연관규칙을 이용한 개인화 시스템 설계

윤종찬* · 윤성대**

Design of Personalized System using an Association Rule

Jong-Chan Yun* · Sung-Dae Youn**

이 논문은 2006년도 부경대학교 기성회 학술연구비를 지원받았음

요 약

최근 웹상에서 사용자들의 요구가 다양해지고 있다. 또한, 웹 사용자들은 보다 편리하고 빠르게 찾고자 하는 자료나 상품을 검색하기를 원하고 있다. 이것은 웹 사용자들마다 검색 기준이나 성향이 다르기 때문에 웹 설계자가 구현한 환경을 사용하려면 불필요한 반복 작업이 따르기 때문이다.

본 논문에서는 로그파일 분석기법을 이용하여 웹상에서 일어나는 사용자 패턴을 분석하여 웹 사이트의 정보를 사용자에게 보다 효과적으로 전달하기 위한 시스템을 제안하였다.

제안한 시스템의 고객 데이터(로그파일분석)은 데이터마이닝의 툴 중의 하나인 EC-Miner를 통해 이뤄지고 각 이동경로에 가중치를 줘서 개인화에 맞도록 적절한 레이아웃을 제공하고자 한다.

ABSTRACT

Currently, user require is diverse on the Web. Furthermore, each web user is wishing to retrieve data or goods that they want to look for more conveniently and more quickly. Because different search criteria and dispositions of web users, they lead to unnecessary repeated operations in order to use implemented by web designer.

In this paper, we suggest the system that analyzes user patterns on the Web using the technique of log file analysis and transfers more effectively the information of web sites to users.

And we analyze the log file for customer data in the system the proposed method are implemented by means of EC-Miner that is one of the tool of datamining, and aims to offer appropriate Layout corresponding with personalization by giving weight to each transport path.

키워드

개인화, 컨텍스트, 웹마이닝, 연관성규칙

Ⅰ. 서 론

최근 새로운 IT(Information Technology : 정보기술) 패러다임으로 인해 실제 세계의 산재해 있는 컴퓨팅 장치들과 인간을 자연스럽게 상호작용하도록 하고 있다. 사용자들이 필요한 정보나 서비스를 언제 어디서나 실시간으로 제공 받을 수 있게 되었다.

현재 웹쇼핑몰이나 전자도서관의 웹환경에서의 상

* 부경대학교 전자상거래시스템(협동)과정 박사과정수료
** 부경대학교 전자계산학과 교수(교신저자)

접수일자 : 2007. 1. 31

품조회(구매)나 도서조회(구매)를 할 때, 사용자들의 웹 검색에 따른 불편이 따르고 있다. 웹 사용자들마다 검색 기준이나 성향이 틀리기 때문에 웹 설계자가 구현한 웹 환경을 사용하려면 많은 시간이나 불필요한 반복 작업이 따르기 마련이다[1,2,3]. 그러므로, 본 연구에서는 이러한 불필요한 반복 작업이나 시간을 줄이기 위해 사용자의 웹 정보를 이용한 패턴 분석을 이용해 사용자의 웹 사용에 대한 편리성(시간 낭비와 불필요한 Click수를 줄인다.)을 부여하고자 한다[4,5,6,7].

인터넷 환경과 웹속성에 대한 이해를 바탕으로 궁극적으로 사용자 중심 웹환경의 레이아웃을 위하여 로그파일 분석기법을 적용하여 웹상에서 일어나는 사용자의 패턴을 분석하고, 이를 통해 웹사이트의 정보를 사용자에게 보다 효과적으로 전달하기 위한 웹사이트 구조를 제안하여 사용성을 높이는데 있다. 이러한 분석방법으로 사용자들이 웹사이트를 사용하는 데 많은 시간을 줄 일수 있도록 레이아웃을 설계할 수 있고, 웹데이터를 관리하도록 한다. 그 결과로서 웹사이트에서 제공되는 메뉴 및 웹페이지의 효율적 배치(사용자의 패턴에 의한 페이지 제공)와 논리적인 구조를 설계하기 위한 실증적 근거를 얻을 수 있도록 한다.

본 논문은 2장에서 관련연구, 3장에서는 시스템 설계, 4장에서는 시스템 분석 및 결과와 5장에서 결론 및 향후 연구과제에 대해서 서술한다.

Ⅱ. 관련연구

2.1 연관성 규칙

웹사이트에서는 이용자에 의해 함께 접속되는 페이지를 연관성 규칙으로 생각할 수 있는데, 예를 들어 어떤 접속자가 사이트 A를 방문한 다음 사이트 B를 방문하는 경우가 자주 발생할 수 있는데 이를 연관성 규칙으로 생각할 수 있다. 즉 연관성 규칙 "IF~THEN"의 "A→B"는 페이지 A를 보는 고객은 페이지 B도 본다는 의미이다. 이와 같이 생성된 연관성 규칙은 고객 데이터베이스로부터 고객들의 구매 품목들 간의 관련성을 발견하고 마케팅 자료로 활용하는데 많이 사용된다. 또한, 본 연구에서 적용한 바와 같이 웹에서의 사용자 접근 패턴을 분석하는 데 사용될 수 있다[6,8].

연관규칙 탐사과정은 크게 두 단계로 진행된다. 첫 번째는 높은 지지도(Support)를 갖는 아이템의 집합을 식별하는 작업이고, 두 번째 단계는 높은 신뢰도(Confidence)를 갖는 연관규칙을 도출하는 작업이다. 여기서 지지도와 신뢰도의 개념은 아주 중요한 개념으로 빈발 항목 집합을 찾아내는데 있어 큰 역할을 한다[7,9,10].

지지도란, 전체 트랜잭션에서 특정 패턴(A→B)이 차지하는 비율이고, 신뢰도란 A를 구매하는 고객 중에 B를 구매하는 고객이 차지하는 비율을 말한다.

$$Support(A \Rightarrow B) = P(A \cap B) = \frac{\text{웹 페이지 } A\text{와 } B\text{를 동시 방문자 수}}{\text{전체 웹 사이트 방문자 수}} \quad \cdots \text{(식 1)}$$

신뢰도는 A의 모든 항목을 포함하고 있는 트랜잭션의 개수에 대하여 B 또한 포함하는 트랜잭션의 비율을 의미한다.

$$Confidence(A \Rightarrow B) = P(B|A) = \frac{\text{웹 페이지 } A\text{와 } B\text{를 동시에 방문자 수}}{\text{웹 페이지 } A\text{의 방문자 수}} \quad \cdots \text{(식 2)}$$

2.2 컨텍스트(Context)

유비쿼터스 컴퓨팅에서 사용되는 컨텍스트의 개념은 아직까지 통일된 정의가 없는 상태이며, 많은 연구자들이 독자적인 정의를 바탕으로 객관화된 컨텍스트의 개념을 정의하기 위하여 꾸준히 시도하고 있다. 사용자의 위치, 사용자 주변에 있는 사람 정보, 그리고 사용 가능한 자원 등의 정보를 컨텍스트라고 정의하였다. 이것은 사용자의 환경이 변화하여도 일관성 있는 컨텍스트의 개념이 적용될 수 있는 특징을 나타낸다[11,12].

컨텍스트는 컨텍스트 정보를 의미한다. 컨텍스트란 한 개체(entity)의 컨텍스트를 나타내는데 사용될 수 있는 어떠한 정보이다. 여기서 한 개체란, 사람, 장소 또는 목적물로서 사용자와 응용물 뿐만 아니라 이들 사이의 인터페이스 모두 포함된다[15]. 인터페이스시 응용가능한 거의 모든 정보는 컨텍스트 정보로 인식될 수 있다. 예를 들면 표 1과 같다[8,13].

표 1. 컨텍스트 정보
Table 1. Context Infomation

정보명	성명
공간정보	위치, 방위, 속도 및 가속도
시간정보	시간, 날짜, 계절
환경정보	온도, 공기의 특성, 밝기
사회적 정보	누구와 함께 있는지
인접한 정보	접근 가능한 장치 및 자원
자원의 이용가능성	베터리, 네트워크 및 통신
물리학적 측정	혈압, 심장박동수, 호흡수
활동력	말하기, 읽기, 걷기, 뛰기 등

컨텍스트 인식이란, 나열한 컨텍스트 정보를 이용할 수 있음을 의미한다. 만약 한 시스템이 정보를 추출해서, 해석할 수 있고 컨텍스트 정보를 이용해서 현재의 컨텍스트에 그것을 기능적으로 적용시킬 수 있을 경우 그 시스템은 컨텍스트를 인식하고 있는 것이다. 이런 시스템의 어려운 점은 컨텍스트 정보를 재현하고 처리, 파악하는 복잡성에 있다. 일반적으로 컨텍스트 정보를 파악하기 위해서는 일부 추가적인 센서들 또는 프로그램이 요구된다. 컨텍스트 정보를 어플리케이션으로 전송하고 다른 어플리케이션이 같은 컨텍스트 정보를 이용할 수 있기 위해서는 이런 정보에 대한 공통적인 표현 형식이 있어야 한다. 또한 컨텍스트 정보를 얻기 위해서는 어플리케이션에 "지능"이 포함되어 정보를 처리하고 그 의미를 추론할 수 있어야 한다. 이것은 컨텍스트이란 종종 간접적이거나 각기 다른 컨텍스트 정보의 조각들을 종합해 추론할 수 있기 때문에 가장 어려운 일이 된다[11, 12].

Ⅲ. 시스템 설계

본 논문에서 제안하는 시스템은 사용자의 컨텍스트를 데이터마이닝 기법(연관규칙)으로 사용자의 구매(조회)성향이나 구매패턴을 파악하여 사용자의 프로파일을 생성하고, 사용자 이동경로 알고리즘을 이용하여 이동경로를 분석하여 개개인의 사용자에게 개인화 웹 페이지를 추천하는 것이다.

3.1 사용자 프로파일

사용자 프로파일은 기계 또는 객체가 사용자를 이해할 수 있도록 사용자와 그의 컨텍스트를 표현한 것이다. 초기 사용자 프로파일의 생성은 웹 사이트에 처음 사용자가 등록을 하게 되면 정적 컨텍스트인 성명, 성별, 나이, 직업 등을 이용한다. 이렇게 생성된 초기 프로파일과 이미 그룹화 되어 있는 그룹을 비교하여 그룹 중 가장 유사한 그룹을 찾아내어 사용자를 그 그룹으로 포함시킨다. 그리고 이후에 생성되는 구매이력과, 이동경로를 분석하여 사용자 프로파일을 갱신하게 된다.

표 2는 사용자 프로파일 생성과 갱신에 사용된 컨텍스트 항목들이다.

표 2. 사용된 컨텍스트 항목
Table. 2 Context Items Used

항목	내용
UID	PKNU2105
성명	황기만
성별	남자
직업	강사
나이	35
선호도	각 상품에 대한 선호도
구매이력	코드, 상품명 구매일, 상품가격
체류시간	각 상품에 대한 체류시간
이동경로	A-A-B-C-C-E-E-F-G-G-E-E

3.2 사용자의 이동패턴 분석 알고리즘

우선, 사용자의 이동경로에 가중치를 주기 위해 몇 초 단위로 사용자의 위치를 모니터링한다. 몇 초 단위로 지정해서 측정하는 이유는 웹 사용자 붐비는 경우 현재 웹 페이지에 관심이 있어 한 장소에 머무는 것이 아니고, 이동하지 못하는 경우를 고려해야 하기 때문이다. 보통 구매(조회)자들이 웹 페이지(상품)을 조회(구매)하기 위해서 머무는 시간이 설문조사에 의해 관심있는 상품일 경우 해당 제품에 대한 정보나 구매후기등을 읽기 위해 해당 페이지에 20초 이상인 것으로 나타났다. 일반적으로 naver.com의 첫 페이지에서는 5초정도, 소설사이트는 1~3분정도이고 웹보안사이트나 게임 사이트는 몇 분에서 몇 시간동안 머무는 것으로 나온다. 사이트 형태에 따라 머무는 시간은 다양하다. 본 논문에서는 학교 도서관의 웹 페이지 검색시 머문 시간이 20초를 넘어가게 되면 사용자가 그 위치에 있는 웹 페이지(상품)에 관심이 있는 것으로 판단하게 된다. 이와 같은 선호도에는 웹 페이지 조회 이력과는 별도의 '가중치'를 주게 된다. 지금 머문 웹 페이지에 관심이 없더라도 잠정적으로 구매(조회)를 할 확률이 높은 것으로 예측할 수 있기 때문이다.

이동패턴의 정확한 분석을 위해 사용자와 비슷한 성향을 가지는 그룹의 이동패턴을 묶어 '클러스터 데이터 세트'를 만들고 각 데이터 세트를 학습하여 '클러스터 모델'을 만든 후 새로운 데이터를 클러스터 모델로 평가해 사용자의 이동패턴을 예측하도록 하였다.

사용자 이동 중에 수집된 위치 데이터 값을 한 번의 이동을 나타낸다. 이렇게 분석된 이동패턴은 사용자의 현재 위치를 파악하여 선호도가 높은 웹 페이지들 중에서 사용자가 관심을 보이는 예측 이동경로에 가장 가까운 곳에 위치한 웹 페이지를 추천하게 된다.

이동경로를 분석하기 위한 알고리즘은 다음과 같다.

사용자의 웹 경로 이동을 20초 단위로 측정하여 경로 이동을 생성하고, 이동 중 중복된 페이지는 한 번만 포함시킨 경로 이동 생성한다. 이 경로이동은 사용자의 웹에서의 이동 패턴을 알려준다. 20초 이상 같은 페이지에 머물렀을 경우 그 페이지에 관심이 있는 것으로 판단한다. 해당하는 경로 이동만 추출하여 경로 이동을 생성한다. 이 경로 이동은 사용자가 웹에서의 관심 있는 페이지로 이루어진 경로 이동이다. 페이지가 중복되는 경우 중 20초 이상 머물지 않은 경우, 다른 페이지로 이동하는 경로 이동 중에 위치한 페이지로 간주한다. 그러나, 20초 이상 머물렀을 경우 그 페이지가 관심 있는 것으로 파악하여 그 페이지에 가중치를 증가시킨다.

생성된 최종 경로 이동에서, A'는 가중치가 한 번 부여된 페이지이고, E''는 가중치가 두 번 부여된 페이지이다. 따라서, 위의 최종 경로 이동의 결과를 가중치가 높은 순으로 웹 페이지를 개인화로 설계하면 각 개인이 선호하는 웹 페이지가 작성된다.

```
bool InIsFavoriteFunction = false ;
//제 한시간 이상 머물었는지를 검사하는 함수.
bool IsFavorite(UrlInfo* CurrentPage, list<UrlInfo*> &MyFavoritePath)
{
        int temp = 0 ;
        srand((unsigned int)time(NULL));
        temp=rand()%3; //몇초를 머물렀을건지를 임의로 지정
        string url = CurrentPage->GetUrl();
        clock_t begin = 0 , end = 0 , timeSec = 0 ;
        begin = clock() ; // 방문이 시작되었을때 시간을 저장

        switch(temp) {
case 0 :
        Sleep(1000 * 300) ; // 5분동안 방문이라 가정
        break ;
case 1 :
        Sleep(1000 * 180) ; // 3분동안 방문이라 가정
        break ;
case 2 :
        Sleep(1000 *20) ; // 20초동안 방문이라 가정
        break ;
        }
```

```
end = clock() ; // 방문이 끝났을 시간 저장

        timeSec = ((end-begin) / CLOCKS_PER_SEC) ; //몇 초를 방문했는지 계산

        //방문이 제한시간을 넘었을 경우 Favorite path에 사이트 URL 을 저장. 이미
        //저장된 URL 일 경우에는 가중치를 증가시킨다.
          if(timeSec >= THRESHOLD_SEC) {
                  InIsFavoriteFunction = true ;
                  if(IsValid(CurrentPage, MyFavoritePath)) {
                          MyFavoritePath.insert(MyFavoritePath.end(), CurrentPage) ;
                  } else CurrentPage->IncrementWeight() ;
                  InIsFavoriteFunction = false ;
                  return TRUE ;
          } else {
                  InIsFavoriteFunction = false ;
                  return FALSE ;
          }
}
```

그림 1. C++로 작성한 Procedure의 일부
Fig. 1. Procedure part that make out by C++

3.3 사용자의 프로파일 Update

사용자의 프로파일에는 사용자의 경로 이동, 체류시간, 같은 패턴에 속한 사용자들의 선호도를 이용한다. 1일 1번 업데이터하여 사용자 프로파일 데이터베이스에 저장한다.

사용자 프로파일 업데이터는 오랫동안 머문 페이지에서의 선호도를 적응시키기 위해 20초 이상 머문 페이지에 선호도를 0.1씩 증가하고 전체 가중치 단위를 네 부분으로 합계는 1이 넘지 않도록 한다($0 \leqq w \leqq 1$). 사용자의 경로이동을 True-FeedBack / False-FeedBack으로 구분한다.

최종적으로 프로파일에 추가 대상은 프로파일에 저장되어 있지 않은 페이지에 대한 체류시간이 발생하면, 이미 존재하는 페이지에 대해서는 기존의 선호도에 업데이터 값을 더한다. 업데이터 후의 선호도가 0 이하일 때에는 프로파일에서 페이지에 대한 선호도를 삭제한다. 대부분의 웹 브라우저는 이전에 요청되었던 페이지들을 캐쉬해 둔다. 그래서 사용자가 '뒤로'버튼을 누르게 되면, 캐쉬된 페이지가 보여지고 웹 서버는 반복된 페이지 요청을 인식하지 않게 된다. 이 과정에서 예외 처리가 발생한다. 또한, 데이터 분석은 상품 추천이 아닌 웹 페이지선호도 분석이다.

위의 내용을 수식으로 표현 한 것이 식 3과 식 4와 같다.

$$F_{u,p} = \sum_{p,s(+)} tr_{u,p} + \sum_{p,s(-)} fr_{u,p} \quad \text{.....식(3)}$$

$$F_{u,p} = \frac{F_{u,p}}{\max(F_{u,p}, p)}$$

$$r_{u,p} = r_{u,p} + F_{u,p}P \quad \text{.....식(4)}$$

각 식(3)과 식(4)의 u는 사용자, s(+)는 사용자가 긍정적 피드백을 준 페이지의 집합, s(-)는 사용자가 부정적 피드백을 준 페이지의 집합, p는 페이지, $tr_{u,p}$는 긍정적 피드백을 얻은 페이지의 가중치 그리고 $fr_{u,p}$는 부정적 피드백을 얻은 페이지의 가중치이다.

3.4 로그데이터 분석 과정

3.4.1 데이터 전처리

일반적으로 처음 저장된 데이터 수준의 로그만으로는 곧바로 유용한 정보가 되지 못한다. 먼저 일정한 정제

과정을 거쳐 산출된 데이터를 데이터베이스에 적재하게 되는데, 이러한 로그 데이터를 데이터베이스에 적재하기 전의 과정을 '전처리 과정(Pre- processing Process)'이라고 한다. 전처리 과정은 원본 로그파일을 정리하는 과정으로 원본 로그파일인 access_log를 그림 2와 같은 형식으로 변환한 후 access라는 이름의 파일로 작성하는 것이다.

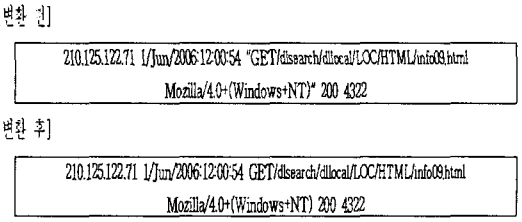


그림 2. 데이터 전처리 과정
Fig. 2. Data Pre-processing

다음의 과정은 그림 3과 같이 로그 파일에서 하나의 텍스트로 되어 있는 것을 각각 열로 나누기 위한 작업이다.

그림 4에서의 SQL문에서 보는 봐와 같이 이 작업에서는 불필요한 데이터 삭제와 분석에 용이한 파일로 전환시킨다. 로그파일에는 *JPG, *.GIF, *.MAP과 같은 파일을 요청하는 경우가 많다. HTML파일이 요청되는 경우 HTML문서의 태그에 포함되어 있는 *JPG, *.GIF, *.MAP 파일들이 자동으로 요청되기 때문이다.

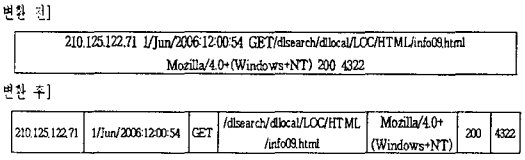


그림 3. 텍스트 나누는 작업
Fig. 3. Divide of Text

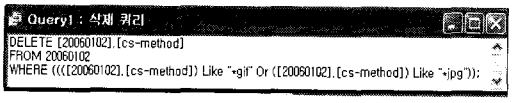


그림 4. 삭제 쿼리 화면
Fig. 4. Deletion Query Screen

이와 같은 파일이 요청된 기록은 삭제하고 에러가 난 경우에 대한 요청기록도 삭제한다.

클리닝 로그(Cleaning Log)과정을 하고나면 로그 사이즈는 보통 1/10에서 1/40정도로 축소된다. 엑셀에서 정제과정이 끝난 엑셀의 행의 수를 보면 65,536에서 19,997과 65,536에서 13,533의 행의 수가 준 것을 볼 수가 있다. 다음으로, access를 이용해 데이터베이스를 작성한다. 데이터베이스를 만들기 위해서 본 연구에서는 MS사의 ACCESS를 사용하였다. 처음 시스템 설계가 끝나면 클리닝 로그과정은 다음부터는 자동실행되게 된다. 각 테이블의 필드는 그림 5와 같다.

Num	IP	URL	Time
11	210.125.122.71	1/Jun/2006:12:02:15	33
12	210.125.122.71	1/Jun/2006:12:02:30	8

그림 5. ACCESS DB 필드
Fig. 5. Access Code Table

표 3은 웹 페이지의 실제 명칭을 분석가가 알기 쉽도록 관계코드로 매핑하여 분석을 용이하게 하였다. 이 관계코드는 각 개인이나 전체 사용자들의 웹이용을 분석하기 위한 웹 사이트간의 관계를 코드로 부여하여 항목(웹 노드)간의 관계를 표현하기 위한 것이다. 표 4에서의 지지도와 신뢰도를 계산하기 위한 기초 값이 된다.

표 3. 관계코드표
Table. 3 Relation Code Table

URL	Site Name	R-Code
/dlsearch/dllocal/LOC/search.html	전체검색	1
/dlsearch/dllocal/LOC/search1.html	단행본검색	7
/dlsearch/dllocal/LOC/new_book.html	신착도서 알림	8
/dlsearch/dllocal/info/info_person.html	개인정보관리	10
:	:	:
/dlsearch/dllocal/LOC/lib_menu.html	전자도서관 메뉴	32
/dlsearch/dllocal/LOC/lib_intro.html	전자도서관 소개	33
/dlsearch/dllocal/LOC/index.html	홈페이지 초기화면	34
/dlsearch/dllocal/info/info_info04.html	전체공지사항	35
:	:	:
/dlsearch/dllocal/lib/lib_intro.html	전자도서관 이용안내	45
/dlsearch/dllocal/lib/new_paper.html	도서관 관련 기사	46

3.4.2 연관성 규칙 분석(EC-Miner이용)

데이터 전처리가 끝난 로그 데이터는 마이닝 분석툴을 사용하여 연관성 규칙 분석을 수행하였다. 사용한 분석 프로그램은 EC-Miner로 조사하였다.

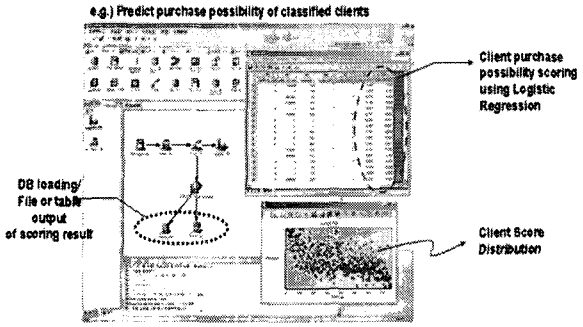


그림 6. EC-Miner의 활용 결과
Fig. 6. Result of Application of EC-Miner

그림6은 EC-Miner의 활용분야를 나타낸 그림이다. 고객의 Scoring이나 Biz-Rule생성, 고객 세분화, 상품 및 서비스 구매 간 연관성 분석된 결과를 DB에 적재하는 기능을 제공하며 마케팅 캠페인에 활용까지 가능하다. Logistic Regression을 이용한 고객 구매 가능성 Scoring과 Scoring결과의 DB적재 및 File or Table 출력, 고객별 스코어 분포까지 가능하므로 사용자 분석함에 있어 충분한 분석기라고 본다.

그림 7에서 사용자가 웹 페이지를 요청하면 웹 서버는 데이터베이스에 저장된 개체들을 마이닝 모듈에서 생성한 지식패턴에 따라 재구성하여 사용자에게 보여준다. 접속한 사용자, 시간대, 웹 페이지에 따라 적절히 대응하여 동적으로 웹 페이지를 생성하여 보여주는 것이다.

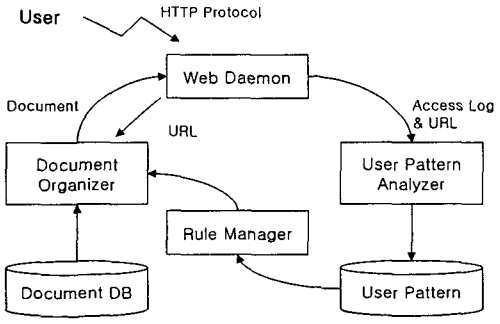


그림 7. 동적인 웹 서버 체제
Fig. 7. Dynamic Web Server System

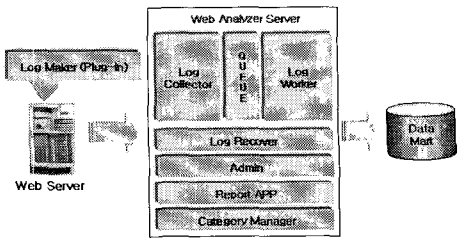


그림 8. 웹 데이타 분석
Fig. 8. Web Data Analysis

설명의 각 노드들 보면, Log Maker는 Web Server에서 로그 생성을 한다. Log Collector는 Log Maker에 의해 생성된 로그 수집을 하고 Log Worker는 로그 데이터를 loader를 이용하여 DB에 적재 및 분석을 한다. Admin은 환경설정 및 admin 기능, WebServer 정보 저장을 한다.

그림 8은 각종 금융, 포탈, 전자상거래, 유료 컨텐츠를 제공하는 e-Business를 위하여 로그파일을 분석하고, 고객별로 웹 사이트에 방문하여 보이는 행동과 패턴 그리고 관심도/선호도에 대한 리포팅을 작성하고 효율적인 처리를 하는 웹 분석을 하는 흐름도이다. 리포팅은 기업 및 조직에서 필요로하는 중요한 의사결정을 위한 분석 정보들을 자동으로 모아주고 결합시켜 원하는 정보를 적시에 제공한다. 또한 전략적 의사결정에 반드시 필요한 정보인 외부 데이터와 과거 데이터를 결합하여 입체적인 분석 정보를 제공하며, 전사차원에서 파악할 수 있는 모든 분석 정보를 제공하기 때문에 데이터 웨어하우스나 e데이터 웨어하우스 시스템을 구축함으로써 경영 또는 각 개인이 의사결정에 응용할 수 있다.

표 4. 2가지 항목의 연관성 규칙 결과
Table. 4 Result of Association Rule of Two Sorts of Items

Relations	Lift	Support(%)	Confidence(%)	TransactionCount	Rules
2	1.03	31.73	76.27	408	메인화면→소장자료검색/전체
2	1.7	25.67	35.37	307	소장자료검색/전체→키워드/분류검색
2	1.7	25.59	95.12	299	소장자료검색/전체→완전일치/분류검색
2	1.04	23.16	90.06	276	소장자료검색/분류검색→메인화면
2	1.04	23.14	25.12	263	메인화면→소장자료검색/분류검색
2	1.07	20.8	21.5	241	메인화면→커뮤니티
:	:	:	:	:	:

분석 과정은 먼저 접속자를 IP로 구분하고, 접속 유치 시간을 설정한 후 접속시간별로 이동 경로를 분석해 낸 뒤에 연관성 규칙을 조사하는 것이다. 연관성 규칙 분석 결과에서 지지도와 신뢰도는 일정 기준이상인 규칙들을 수집하여 유용한 연관성 규칙으로 선정하였다. 선정된 분석 결과들의 연관성 규칙, 지지도, 신뢰도와 향상도는 다음의 표4와 같다.

표 4의 규칙을 예로 들면, 사용자의 31.73%가 메인화면과 전체검색을 동시에 이용하고, 76.27%가 메인화면에서 소장자료검색 페이지의 전체검색으로 이동함을 보여준다. 그림 10은 위의 결과를 EC-Miner에서 나타낸 화면이다.

위의 결과들을 종합하면, 첫째, 해당 홈페이지를 방문하는 사람들의 주된 목적은 '소장자료검색/전체'로 이동하기 위함이다. 둘째, 접속횟수가 높은 '키워드' 또는 '분류검색' 등은 1~2회의 링크를 통해야 접속이 가능하다. 셋째, 사용자들의 이용 순으로 접속횟수로 판단이 가능하다.

표 4의 패턴을 이용하여 개인 또는 전체 이용자에게 향후 사용자들의 이용패턴을 분석해서 각 사용자들에게 맞는 웹 페이지를 설계하기 위한 항목(경로)들의 연관성 규칙결과이다.

3.4.3 시스템 분석

데이터 정제과정을 끝낸 데이터를 이용하여 EC-Miner를 가지고 웹 로그 분석을 하고자 한다. EC-Miner 측정을 이용하여 입력된 고객 데이터가 추출되는 단계를 나타내고, 그 결과를 고려하여 고객 데이터를 추출하는 과정을 분석, 기존의 시스템과 추천 시스템과의 비교 및 연관성 규칙에 대해 상품을 추출하는 것을 살펴보고자 한다.

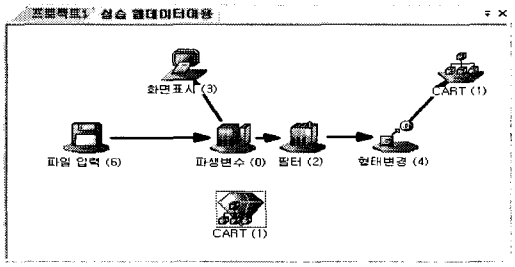


그림 9. EC-Miner의 Layout
Fig. 9. EC-Miner Layout

그림 9는 정제된 데이터를 이용해 결과를 나타낼 EC-Miner의 레이아웃 화면이다.

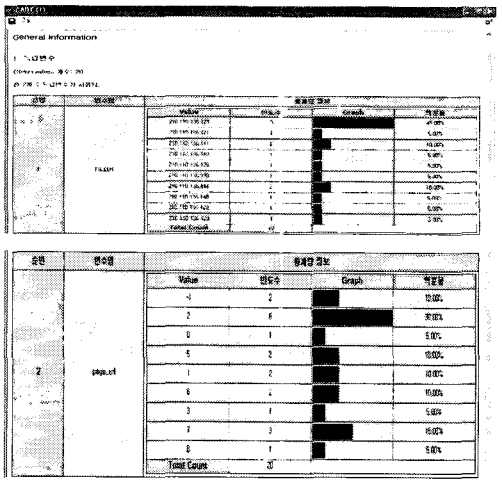


그림 10. EC-Miner의 수행결과
Fig. 10. Result of Implementation of EC-Miner

그림 10에 있는 'CART'노드는 정제된 데이터들을 EC-Miner에서 각각 노드 단계를 거쳐 최종적으로 나타난 결과 화면이다. 변수형태의 결과와 의사결정트리로 결과를 나타내고 있다. 결과 화면의 빈도수를 보면 각 이동페이지의 이용율을 보고 사용자의 이동패턴을 알 수가 있다. 그림 11은 각 패턴과 예측을 구분할 수 있는 것을 의사결정트리형태로 나타낸 결과 화면이다.

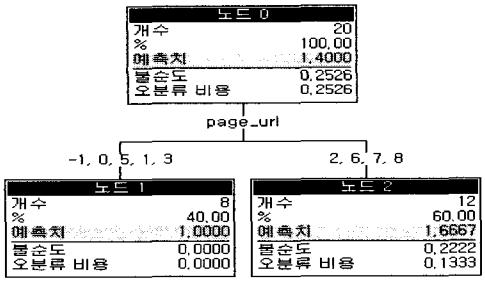


그림 11. EC-Miner의 의사결정트리
Fig. 11. Decision Tree of EC-Miner

3.4.4 시스템 결과

표 5. 각 연결 상태 비교
Table. 5 Each connection condition comparison

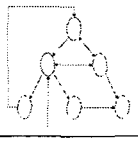
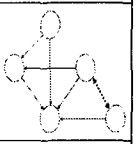
	일반경로	제안연관규칙 (Minsup→ maxSup)
이동 경로 형태		
이동경로 (T_ID : 1, 2, 3, 4, 5)	1. A→B→D 2. A→B→C→A→B→E 3. B→E→F 4. A→C→F→B→E 5. C→F→B→E	1. A→B 2. A→B→E 3. B→E→F 4. A→E 5. C→F→E
사용자 Click수	많음	보통(최적화)
사용자 이용시간	많음	보통(최적화)
웹 디자인 설계	복잡	간편
고객 사용 만족도	복잡	간편

표 5는 웹 이동경로를 기존의 일반 웹 이동경로, 연관규칙의 최소 지지도를 이용한 웹 이동 경로 그리고 기존의 연관규칙에 시차를 적용(제안한 연관규칙, 그림 12의 알고리즘 참조)한 최대 지지도를 이용한 경로 지정을 비교한 결과이다. 표 5의 이동 경로(T_ID : 1, 2, 3, 4, 5)의 각 경로를 보면 시차 연관규칙에서 웹 이동 경로가 최적화 되는 것을 볼 수 있다. 또한 표 5에서 보듯이 일반 연관 규칙보다도 제안한 연관 규칙이 고객 만족과 사용자의 마우스 클릭수를 줄 일수 있었고, 개인에게 맞는 웹 개인화 시스템 설계에서도 간편한 설계를 할 수 있다는 것을 알 수 있다.

SL : 시차 지지도
SL_i = {large i-itemsets};
for (k=2;$SL_{k-1} \neq \emptyset$;k++) do begin
  C_k=apriori_gen(SL_{k-1}); 후보 itemset 생성
  get_count(C_k)
  SL_k={c∈C_k | c.count ≥maxSup}
end
result = ∪ SL_k

그림 12. 시차 연관규칙을 이용한 maxSup발견하는 알고리즘
Fig.12. MaxSupport Algorism using A Squence Association Rule

그림 12는 itemsets이 최소 지지도 이상인 항목 집합에서 maxSup을 가진 itemsets을 발견하는 알고리즘이다.

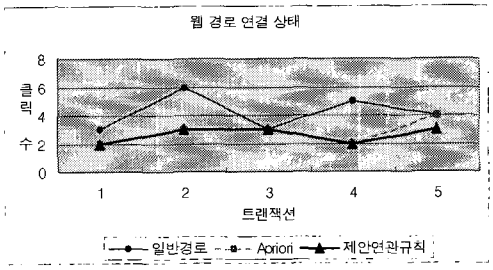


그림 13. 데이터셋에 따른 클릭 수
Fig. 13. The click possibility of folowing in data-set

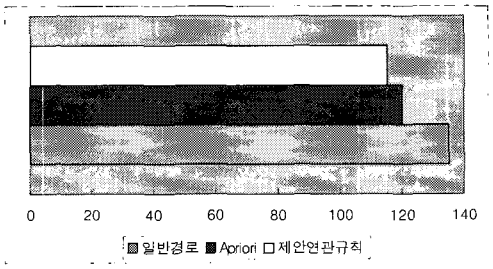


그림 14. 빈번한 항목 집합 검색 결과
Fig. 14. The item rathering search result which is frequent

그림 13에서 보듯이 각 웹 경로에서의 클릭 수는 큰 변화는 없지만 제안한 연관규칙에서는 Apriori연관규칙을 이용한 웹 경로보다는 최적화 되어진 것을 볼 수가 있다. 트랜잭션의 마지막 5단계에서 모든 웹 경로를 지원하는 단계에서 다른 경로를 지정한 것보다 개인화에 맞게 웹 경로를 지정할 수 가 있고 또한 전체 웹 경로 지정을 최적화 할 수 가 있다. 따라서 고객의 사용 만족도가 제안한 연관규칙을 이용해 작성된 웹 경로가 최적인 것이다.

그림 14의 결과에서 보듯이 웹 사용자가 웹을 이용하는 데 걸리는 시간에서 제안 연관규칙이 일반경로나 일반 연관규칙(Apriori)보다 좀 더 최적화된 것으로 나타났다. 그림 14는 세 가지 웹 경로 연결상태에서 찾은 빈번한 항목 집합의 평균 수를 보여준다. 제안 연관규칙에서는 115개의 다른 빈번한 k항목 집합을 찾고 Apriori연관규칙에서는 120개를 찾으며 일반경로에서는 135개를 찾는다. 제안 연관규칙과 Apriori연관규칙의 차이는 일반 연관규칙의 결과에 MaxSup(최대 지지도)값을 적용하였기 때문이다.

Ⅳ. 결론 및 향후 연구 과제

본 논문에 사용한 EC-Miner의 수행결과를 이용하여 접속자별로 페이지 간 이동과 접속 유지시간에 대한 연관성 규칙 분석 결과를 근거로 상호 접속빈도가 높은 페이지를 찾을 수 있다. 웹 관리자는 사용자들이 웹 게시판을 많이 이용하므로 커뮤니티 형성과 웹 모니터링을 적절하게 운영하는 것이 활용되어야 된다는 것이다. 그리고 최소한의 마우스 Click수로 페이지 이동이 가능하도록 페이지 링크를 걸어주며, 사용성을 향상시킬 수 있을 것이다. 결과적으로 제안한 시스템은 효율적인 웹사이트 설계(웹 서버관리, 웹 사이트 디자인 등)에 사용될 수 있을 것이다.

웹 구조를 개선한다는 것은 웹 문서간의 관계가 깊은 문서끼리 링크를 연결시켜주는 것이라 할 수 있다.

실험결과 제안된 연관규칙은 웹 경로를 찾아내는데 있어서 기존의 일반경로보다 빠른 측정 결과를 보여주지는 못하지만 반복적이고 빈번한 빈발항목을 찾는 데는 더 우수하기 때문에 반복적인 웹 이동경로를 개인화 설계할 때는 효과적이다. 마이닝의 대상이 되는 경로가 동일한 경로를 이용하거나 의미적으로 유사한 집합일 경우 웹 경로 개인화를 최적화하는 데 적용할 수 있다.

본 연구에 의해 도출된 분석을 이용하면 간단한 웹 페이지 뿐만 아니라, 많은 웹 페이지로 작성된 학교 홈페이지를 이용하는 데 학생들에게 효율적인 이용과 정보검색에 따른 시간 낭비를 줄 일 수 있을 것이다.

참고문헌

[1] U. Fayyad and R. Uthurusamy, eds., "Data Mining,," Special Issue, Communications of the ACM, 39(11), Nov, 1996.

[2] C Oosthuizen, J Wesson, C Cilliers, "Visual Web Mining of Organizational Web Sites," Proceedings of the Infomation Visualization (IV'06), 2006.

[3] R. Cooley, B. Mobasher, and J. Srivastava., Automatic Personalization Based on Web Usage Mining, Communications of ACM, No. 8, Vol. 43, 2000.

[4] B. Mobasher, N.Jain, E-H Han, and J.Srivastava, "Web Mining : Pattern Discovcovery from World Wide Web Transaction," Tech. Report, 1996.

[5] Bamshad. Mobasher, Robert Cooley, Jai-dpeep Srivastava, "Automatic Persona lization Based on Web Usage Mining," COMMUNICATIONS OF THE ACM, vol. 43, No. 8, Page 142-151, August 2000.

[6] 황정민, 승현우, "영화정보 시스템을 위한 데이 터 마이닝기법연구," J,Natural Science, SWINS, Vol. 13, p.79-86, 2001.

[7] 박원환, 박두순, "데이터마이닝에서 탐사범위 제한 수량연관규칙," 순천향산업기술연구소 논문집 제9권 1호, 2003.

[8] DE-XING WANG, XUE-GANG HU, HAO-WANG, "The Research on Model of Mining Association Rules based on Quantitative extended Concept Lattice," Proceedings of the First International Conference on Machine Learning and Cybernetics, Beijing, 4-5 November 2002.

[9] G. Piatetsky-Shapiro and W. Frawley, eds., "Knowledge Discovery in Databases," Menlo Park, CA, AAAI, 1991.

[10] 조일래, "영역 연관규칙 탐사를 위한 효율적 알고리즘," 한국해양정보통신학회논문지 제1권 제2호, 1997.

[11] 천인국, "컨텍스트 인식 기반 모바일 상품추천시스템 설계," 순천향산업기술연구소논문집 제8권 1호, 2002.

[12] 임재현, "Ubi-Class를 위한 컨텍스트 인식 시스템," 공주대학교 생산기술연구소 논문집 제12권, 2004.

[13] Cooley R., Mobasher B, and Srivastava J. 1998, "Web Mining : Information and Pattern Discovery om the World Wide Web," University of Minnesota.

저자소개

윤 종 찬(Jong-Chan Yun)

2003년 동명정보대학교 경영정보학과 졸업(경영학사)
2005년 부경대학교 대학원 전산정보학과졸업(공학석사)
2007년 부경대학교 대학원 전자 상거래시스템학과 박사과정수료
※관심분야 : 전자상거래, 데이터마이닝, 유비쿼터스, e-CRM 등

윤 성 대(Sung-Dae Youn)

1980년 경북대학교 컴퓨터공학과 졸업(공학사)
1984년 영남대학교 대학원 전자계산학과 졸업(공학석사)
1997년 부산대학교 대학원 전자계산학과 졸업(이학박사)
1989년 - 현재 부경대학교 전자컴퓨터정보통신공학부 교수
※관심분야: 병렬처리, 멀티캐스트통신, 데이터마이닝 등