리즘이 있다. 그리고 항목집합(itemset)과 빈번 항목집합(large itemset)의 저장 구조는 [6]에서의 MLST 방법을 고려한다.

**정의 5 : k-페이지 항목집합(Page Itemset) $I_i$**

$I_i=\{url_1, url_2, \ldots, url_n\}$

$I_i$의 $url_j$(j=1,2,…,n)는 트랜잭션 $T_i$의 요구(request) 페이지 $R_i$.url의 부분집합이다.

예. $T_1$ = {$UIP_1$, {$url_1$, $time_1$, $segment_1$}, {$url_2$, $time_2$, $segment_2$}, {$url_3$, $time_3$, $segment_3$}}일 때 페이지 항목집합(Page Itemset) $I_1$는 다음과 같다.

1-페이지 항목집합 - {$url_1$} , {$url_2$} , {$url_3$} : 총 3개

2-페이지 항목집합 - {$url_1$, $url_2$}, {$url_1$, $url_3$}, {$url_2$, $url_3$} : 총 3개

3-페이지 항목집합 - {$url_1$, $url_2$, $url_3$} : 총 1개

**정의 6 : k-빈번 페이지 항목집합(Lareg Page Itemset) $L_i$**

**k-페이지 항목집합 중에서 최소 지지도(support)와 최소 신뢰도(confidence)를 만족하는 것이다.**

[9] 예를 들어 2-빈번 페이지 항목집합 { A, B }를 얻은 경우 이 패턴의 의미는 사용자가 이 사이트에 들어왔을 때 A, B 페이를 동시에 보는 경향이 있음을 의미한다.

## 4. 새로운 유형의 패턴

### 4-1 시간 구간을 적용한 패턴

사용자가 웹 페이지에 머문 시간은 웹 페이지에 대한 사용자 관심도를 반영하는 요소로써 중요한 의미를 가진다. 따라서 페이지에 머문 시간을 통해 앞에 제시한 내용 페이지(content page)와 추가 페이지(auxiliary page)를 나눌 수 있으며, 페이지에 새로운 의미를 부여할 수 있게 된다.

본 논문은 패턴1,2와 같이 사용자가 페이지에 머문 시간(time segment)을 부여하여, 새로운 패턴(pattern)을 제안한다.

**패턴 1 : 유용하지 않은 페이지 발견**

A(20)→B(45)→C(5)→B(10)→D(29)(여기에서 A(a)는 A page에서 머문 시간이 a라는 것을 의미)와 같은 패턴이 발견되었는데, C의 시간 구간(time segment)이 매우 짧을 때에는 페이지 C에는 유용한 정보가 없다고 볼 수 있다. 이를 통하여 C 페이지의 내용

을 B나 D 페이지에 포함시키거나 페이지를 제거하여 사이트의 구조를 재구조화 하는데 도움을 줄 수 있다.

**패턴 2 : 베이스 페이지 발견**

하나의 페이지가 반복적으로 나타나면서, 그 페이지에 머문 시간(time segment)이 짧을 수 있다. 즉 다음과 같은 패턴(pattern)을 분석한다. B(12)→C(25)→B(10)→D(29)→E(35)→B(15)→F(33)와 같은 패턴이 발견된 경우 B 페이지를 베이스 페이지(base page)라 말한다. 베이스 페이지는 내용은 제공하지 않지만 내용의 인덱스를 제공하는 페이지이다. 예를 들어 신문의 상위 페이지, 사이트를 구조를 나타내는 사이트맵 페이지 등이 있을 수 있다. 베이스 페이지를 찾음으로 베이스 페이지의 이미지를 줄이 거나, 빠른 서버나 디스크에 베이스 페이지를 배치함으로써 사용자에게 빠른 사이트의 느낌을 줄 수 있으며, 베이스 페이지에서의 광고 배너의 위치를 링크 주위에 배치시킴으로 광고 효과를 극대화 시킬 수 있다.

시간 구간을 적용한 패턴 발견을 위해 최소 세그먼트(Minimal Segment), 최대 세그먼트(Maximal Segment), 페이지/시간 항목집합(Page Itemset)을 정의 한다.

**정의 7 : 최소 세그먼트(Minimal Segment)**

**페이지에 머문 시간 중에서 사용자가 정보를 얻을 수 없을 만큼 적은 시간.**

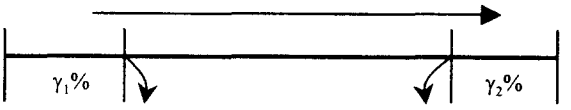
사용자가 주는 요소이다. 모든 트랜잭션의 페이지 세그먼트(segment)의 분포에서 하위 $\gamma_1$%의 분포를 만족시키는 지점에서 최소 세그먼트(segment)를 구할 수 있다. 이를 통하여 추가 페이지(Auxiliary Page)를 찾을 수 있다.

**정의 8 : 최대 세그먼트(Maximal Segment)**

**페이지에 머문 시간 중에서 사용자가 정보를 얻을 수 있을 만큼의 시간.**

사용자가 주는 요소이다. 모든 트랜잭션의 페이지 세그먼트(segment)의 분포에서 상위 $\gamma_2$%의 분포를 만족시키는 지점에서 최대 세그먼트를 구할 수 있다. 이를 통하여 내용 페이지(content page)를 찾을 수 있다.

그림 6은 모든 트랜잭션에 속하는 페이지의 세그먼트의 크기를 내림차순으로 정렬한 다음, 상위 $\gamma_1$%, 하위 $\gamma_2$%를 찾아 최대/최소 세그먼트의 크기를 찾는 것이다.




최대 세그먼트 최소 세그먼트

[전체 트랜잭션의 페이지 세그먼트 크기를 내림차순으로 정렬]

**그림 6 : 최소/최대 세그먼트**

**정의 9 : 페이지/시간 항목집합(Page Itemset) $TI_i$**

**$TI_i$ =<{$url_1$,$duration_1$}, {$url_2$,$duration_2$}, ..., {$url_n$, $duration_n$}>**

$TI_i$의 $url_k$(k=1,2,...,n)는 트랜잭션 $T_i$의 요구(request) $R_i$.url의 부분집합이다. 그리고 $duration_k$(k=1,2,...n)은 트랜잭션 $T_i$의 $url_k$에 해당하는 세그먼트(segment)로 이루어 진다.

$duration_k$(k=2,...,n-1) 중에서 적어도 하나는 최소 세그먼트를 만족시켜야 한다. 그리고 $duration_k$(k=j,...,n) 중에서 적어도 하나는 최대 세그먼트를 만족시켜야 한다. j는 최소 세그먼트를 가지는 최소 duration의 인덱스이다.

예. 최소 세그먼트가 25, 최대 세그먼트가 6일 때 다음과 같은 페이지/시간 항목집합이 있을 수 있다.

TI = < {A,20}, {B,45}, {C,5}, {B,10}, {D,29} >

## 4-2 떠남(exit)의 패턴

사용자가 사이트에 접속한 뒤 사이트를 떠나는 패턴을 분석함으로써 사이트의 문제점(예를 들어 떠남 페이지에 이미지가 많아 속도가 느림 등)을 파악할 수 있다. 또한 쇼핑몰인 경우 유용한 정보를

가진 페이지를 추가함으로써 혹시 사이트를 떠날 고객을 좀 더 자신의 사이트에 붙들어 둘 수 있다.

**정의 10 : k-떠남 페이지 항목집합(Exit Page Itemset) $EI_i$**

$EI_i = \{url_1, url_2, \ldots, url_k\}$

$EI_i$의 $url_k$(k=1,2,...,n)는 세션 $S_i$의 마지막 요구(request) 페이지 $R_n$.url 부터 시작하여 $|I_i|$ 만큼의 요구(request)로 이루어 진다.

예. $S_1 = \{IP_1, \{url_1, time_1\}, \{url_2, time_2\}, \{url_3, time_3\}\}$일 때 떠남 페이지 항목집합(Exit Page Itemset) EI는 다음과 같다.

1-떠남 페이지 항목집합 - $\{url_3\}$

2-떠남 페이지 항목집합 - $\{url_2, url_3\}$

3-떠남 페이지 항목집합 - $\{url_1, url_2, url_3\}$

**정의 11 : k- 빈번 떠남 페이지 항목집합(Large Exit Page Itemset) $EL_i$**

k-떠남 페이지 항목집합 중에서 최소 지지도(support)를 만족하는 것들의 집합이다. 전체 지지도(support)에 대한 전체 항목은 세션의 마지막 k개의 요구(request)들이다. 예를 들어 표 1에서 최소 지지도를 2라고 했을 경우, k-빈번 떠남 페이지 항목집합은 다음과 같다.

1-빈번 떠남 항목집합 {D}

2-빈번 떠남 항목집합 {C,D}

위의 패턴이 가지는 의미는 1-떠남 항목집합에서 D 페이지에서 가장 많이 떠남이 발생함을 보여준다. 그리고 2-떠남 항목집합을 통하여 C, D 페이지를 보고 떠나는 패턴을 알 수 있다.

**정리 1 : $k \geq 2$, k-빈번 떠남 항목집합이 $\{url_1, \ldots, url_k\}$일 때, $\{url_2, \ldots, url_k\}$도 (k-1)-빈번 떠남 항목집합이 된다.**

| | |
|---|---|
| S1 | {u1, {A,10}, {B,20}, {C, 30}, {D,20}} |
| S2 | {u2, {A,6}, {B,30}, {C, 30}, {E,12}} |
| S3 | {u3, {B,7}, {C,22}, {B, 3}, {A,21}} |
| S4 | {u4, {A,19}, {B,5}, {C, 24}, {B,19}} |
| S5 | {u5, {A,10}, {F,21}, {C, 21}, {D,20}} |

**표 1 : k-빈번 떠남 페이지 항목집합의 예**

## 5. 결론

인터넷 사용자들의 서비스 이용 패턴과 경향을 분석하기 하기 위해 먼저 선행되어야 할 것은 웹 로그에서 분석 대상이 되는 의미 있는 트랜잭션을 정의하는 것이다. 본 논문은 웹 마이닝을 위한 트랜잭션을 웹 로그, 사용자 데이터베이스 그리고 사용자 히스토리(history) 정보를 통하여 웹 마이닝 시스템에 적합한 트랜잭션 데이터 구조를 제시하였다 이를 바탕으로 유용한 패턴을 얻기 위한 방법으로 개념 계층을 이용한 3-스타 스키마 구조를 제시하였으며, 웹에서의 연관규칙 탐사에 대해 알아보았다. 그리고 끝으로 시간 구간을 적용한 패턴, 사용자가 사이트를 떠나는 경향을 파악하는 떠남 패턴에 대해 알아 보았다.

웹 마이닝을 위해 의미 있고 정확한 트랙잭션의 구성과 유용한 새로운 패턴을 정의함으로써 실제적인 도움을 줄 수 있는 가치 있는 정보를 창출한다고 생각한다.

## 참고문헌


[1] weblog. http://www.weblog.com

[2] webtrends. http://www.webtrends.com

[3] Accrue. http://www.accrue.com

[4] Jaideep Srivastava, Robert Cooley, Mukund Deshpand. Web Usage Mining: Discovery and Applications of Usage Patterns from Web Data. *SIGKDD Explorations, Vol. 1, Issue , 2*, 2000.

[5] Robert Cooley, Bamshad Mobasher, and Jaideep Srivastava, Data Preparation for Mining World Wide Web Browsing Patterns, *Knowledge and Information Systems*

*VI(1)*, 1999

[6] 이동하, "다단계 선순열 트리와 개념 계층을 이용한 대용량 관계형 데이터베이스에서의 연관 규칙 추출 기법", 박사학위 논문, 포항공과대학교, 2000

[7] S K Madria, S S Bhowmick, W K Ng, E P Lim, "Research Issues in Web Data Mining", Proceedings of the 1st International Conference on Data Warehousing and Knowledge Discovery (DAWAK 99), 1999

[8] M.S. Chen, J.S. Park, and P.S. Yu. Data mining for path traversal patterns in a Web environment. In *Proceedings of the 16th International Conference on Distributed Computing Systems*, pages 385-392, 1996

[9] R. Agrawal, T. Imielinski, and A. Swami, "Mining association rules between sets of items in large database" In proceedings of the ACM SIGMOD conference on Management of Data, pp. 207-216, Washington, D.C., May 1993

[10] J.-S. Park, M.-S. Chen, and P. S. Yu, "An Effective Hash Based Algorithm for Mining Association Rules", ACM SIGMOD Conference on Management of Data, pages 175-186, My,, 1995

[11] A. Luotone, The common log file format, http://www.w3.org, 1995

[12] D.-H. Lee, D.-Y. Seo, N.-H. Kim, J.-Y. Lee, "Discovery and Application of User Access Patterns in the World Wide Web.", Proceedings of the 4th World Congress on Expert Systems, March 16-20, 1998, pp.321-327

[13] 남도원, 이동하, 서동렬, 이전영, "웹로그에서의 사용자 접근 패턴 분석", HCI '98 학술대회 논문집, 한국정보과학회 HCI 연구회, 1998

1. 소속기관

한글 : 포항공과대학교 컴퓨터공학과

영문 : Dept. of Computer Science and Engineering, Pohang University of Science and Technology

2. 전화번호

0562)279-5660

3. E-mail

lark@postech.ac.kr

4. 주소

포항시 남구 효자동 산31번지 포항공과대학교 정통연 IIS Lab.