웹 사용 정보 마이닝 기반의 동적 사용자 프로파일 생성

안 계 순[†] · 고 세 진[††] · 정 준[†] · 이 필 규[†††]

요 약

동적 웹 컨텐츠 제공에서 고객을 위한 추천서비스에 이르는 인터넷 기반의 전자상거래 애플리케이션에서는 고객이 어떤 성향을 가지고 있는 가에 대한 정보를 획득하는 것이 중요하다. 웹 개인화의 대표적인 기술인 협력적 여과는 사용자의 정보를 정적인 프로파일 형태로 저장하여 사용자의 성향 변화를 빨리 획득할 수 없다. 또한 사용자의 명시적 평가 의존성, 확장성 부족, 다차원 공간 데이터에 대한 적용 어려움 등의 문제점을 가지고 있다. 이와 같은 단점을 해결하기 위한 해결 방안으로 웹 사용 정보 마이닝(web usage mining)이 쓰이고 있다. 웹 사용 정보 마이닝은 서버에 축적된 웹 사용 데이터(web usage data)를 이용하여 패턴을 발견하는 기술이다. 특히 연관 규칙 생성 알고리즘으로 웹 사용 패턴(web usage pattern)을 찾고 패턴을 클러스터링하는 기술이 사용되고 있다. 그러나 연관 규칙 생성 알고리즘은 많은 수의 패턴들을 찾고 또 유용하지 못한 패턴을 발견하는 단점이 있다. 본 논문에서는 검증된 웹 사용 패턴을 이용한 동적 사용자 프로파일 생성 방법을 제안한다. 먼저 패턴 발견을 위해 연관 규칙 생성 알고리즘인 Apriori를 이용하고 사용자 프로파일을 위한 클러스터를 생성하기 위해 ARHP를 채택하였다. 클러스터를 생성하기 전에 Dempster-Shafer 이론을 이용하여 유용하지 못한 패턴을 제거하는 패턴 검증 과정을 수행한다. 검증된 패턴을 이용하여 클러스터를 생성하고 사용자의 현재 활성화된 세션에 따라 동적으로 사용자 프로파일이 생성된다.

Generator of Dynamic User Profiles Based on Web Usage Mining

Kye-Sun An[†] · Se-Jin Go[††] · Jun jiong[†] · Phill-Kyu Rhee[†††]

ABSTRACT

It is important that acquire information about if customer has some habit in electronic commerce application of internet base that led in recommendation service for customer in dynamic web contents supply. Collaborative filtering that has been used as a standard approach to Web personalization can not get rapidly user's preference change due to static user profiles and has shortcomings such as reliance on user ratings, lack of scalability, and poor performance in the high-dimensional data. In order to overcome this drawbacks, Web usage mining has been prevalent. Web usage mining is a technique that discovers patterns from We usage data logged to server. Specially, a technique that discovers Web usage patterns and clusters patterns is used. However, the discovery of patterns using Apriori algorithm creates many useless patterns. In this paper, the enhanced method for the construction of dynamic user profiles using validated Web usage patterns is proposed. First, to discover patterns Apriori is used and in order to create clusters for user profiles, ARHP algorithm is chosen. Before creating clusters using discovered patterns, validation that removes useless patterns by Dempster-Shafer theory is performed. And user profiles are created dynamically based on current user sessions for Web personalization.

키워드 : 웹 사용 마이닝(Web Usage Mining), 개인화(Personalization), 추천 시스템(Recommendation System), 패턴 탐색(Pattern Discovery), 지지도(support), 신뢰도(Confidence)

1. 서 론

인터넷의 대중화와 컨텐츠의 다양화로 웹은 더 이상 단순히 정보를 찾기 위한 수단이 아니며 비즈니스를 위한 매개체로 이용되어지고 있으며 삶의 일부분으로 정착되어지고 있다. 따라서 인터넷 서비스 제공자는 웹을 통하여 고객의 정보를 습득하고 분석하여 고객에게 최적의 개인화된 상품 및 서비스를 제공하여 고객의 충성도를 확보하려하고 있으며 사용자는 보다 개인적인 정보를 원하고 있다. 위와 같은 목적을 위해 인공지능 및 데이터 마이닝 기법에 기반한 개인화 기술에 대한 연구가 활발히 진행되고 있다.

이러한 개인화 서비스는 사용자에 대한 명확한 인식에서부터 시작된다. 즉 사용자가 누구이며 사용자는 어떤 행동을 하는 가에 대한 정보 구축이 선행되어져야 하며 이러한 정보를 사용자 프로파일 형태로 저장한다. 사용자 프로파일은인류 통계학적 정보(나이, 성별, 주소 등), 일정 기간 동안 사용자가 행한 정보, 특정 group에 속한다는 정보, 사용자가 가지는 규칙 또는 패턴 등을 가지고 있다[3]. 그러나 위와 같은 사용자 프로파일은 개인화 서비스를 위해 한번 작성된 후 일정 시간이 지난 후에 업데이트가 되므로 사용자의 성향의 변화를 빠르게 탐지하지 못하는 정적 프로파일이다. 따라서 사용자의 현 상태에 따라 동적으로 생성이 되는 프로파일이 필요하다.

† 준 회 원 : 인하대학교 대학원 전자계산공학과
†† 정 회 원 : 인하대학교 대학원 전자계산공학과
††† 종신회원 : 인하대학교 전자계산공학과 교수
논문접수 : 2002년 5월 7일, 심사완료 : 2002년 7월 5일

정적 사용자 프로파일을 적용하여 개인화 서비스를 제공하는 협력적 여과는 웹 서비스에 대한 사용자의 평가 값을 명시적으로 받아들이고, 사용자간 상관 엔진(correlation engine)을 통해 사용자의 선호도와 유사하게 일치하는 정보를 제공한다[4]. 그러나 이 기술은 아이템 수가 많을 경우 알고리즘의 확장성에 문제가 있으며 사용자의 명시적인 피드백 정보를 요구해 사용자의 불편을 초래하고 cold-start 문제점이 있다[9].

협력적 여과의 문제점을 해결하기 위해 다양한 방법이 연구되어지고 있다. 특히 웹 사용 정보 마이닝[4, 5]은 일반적인 데이터 마이닝 방법을 웹 도메인에 적용하여 웹 사용 패턴을 찾아내고 이를 바탕으로 개인화 된 정보를 사용자에게 전달하는 기술이 활발히 연구되어지고 있다. 이는 사용자의 명시적인 피드백을 필요로 하지 않으며 cold-start 문제를 해결할 수 있는 기술로 인정받고 있다. 이러한 웹 사용 마이닝에 기반한 개인화 서비스의 장점 때문에 많은 사이트에서 마이닝 기술을 채택하여 개인화 서비스를 시도하고 있다. 특히, 연관 규칙 생성을 통한 웹 오브젝트간의 연관성을 고려한 개인화 서비스를 시도하는 사이트가 점차 많아지고 있다. 하지만, 현재의 연관 규칙 생성을 통한 개인화 서비스를 시도하려는 사이트에서는 웹 사이트 전체 사용자의 웹 접근 정보를 웹 오브젝트 연관 규칙 발견을 위한 정보원으로 사용한다. 이러한 방법은 각 사용자와 유사한 사용자들의 웹 접근 정보를 이용하여 사용자 프로파일을 통한 개인화 서비스보다 그 성능이 저하되고 진정한 의미의 개인화 서비스라고 정의할 수 없다. 따라서 효과적인 개인화 서비스를 위해서 개선된 알고리즘을 통한 개인화 서비스를 제공해야 한다.

웹 사용 정보 마이닝은 사용자가 웹을 항해할 때 생성되는 데이터로부터 유용한 정보를 발견하는 것이다. 이것은 사용자가 웹 서버와의 상호작용을 하는 동안에 사용자 행위를 예측할 수 있는 기법들에 초점을 맞추고 있다. 웹 사용 정보 마이닝의 데이터 전처리 부분에서 웹 문서 내용과 웹 사이트 구성 정보는 정보 자료로 사용될 수 있고, 웹 사용 정보 마이닝은 웹 내용 정보 마이닝과 웹 구조 정보 마이닝과의 상호 작용을 한다. 또한 패턴 탐색 결정 과정에서 클러스터링은 웹 사용 정보 마이닝으로부터 웹 내용 정보 마이닝과 웹 구조 정보 마이닝의 연결 역할을 한다.

웹 사용 정보 마이닝을 위해서 사용되는 입력자료는 사용자들의 접근 기록이 저장되어 있는 웹 서버의 로그 파일이다[13]. 웹 서버 로그 파일의 포맷은 여러 가지가 있지만, 일반적으로 로그 파일에는 접속한 사용자의 IP 주소, 요청 시각, 요청 방법, 요청한 웹 문서의 URL, 상태 정보 등이 기록되어 있다. 로그 파일에 대한 분석 도구들은 많이 존재하고 있지만, 대부분의 분석도구들은 사용자들이 웹 서버에 가장 많이 접근하는 시간대나 가장 많이 요청되는 웹 문서들에 대한 분석 등 단순한 통계적인 분석에 그칠 뿐, 사용자들의 브라우징 패턴에 대한 분석은 고려하지 않고 있다[10]. 이러한 웹 로그 파일을 이용하여 웹 사이트 디자인 같은 디자인 작업에 이용될 수도 있고, 웹 서버 디자인, 웹 사이트를 이용하는 사용자의 탐색하는 방법을 생성할 수 있다. 데이터 마이닝 알고리즘을 적용하기 전에 원시 데이터(raw data)로부터 처리를 위한 데이터 추상화를 위한 변형 작업을 해야 한다.

사용자의 패턴을 발견하는 단계는 웹 마이닝의 핵심 부분이다. 패턴 발견은 데이터 마이닝, 기계학습, 통계학, 패턴 인식 등과 같은 여러 분야에 사용하는 알고리즘이나 기술들을 이용한다. 데이터 마이닝에서 사용하는 기법들은 연관 규칙, 순차 패턴, 클러스터링, 분류 등이 있다

패턴 분석은 웹 사용 정보 마이닝의 최종 단계이다. 패턴 분석 처리의 목표는 관련 없는 규칙을 제거하고 흥미로운 규칙이나 패턴 발견 단계의 산출물로부터 패턴을 추출하는 것이다. 웹 마이닝 알고리즘의 산출물은 사용자가 직접 사용하기에 적절한 형식이 아니기 때문에 사용자가 쉽게 알 수 있는 형식으로 변형시켜야 한다,

본 논문에서는 검증된 웹 사용 패턴을 이용한 동적 사용자 프로파일 생성 방법을 제시한다. 먼저 연관 규칙 생성 방법을 이용하여 웹 페이지로 이루어진 빈발항목집합을 찾아내고 ARHP(Association Rule Hypergraph Partitioning) 알고리즘을 이용하여 웹 페이지 클러스터를 생성한다[6]. 생성된 각 클러스터는 유사한 사용자들이 접근한 웹 페이지로 구성되어지므로 사용자가 웹사이트에 현재 접근한 패턴과 클러스터 내 정보와 일치할 경우 실시간으로 패턴으로 이루어진 사용자 프로파일이 생성된다. 그러나 연관 규칙 생성 방법을 이용한 빈발항목집합은 사이트 전문가가 이미 알고 있거나 흥미롭지 못한 경우가 많아 개인화 서비스의 성능을 저하시킨다. 따라서 통계적 추론 방법인 Dempster-Shafer[8, 14]를 이용하여 빈발항목집합의 유용성 여부를 검증하여 유용한 집합만을 ARHP에 적용하고 코사인 유사도 측정 방법을 이용해 동적으로 사용자 프로파일을 생성하여 웹 개인화 시스템에 적용함으로써 개인화 성능을 향상시킬 수 있다.

본 논문의 구성은 다음과 같다. 제 2장에서는 웹 사용 마이닝 설명하고 제 2장에서는 웹 사용 패턴을 이용한 클러스터링 알고리즘을 소개한다, 3장에서는 패턴의 검증을 위한 Dempster-Shafer 이론에 대해 기술한다. 제 4장에서는 생성되어진 클러스터를 이용한 사용자 프로파일 생성에 대해 기술하고 제 5장에서는 실험 평가 기준 제시 및 성능을 평가하고, 제 6장에서 결론을 기술한다.

2. 웹 사용 패턴 클러스트링

2.1 연관 규칙 생성

본 논문에서 웹 로그 파일을 전처리(preprocessing)하여 얻어진 트랜잭션 파일을 이용하여 각 익명의 사용자가 접

근한 페이지간의 연관 규칙 생성을 위해 Apriori 알고리즘 [2]을 이용하였다. 아래 (그림 1)은 전처리 되어진 트랜잭션 파일을 보여준다.

```
익명 사용자 1 트랜잭션 :
A.html F.html D.html E.html G.html H.html D.html

익명 사용자 2 트랜잭션 :
A.html D.html C.html F.html

익명 사용자 3 트랜잭션 :
E.html D.html F.html G.html S.html

익명 사용자 4 트랜잭션 :
H.html D.html F.html J.html R.html
                    .
                    .
                    .
익명 사용자 n 트랜잭션 :
A.html F.html G.html J.html R.html
```

(그림 1) 전처리후 생성된 트랜잭션 화일

연관 규칙이란 "어떤 사건이 일어나면 다른 사건이 일어난다"와 같은 연관성을 나타낸다. 주어진 트랜잭션 집합이 있고 X와 Y를 트랜잭션에 속하는 페이지 즉, 아이템의 집합이라고 하면, 연관 규칙(association rule)은 $R: X \rightarrow Y$ 형식의 함축이고, 이때 X와 Y는 서로 같은 오브젝트를 갖지 않는 항목집합이다. 즉, 트랜잭션 집합을 I라고 표현했을 경우, $X, Y \subseteq I$, $X \cap Y = \varnothing$, $Y \neq \varnothing$를 만족해야 한다. X를 규칙의 조건부(antecedent)라 하고 Y를 결과부(consequent)라 한다. 만일 한 트랜잭션이 X를 지지한다면, 또한 어떤 확률에 의해 Y도 지지할 것이라는 예측으로 이해될 수 있는 것이 연관 규칙이다. 생성된 연관 규칙이 트랜잭션들의 상황을 얼마나 잘 뒷받침해 주는가는 다음의 두 가지의 척도로서 측정한다.

- 지지도(support degree) : 생성된 연관 규칙이 전체 트랜잭션에서 차지하는 비율을 말한다. 즉, 트랜잭션 화일에 속한 전체 트랜잭션의 개수중 그 연관 규칙을 지지하는 트랜잭션의 비율을 의미한다.
- 신뢰도(confidence degree) : 연관 규칙의 강도를 의미하며 조건부를 만족하는 트랜잭션의 결과부 까지를 만족하는 비율을 말한다.

이러한 신뢰도는 조건 x의 경우 결과 y의 규칙에 대한 정확도를 측정할 수 있는 지표가 된다.

결론적으로 어떤 규칙의 신뢰도는 얼마나 조건부에 대해 결과부가 자주 적용될 수 있는지를 나타내고, 반면에 지지도는 그 규칙의 전부가 얼마나 믿을만한지를 나타낸다고 말할 수 있다. 그러므로, 규칙이 트랜잭션 파일 안에서 적절해지려면 충분한 지지도와 신뢰도를 가져야 한다.

2.2 ARHP(Association Rule Hypergraph Partitioning)

웹 개인화를 위해 집합적인 웹 사용 패턴으로 이루어진 클러스터 생성에 관련한 연구가 진행 중이다[11, 12]. 그러나 [11, 12]들은 클러스터 생성 자체와 관련한 프레임워크만 제시할 뿐 웹 개인화 적용에 확장이 어렵다[4]. 따라서 본 논문에서는 B. Mobasher가 제시한 집합적 웹 사용 패턴을 이용한 클러스터 생성 및 웹 개인화 적용 방법을 적용하였다[4].

먼저 전처리 되어진 트랜잭션 파일은 클러스터링 기법을 적용하기 위해 n-차원 공간의 벡터로 이루어지며 각 벡터는 가중치를 갖게되며 식 (1)처럼 표현되어진다.

$$t = w(p_1, t), w(p_2, t), \cdots, w(p_n, t) \quad (1)$$

t : 익명 사용자 트랜잭션, w : 가중치, p : page

가중치는 다양한 방법으로 얻을 수 있으며 본 논문에서는 해당 페이지의 존재 여부에 따라 1 또는 0 값을 할당하였다. 그리고 ARHP 알고리즘을 이용하여 생성될 클러스터 역시 n-차원의 벡터로 표현되어 진다.

$$C = W_1^c, W_2^c, \cdots, W_n^c \quad (2)$$

$$W_i^c = weight(p_i, C), \text{ if } p_i \in C. \ otherwise\ 0$$

ARHP 알고리즘은 연관 규칙과 Hypergraph partitioning을 이용하여 트랜잭션 기반의 데이터베이스에서 연관된 항목들을 클러스터링하는 방법이다[6]. 먼저 연관 규칙 탐사 방법인 Apriori 알고리즘을 이용하여 트랜잭션들 안에서 빈번하게 동시에 출현하는 항목들의 집합을 찾는다. 그리고 Hypergraph Partitioning 알고리즘[6]은 항목 클러스터를 찾기 위해 사용된다. 빈번한 항목 집합에 의해 항목들간의 유사도가 계산된다. Hypergraph $H = (V, E)$는 문서들로 구성된 정점(vertex)들의 집합 V와 빈번한 항목 집합들을 나타내는 hyperedge들의 집합 E로 구성된다. Hypergraph partitioning 알고리즘은 빈발항목집합의 가중치로 빈발항목집합으로 생성될 모든 연관규칙의 신뢰도의 평균을 적용한다[7].

생성된 Hypergraph H는 다단계의 과정을 걸쳐서 클러스터 집합을 만들어 낸다. Hypergraph를 축소하는 단계(Coarsening Phase), 최소화된 Hypergraph를 partitioning하는 단계(Partitioning Phase), partitioning한 Hyperedge를 원래 크기의 Hypergraph로 확대하는 단계(UnCoarsening Phase)를 반복하면서 클러스터링 한다[6]. 특히 두 번째 단계에서는 클러스터 내의 각 점들 간의 연결 정도를 계산하여 연결 정도가 특정 임계값(threshold) 값보다 적은 점은 클러스터에서 제거되며 클러스터(c) 내 한 점(v)의 연결 정도(connectivity)는 아래 식 (3)에 의해 결정된다. 각 클러스터 내의 페이지들은 connectivity 값을 가중치로 표현되는 n-차원의 벡터로 나타내어진다.

$$conn(v, c) = \frac{\sum_{e \subseteq c \wedge v \subseteq e} weight(e)}{\sum_{e \subseteq c} weight(e)} \quad (3)$$

e = Hyperedge, weight = 신뢰도

3. 웹 사용 패턴 검증

앞 절 2.1에서 언급한 연관 규칙 생성 알고리즘인 Apriori는 전체 트랜잭션에서 공통적으로 발생하는 규칙 즉 패턴을 찾게 해주는 유용한 알고리즘이며 널리 이용되고 있다.

그러나 Apriori 알고리즘은 많은 수의 패턴 탐색 결과를 가져올 뿐만 아니라 의사 결정에 참여한 도메인 분석가에게 이미 알려진 패턴이거나 불필요한 패턴, 즉 흥미롭지 못한 탐색결과를 가져온다. 이 문제점은 발견된 패턴을 입력 데이터로 이용하는 ARHP에서는 이미 알려져 있거나 흥미롭지 못하며 관련이 없는 웹 페이지로 구성된 클러스터를 생성할 수 있다. 이는 이 클러스터를 바탕으로 사용자 프로파일을 구성하여 개인화 서비스를 제공할 경우 사용자의 추천 만족도를 저하시킬 수 있다. 따라서 클러스터 생성 전에 패턴 검증 과정이 필요하다. 본 절에서는 패턴의 검증을 위한 패턴의 유용성 측정 기준 및 검증 알고리즘을 제시한다.

3.1 Measure of Interestingness

데이터 마이닝 분야에서 패턴의 유용성 여부를 언급하기 시작했으며 그 결과로 Interestingness[1, 14]라는 개념을 도입하였으며 이를 결정하는 척도 두 가지를 정의하였다. 즉, 객관적 척도(Objective Measures)와 주관적 척도(Subjective Measures)를 정의하였다. 전자는 패턴 발견 프로세스에서 사용되는 데이터와 그 구조에 의해 측정되어지며 대표적인 객관적인 척도는 패턴의 지지도(support), 신뢰도(confidence)[1, 14] 등이다. 이 척도들은 주로 발견되어지는 패턴의 수를 감소시키는 역할을 하며 패턴에 대한 명시적인 interestingness를 위한 평가 기준을 제시하지 못한다. 반면 후자는 정의되어진 믿음(Belief)를 이용하여 패턴을 평가하는 기준으로 제시된다.

패턴을 이루는 데이터와 그 구조를 이용할 뿐만 아니라 정의되어진 믿음(Belief)을 이용하는 패턴 검증의 척도로 본 논문에서는 주관적 척도[1, 14]를 이용한다. 주관적 척도는 다시 두 가지 요소로 나뉘어 진다.

1. Unexpectedness : 주관적인 척도로서 발견되어진 패턴이 전문가에게 놀란만한 사실을 전해주는 정도.
2. Actionability : 발견되어진 패턴이 전문가로 하여금 특정 행위를 하여 이익을 얻게 하는 정도.

Unexpectedness와 Actionability 모두 패턴의 interestingness를 결정하는 중요한 요소이나 Actionability는 개념적인 요소로서 실제로 사용되기는 어렵다. 두 요소는 상호 배타적이 아니며 일반적으로 Unexpectedness 패턴은 Actionability라하며 그 역도 성립하는 것으로 받아들여진다[33]. 따라서 추출된 패턴의 Unexpectedness를 패턴 분석의 주관적 척도의 개념적 요소로 받아들이고, 곧 패턴의 interestingness를 식별하기 위한 척도로 사용한다.

패턴 분석가가 이미 알고 있는 믿음은 그 믿음을 신뢰할 수 있는 확률값, 즉 신뢰도(degree of belief)가 필요하며 신뢰도를 결정하기 위한 증거(evidence)가 요구된다. 즉 모든 믿음에는 어떤 증거가 존재하며 그 증거를 근거하여 신뢰도를 알 수 있게 된다. 만약 아이템 연관성에 대한 새로운 증거들이 믿음에 적용되어 믿음 값의 변화가 생기면 새로운 증거가 유용한 패턴이 된다.

3.2 Dempster-Shafer 이론

앞 절에 설명한 패턴의 유용성 여부는 패턴의 신뢰도(degree of belief) 변화 여부에 따라 결정된다. 따라서 주어진 증거를 이용하여 가능성에 근거한 패턴의 유용성을 찾기 위한 통계적 추론 방법을 사용한 모델의 구성이 필요하다.

통계적 추론의 대표적인 예가 베이지안 이론(Bayesian Theory)이지만 Apriori 알고리즘으로 생성된 빈발 항목 집합일 경우에 적용하기가 어렵다. 따라서 가설 집단의 모든 부분집합들을 하나의 가설로 취급하여 믿음(Belief)에 대한 통계적인 추론 과정을 수행하는 모델인 Dempster-Shafer 이론[8]이 이에 해당한다.

Dempster-Shafer 이론은 가설 집단(frame of discerment)의 모든 부분집합들을 하나의 가설로 인식하게 할 수 있으며 이 가설들에 신뢰도를 배정할 수 있다. 먼저 상호 배타적인 단일 가설들로 구성이된 유한 집합 $\odot$가 주어질 때 $2^{\odot}$ $\odot$로 구성된 집합 Λ을 구할 수 있다. 예를 들면 가설 집단은 4개의 입체 모형들로 구성되어 있다.

$$\odot = \{ cube, cylinder, shpher, prism \}$$
$$\Lambda = 2^{\odot} = \{ \varnothing, cube, cylinder, sphere, prism, \{ cube, cylinder \}, \cdots, \{ cube, cylinder, sphere \}, \cdots \{ cube, cylinder, sphere, prism \} \}$$

만약 베이지안 이론을 적용할 경우 집합 $\odot$의 각각 하나의 원소에만 신뢰도 값을 부여할 수 있지만 Dempster-Shafer 이론을 적용하면 복수개의 원소로 구성된 가설에 신뢰도 값을 적용할 수 있다[8].

집합 Λ의 원소들은 집합 $\odot$와 달리 복수개의 원소로 구성되고 기본 확률값(basic probability) m을 가지며 식 (6)을 만족한다.

$$\sum_{A \subset \Lambda} m(A) = m(\oplus) + m(cube) + \cdots = 1 \qquad (6)$$

집합 Λ의 원소들은 가설, 즉 발생할 사건으로 볼 수 있으며 각 사건이 발생할 수 있는 가능성을 나타내는 신뢰도는 아래 식 (7)로 구해진다.

$$Belief(A) = \sum_{\forall B;\, A \subseteq B} m(B) \qquad (7)$$

사건 A에 대한 신뢰도는 A를 포함하는 집합 Λ의 원소들의 m값을 모두 합한 값이 된다.

Dempster-Shafer 이론은 특정 사건에 대한 신뢰도를 구할 수 있을 뿐만 아니라 신뢰구간(interval)을 추정할 수 있으며 그 구간은 아래 식 (8)과 같다.

$$\bigcap val = [Belief, Plausibility] \tag{8}$$

식 (10)에서 *Plausibility*는 사건 A에 대해 알려지지 않은 증거들이 존재한다면 그 사건이 일어날 신뢰도이며 식 (9)로 계산된다.

$$Plausibility(A) = 1 - Belief(\neg A) = 1 - \sum_{\forall B\,;\,ANSUBSETB} m(B) \tag{9}$$

식 (8), 식 (9)에서 계산한 Belief와 Plausibility는 식 (10)와 같은 성질을 만족한다.

$$Belief(A) \le Plausibility(A) \tag{10}$$

이처럼 Dempster-Shafer 이론은 불확실성을 베이지안 이론처럼 하나의 수치로 나타내기보다는 신뢰구간을 지정하여 유효한 표현이 가능하게 한다.

가설 집단(frame of discernment)의 각 원소를 복수 개로 갖는 가설은 기본 확률값(basic probability) m을 갖게 되며 단일한 증거 공간으로부터 얻게 된다. 만약 빈발항목집합들의 신뢰도를 계산을 할 경우 기본 확률 m은 웹 사이트의 구조적 정보로 계산할 수 있다. 빈발항목집합의 연관성에 대한 신뢰 정도는 그 빈발항목집합의 각 항목들이 웹 사이트 구조상 서로 얼마나 링크가 형성이 되어 있는 가를 통해 얻을 수 있다. 또한 빈발항목집합의 아이템 간의 연관성은 연관 규칙 생성 과정에서 사용된 지지도와 신뢰도를 이용하여 기본 확률을 구할 수 있다. 즉 하나의 빈발항목집합의 아이템 연관성에 대한 신뢰도를 추정하기 위해서는 하나의 증거 공간만을 사용하지 않고 다양한 증거공간을 통해서 더욱더 정확한 추정을 할 수 있다[8].

Dempster-Shafer 이론은 위에서 언급한 다양한 증거 공간(multiple source)로부터 각각의 서로 다른 증거 공간에서 추정된 기본 확률값을 병합할 수 있게 한다. 아래 식 (11)은 서로 다른 기본 확률값이 병합이 되어 새로운 기본 확률값이 계산됨을 보인다.

$$m_1 \oplus m_2 = \frac{\sum_{\forall B,C\,;\,B\cap C=A} m_1(B)m_2(C)}{1 - \sum_{\forall B,C\,;\,B\cap C=\varnothing} m_1(B)m_2(C)} \tag{11}$$

따라서 식 (11)에서 볼 수 있듯이 가설 A에 대한 m_1으로 이루어진 신뢰함수 $Belief_1(A)$와 m_2로 이루어진 신뢰함수 $Belief_2(A)$가 병합하여 새로운 신뢰함수 $Belief_{m_1 \oplus m_2}(A)$이 생성된다. 신뢰함수의 병합은 병합이전의 신뢰도와 비교를 통해 추정값의 정확성을 얻을 수 있다.

3.3 빈발항목집합의 검증

Apriori로 생성된 빈발항목집합을 ARHP의 입력데이터로 여과없이 사용할 경우 클러스터의 수가 많아지며 뿐만이 아니라 각 클러스터 내의 웹 사용 패턴은 유용하지 못해 개인화 서비스의 질을 저하시킨다. 따라서 찾아진 빈발항목집합에 대한 유용성 검증을 통해 유용한 웹 사용 패턴을 가진 클러스터를 생성하여 개인화의 효율을 높이는 것이 중요하다.

본 논문에서는 3.1에서 언급한 패턴의 유용성 여부를 결정하기 위해 패턴의 interestingness를 적용한다. 빈발항목집합에 대한 신뢰도(degree of belief)가 임계값(threshold) 이상인 빈발항목집합으로 선택하며 신뢰도 추정을 위해 Bayesian 이론의 단점을 해결한 Dempster-Shafer 이론을 적용한다. 이 이론은 빈발항목집합으로 구성된 가설에 대한 신뢰도를 추정하기에 적합하며 웹 사이트의 구조 정보와 웹 사용 정보를 이용하여 신뢰함수 병합을 수행할 수 있다.

먼저 가설 집단(frame of discerment) $\odot$는 전처리 되어진 익명 사용자 트랜잭션을 이루는 전체 페이지이며 기본 확률값이 배정될 $\wedge = 2^{\odot}$인 집합은 빈발항목집합 i를 원소로 갖는 집합 I으로 구성이 되며 i의 원소의 개수가 하나인 i는제외된다. 각각의 빈발항목집합 i는 Dempster-Shafer 이론에 따라 불확실성을 가진 발생 가능한 가설 또는 사건이 된다.

빈발항목집합 i의 기본 확률값 m은 웹 구조 정보로 얻어진 m_s와 웹 사용 정보로 얻어진 m_u 두 가지를 갖는다. m_s는 빈발항목집합 i가 현 사이트 구조 정보, 즉 빈발항목집합을 이루는 페이지간의 링크 수를 통해 m_s를 계산할 수 있다. m_s의 계산 방법은[14]이 제시한 *lfactor*와 *cfactor*를 적용하며 아래 식 (12)와 식 (13)을 통해 구한다.

$$lfactor = \frac{L}{N(N-1)} \tag{12}$$

$$cfactor = \begin{cases} 1, & \text{if } G(i) \text{ is connected} \\ 0, & \text{otherwise} \end{cases} \tag{13}$$

*lfactor*에서 N은 빈발항목집합 i의 총 페이지 수이며 L은 각 페이지 간에 직접 연결된 링크의 수이다. 그리고 *cfactor*의 $G(i)$는 i의 페이지가 그래프를 형성하는 가에 대한 여부를 나타낸다. 이 두 값 *lfactor*, *cfactor*을 이용하여 m_s를 구한다.

$$m_s = lfactor * cfactor \tag{14}$$

웹 사이트의 구조 정보를 이용한 m_s뿐만이 아니라 전체 트랜잭션에서 해당 빈발항목집합의 빈도수를 이용하여 m_u를 계산할 수 있다. [14]는 빈발항목집합 자체의 지지도와 신뢰도보다는 좀 더 정확한 기본 확률값을 구하기 위해 커

버리지(coverage) 개념을 이용하였다. 커버리지는 지지도와 달리 전체 트랜잭션에서 빈발항목집합의 페이지들이 적어도 한 번 이상 나타나는 트랜잭션의 수를 적용한 것이다. 본 논문에서도 [14]이 제시한 커버리지와 지지도를 이용해 m_u 값을 구한다.

$$Support = \frac{Count(page_1 \wedge page_2 \cdots \wedge page_n)}{N_T} \quad (15)$$

$$Coverage = \frac{Count(page_1 \vee page_2 \cdots \vee page_n)}{N_T} \quad (16)$$

$$m_u = \frac{Support}{Coverage} \quad (17)$$

최종적인 빈발항목집합의 검증은 식 (14), 식 (17) 빈발항목집합 i에 대한 m_s으로 이루어진 신뢰함수 $Belief_{m_s}(i)$와 m_u로 이루어진 신뢰함수 $Belief_{m_u}(i)$를 병합하여 새로운 신뢰함수 $Belief_{m_s \oplus m_u}(i)$이 생성하고 각각 신뢰도를 비교하여 그 차이에 따라 빈발항목집합의 interestingness를 결정한다. 웹 사용 패턴의 전체 알고리즘은 (알고리즘 2)와 같다.

입력 : 빈발항목집합 i를 원소로 갖는 집합 I
출력 : 빈발항목집합의 신뢰도 (*Belief*)및 interestingness (*Interest*)

1. 빈발항목집합 i의 웹 구조 기반 기본 확률값 m_s를 구한다.
   1.1 $lfactor = \frac{L}{N(N-1)}$
   (단, N은 $|i|$, L은 페이지간 직접 링크 수)
   1.2 i가 그래프를 이루면 $cfactor = 1$ 그렇지 않으면 $cfactor = 0$
   1.3 $m_s = lfactor * cfactor$
2. 빈발항목집합 i의 웹 사용 기반 기본 확률값 m_u를 구한다.
   2.1 빈발항목집합의 커버리지(Coverage)를 구한다.
   $$Coverage = \frac{Count(page_1 \vee page_2 \cdots \vee page_n)}{N_T}$$
   (단, N_T는 총 트랜잭션 수)
   2.2 빈발항목집합 생성 시 발생한 지지도(Support)를 이용하여 m_u 계산
   $$m_u = \frac{Support}{Coverage}$$
3. 총 빈발항목집합(I)의 m_s, m_u, 계산을 위해 $|I|$만큼 1-2 수행
4. 빈발항목집합 i의 웹 구조 기반 신뢰도 $Belief_{m_s}(i)$를 구한다.
   $$Belief_{m_s}(i) = \sum_{\forall B: i \subseteq B} m_s(i)$$
5. 빈발항목집합 i의 웹 사용 기반 신뢰도 $Belief_{m_u}(i)$를 구한다.
   $$Belief_{m_u}(i) = \sum_{\forall B: i \subseteq B} m_u(i)$$
6. 빈발항목집합 i의 신뢰함수 병합을 수행하여 $Belief_{m_s \oplus m_u}(i)$ 계산
7. $Belief_{m_s}(i), Belief_{m_u}(i)$,를 $Belief_{m_s \oplus m_u}(i)$ 비교하여 *Interest* 계산
8. *Interest*가 임계값(σ)보다 크면 interestingness로 결정
9. 총 빈발항목집합(I)의 *Interest* 계산을 위해 $|I|$만큼 4-8 수행

(알고리즘 2) 웹 사용 패턴 검증 알고리즘

4. 동적 사용자 프로화일 생성

웹 개인화의 대표적 기술인 협력적 여과는 사용자의 성향을 파악하기 위해 사용자의 피드백 정보를 이용하여 사용자 간의 유사도에 기반하여 클러스터를 생성하여 이를 이용하여 정적 사용자 프로파일을 생성한다. 그러나 협력적 여과에 사용되는 사용자 프로파일은 생성된 후 일정 시간이 지나 업데이트가 되므로 사용자의 성향 변화를 적시에 습득하기 어렵다. 따라서 웹 사이트에 접속한 현재의 사용자의 성향을 파악할 수 있는 동적 사용자 프로파일이 필요하다.

앞 절에서 설명되어진 집합적인 웹 사용 패턴, 즉 클러스터들은 사용자의 현재 활성화 세션을 이용하여 동적으로 사용자 프로파일을 생성한다. 본 절에서는 동적 사용자 프로파일의 형태와 그 생성 방법을 설명한다.

4.1 패턴 기반 사용자 프로파일

본 논문에서 제안하는 사용자 프로파일은 사용자의 활성화된 세션을 이루는 페이지 정보를 기반으로 연관 규칙 형태로 구성되어 지며 아래와 같이 표현된다.

$$Page_1, Page_2, \cdots, Page_n \rightarrow Page_R \; ; \; RecScore = S$$

사용자 프로파일을 이루는 패턴, 즉 규칙은 두 부분으로 나뉘어 지며 규칙의 좌측 부분을 선행부(antecedent)라 하고 우측 부분을 결과부(consequent)라 한다. 선행부는 현재 사용자의 활성화된 페이지로 구성이 되며 결과부는 한 개의 페이지로 구성된다. 결과부는 ARHP에 의해 생성된 클러스터의 페이지가 된다. 그리고 *RecScore*는 4.2에서 설명될 방법에 의해 구해진다. 실제 웹 개인화를 위해 사용되어지는 프로파일의 예는 (그림 2)와 같다.

(그림 2)는 현재 활성화된 세션의 페이지가 $Page_1, Page_2, Page_3$ 일 때 동적으로 생성된 사용자 프로파일을 보여주고 있다. 결과부에는 추천되어질 후보 추천 리스트들이 나타나며 각 후보 추천 페이지는 *RecS*값을 가지고 있다.

선행부(antecedent) :
$\langle page_1, page_1, page_3 \rangle$
결과부(consequent) :

$page_{R1}$	$RecS = 0.97$
$page_{R2}$	$RecS = 0.93$
$page_{R3}$	$RecS = 0.91$
$page_{R4}$	$RecS = 0.88$
$page_{R5}$	$RecS = 0.85$

(그림 2) 동적 사용자 프로파일

이처럼 동적으로 생성된 사용자 프로파일은 정적으로 생성되어 개인화에 사용되는 사용자 프로파일보다 장점이 있다. 즉 현 사용자의 성향을 파악할 수 있어 사용자의 변화하는 성향에 적시에 적응할 수 있기 때문이다.

4.2 프로화일 생성

위에서 언급한 동적 사용자 프로파일을 생성하기 위해서는 현재 활성화된 세션과 생성된 클러스터 정보를 이용하여 $RecS$값을 구해야 한다. 본 절에서는 이에 대해 설명한다.

검증되어진 패턴을 이용한 ARHP 알고리즘은 여러 사용자의 집합적인 패턴을 나타내는 클러스터를 생성하고 각 클러스터는 웹 페이지와 가중치 값으로 이루어진 n-차원의 벡터로 구성하게 하며 아래와 같이 표현된다.

$$C = W_1^c, W_2^c, \cdots, W_n^c$$

$$W_i^c = weight(\pi, C), \text{ if } pi \in C, otherwise\ 0$$

위와 마찬가지로 사용자의 활성화된 세션 S도 n-차원의 벡터로 표현되어지는데

$S = s_1, s_2, \cdots, s_m$ s_i는 현재 페이지의 중요도(significant degree)로서 사용자의 현재 활성화된 세션의 페이지가 클러스터 내의 페이지에 존재하면 1의 값을 가지며 존재하지 않을 경우 0값을 갖는다. 현재 활성화된 세션에서 직접 추천 페이지를 생성하는데 영향을 미치는 세션을 크기가 m인 슬라이딩 윈도우(m-sliding window)라고 하며 현재 세션에서 가장 마지막으로 접근한 m개의 페이지를 말한다[4]. 즉 사용자가 <A, B, C>세션이 있고 3-sliding window를 적용할 경우 다음 접근한 페이지가 D라면 현재 사용자 세션은 <B, C, D>가 된다.

n-차원 벡터로 표현된 클러스터와 사용자 세션은 추천 리스트를 생성하기 위해 matching score를 계산하며 정규화된 코사인 유사도 값을 이용한다. matching score는 아래 식 (18)로 표현된다.

$$match(S, C) = \frac{\sum_k w_k^c \cdot S_k}{\sqrt{\sum_k (S_k)^2 \cdot \sum_k (w_k^c)^2}} \quad (18)$$

식 (7)에서 구해진 matching score를 이용하여 사용자 세션 S와 클러스터 C의 페이지 p의 recommendation score, $RecS(S, p)$을 아래 식 (19)으로 계산한다.

$$RecS(S, p) = \sqrt{weight(p, C) \cdot match(S, C)} \quad (19)$$

만약 클러스터 내의 페이지 p가 사용자의 세션에 존재한다면 $Rec(S, p)$값은 0이 된다. 사용자 프로파일을 생성하기 위한 사용자 세션 S에 대한 최종적인추천 집합인 $UREC(S)$는 추천 임계값(recommendation threshold, ρ) 보다 큰 recommendation score를 가지는 전체 클러스터 TC의 부분집합 C의 페이지 p 집합들이며 아래의 식 (20)로 표현된다.

$$UREC(S) = w_i^c | C \in TC, \text{ and } Rec(s, w_i^c) \geq \rho \quad (20)$$

검증되어진 웹 사용 패턴을 이용해 생성된 클러스터는 여러 사용자의 공통된 웹 사용 패턴을 보여 주며 사용자의 활성화된 세션을 이용하여 동적으로 사용자 프로파일을 구성할 수 있게 한다. 즉 개인화 서비스 시스템은 사용자의 활성화 세션에 식 (18)~식 (20)을 적용해 추천리스트를 동적으로 생성하여 사용자 프로파일을 구축하게 된다.

5. 실험 결과

5.1 실험 환경 및 데이터

본 논문의 실험환경에서 오프라인 작업(웹 로그 파일 전처리, 연관 규칙 생성, 패턴 검증, 클러스터 생성)을 위해 Java(J2SE 1.3)를 이용하여 통합환경을 구축하였으며 웹 로그 파일 전처리를 위해 Oracle 8i DBMS를 사용하였다. 온라인 작업(동적 사용자 프로파일 생성)은 웹 환경에서 적용되어야 하므로 Windows 2000 서버에 Tomcat 서블릿 엔진을 이용하였으며 사용자 세션 트래킹을 위해 [5]에서 제공하는 Java Applet를 이용하였다.

웹 로그 파일은 A증권 사이트에서 2001.7.1~2001.7.31까지의 축적된 로그 데이터를 이용하였다. 웹 로그 파일의 전처리는 [13]에서 제시한 방법을 사용하여 456개의 트랜잭션을 식별하였으며 총 페이지의 수는 125개였으며 전체 트랜잭션에서 10% 이상 90% 이하의 비율로 나타나는 페이지만을 트랜잭션의 페이지로 결정하였다. 또한 트랜잭션의 길이가 3 이상인 트랜잭션만 고려하였다. 또한 전체 트랜잭션에서 20%를 평가 데이터(evaluation data)으로 설정하였고 나머지 트랜잭션 데이터는 트레이닝 데이터(training data)로 사용하였다. 전처리되어 생성된 트랜잭션은 <표 1>과 같다

<표 1> 456개의 사용자 트랜잭션 테이블

트랜잭션	URL 스트림
TID1	investment -> list_dbno_33 -> list_db_no_41 -> inv_info
TID2	investment -> list_dbno_33 -> liv_info -> unstock
TID3	list_dbno_39 -> bsnewlist_BA -> unstock
.	.
.	.
.	.
TID456	3market -> list_dbno_104 -> inv_info -> left_menul

5.2 성능 평가 기준 및 평가

본 연구의 최종적인 결과물은 현재 사용자 세션에 따라 동적으로 생성된 사용자 프로파일이다. 사용자 프로파일은 4.1의 (그림 2)처럼 추천되어진 웹 페이지이다. 사용자 프로파일 내의 추천 리스트들이 얼마나 사용자의 성향을 반영하는가에 대한 평가 기준이 필요하다.

본 실험에서는 B. Mobasher[8]가 제시한 추천의 정확도 측정 방법을 적용하였다. 앞서 언급한 평가 데이터는 실제 사용자들의 웹 사용 패턴을 보여주므로 생성된 추천 리스트의 정확성을 파악할 수 있다. 기본적인 방법은 다음과 같

다. 먼저 주어진 평가 데이터에서 임의의 트랜잭션 t를 선택하고 현재 사용자의 활성화 슬라이딩 윈도우 크기를 n이라고 할 때 임의의 |t|-n+1개의 페이지로 구성된 페이지 그룹을 t로부터 구하고 이 페이지 그룹을 현재 사용자의 대표 사용자 세션으로 정한다. 각각의 페이지 그룹과 생성되어진 클러스터를 이용하여 추천 리스트를 생성하고 t에서 페이지 그룹의 뺀 나머지 페이지가 페이지 리스트에 나타나는 비율에 대한 평균이 t에 대한 평가 스코어가 된다. 따라서 전체 트랜잭션에 대한 추천 정확성은 각각의 트랜잭션 t에 대한 평가 스코어 평균값이 된다.

실험은 먼저 전처리 되어진 파일로부터 Apriori 알고리즘을 적용하여 빈발항목집합을 찾았다. 빈발항목집합 발견 시 최소 지지도는 빈발항목집합의 수를 감소시키는 패턴의 interestingness에 대한 객관적인 척도(objective measure)이며 최소 신뢰도는 ARHP 알고리즘을 이용하여 클러스터를 생성할 때 빈발항목집합 즉, edge의 가중치를 설정하는데 사용된다. 본 실험에서 적용한 최소 지지도와 최소 신뢰도는 <표 2>와 같다. 그리고 패턴 검증이 적용되지 않은 Apriori 알고리즘 및 ARHP를 이용하여 생성된 내용은 <표 3>과 같다.

〈표 2〉 최소 지지도와 최소 신뢰도

최소 지지도(min_supp)	최소 신뢰도(min_conf)
20%	50%

〈표 3〉 패턴 검증이 적용되지 않은 결과물

빈발항목집합수	총 페이지 수	클러스터 수
87	68	14

<표 3>에서와 같이 전체 빈발항목집합이 14개의 클러스터로 구분되었으며 각 클러스터는 페이지간의 연결정도에 따라 가중치를 갖는다. 추천 리스트를 생성하기 위해 평가 데이터에서 임의로 하나의 트랜잭션을 선택하였으며 현 사용자 세션의 슬라이딩 윈도우는 2로 설정하였으며 추천 리스트 설정 시 추천 스코어에 대한 임계값(threshold)은 0.5로 하였다. 아래 <표 4>는 현 사용자 세션이 investment -> list_dbno_33일 때 생성된 추천리스트 이며 추천 정확성 평가 기준인 평가 스코어는 평균값 0.722이였다. 전체 트랜잭션에 대한 추천 정확성에 대한 평가 스코어는 0.714였다.

〈표 4〉 패턴 검증이 적용되지 않은 추천 리스트

세션 윈도우	추천 페이지	추천 스코어
investment-> list_dbno_33	inv_info	0.785
	list_dbno_88	0.752
	list_dbno_117	0.714
	3market	0.664
	list_dbno_89	0.591
	list_dbno_114	0.53

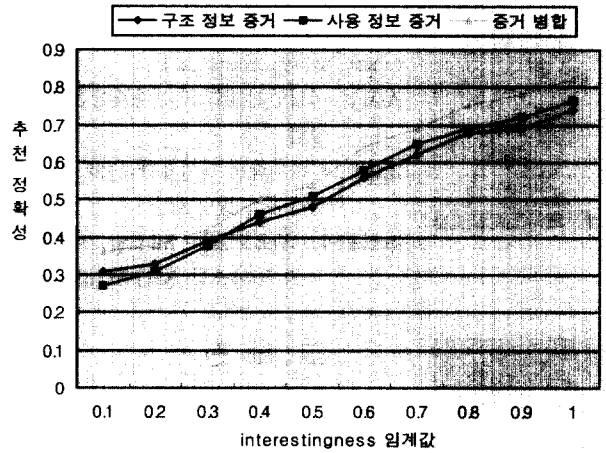
빈발항목집합에 대한 검증의 수행은 Dempster-Shafer 증거 이론을 적용하여 수행하였다. 즉 찾아진 빈발항목집합에 대한 신뢰도(degree of belief)를 웹 구조 정보 증거와 웹 사용 정보 증거를 이용하여 구한다. 그리고 새로운 증거 집합으로 앞서 적용한 구조 및 사용 정보 증거를 병합하여 신뢰도를 구하였다. <표 5>는 전체 빈발항목집합의 사용 정보 증거와 증거 병합 검증 결과를 보이며 interestingness 임계값은 0.4이다. 지지도가 0.2 이상인 빈발항목집합 중에서 지지도가 높은(0.8 이상) 항목은 너무 자주 전체 트랜잭션에 나타나므로 흥미롭지 못한 패턴이 될 수 있으며 지지도가 Apriori 알고리즘 시 적용한 지지도(0.2) 보다는 상대적으로 높지만 지지도가 0.5 이하는 빈발항목집합은 병합으로 인하여 유용한 패턴에서 제거됨을 <표 5>와 <표 6>을 통해 알 수 있다. <표 5>와 <표 6>의 차이는 사용 정보 증거와 증거 병합을 이용한 유용성 검증, 구조 정보 증거와 증거 병합을 이용한 유용성 검증으로 나누어 실험을 수행한 데이터이다. 구조 정보 증거는 페이지간의 링크 수 및 그래프 형성여부가 신뢰도를 계산하는데 영향을 미쳐 사용 정보 증거에 기반한 유용성 검증에서 interestingness 값이 다소 차이가 나는 것을 실험을 통해 알 수 있다. 이는 사용자의 웹 페이지 접근 경향은 사이트의 구조적인 정보에 많은 영향을 받지 않음을 의미한다.

〈표 5〉 사용 정보 증거와 병합의 임계값 0.4 검증

사용	병합	interes-tingness	지지도	빈 발 항 목 집 합
0.653	0.09	0.583	0.74	investment, list_dbno_41, 3market, list_dbno_89
0.672	0.112	0.56	0.713	investment, list_dbno_44, inv_info
0.665	0.12	0.545	0.692	3market, list_dbno_104, inv_info, list_dbno_118
0.66	0.14	0.52	0.65	unstock, list_dbno_113, list_dbno_BK, newlist
0.716	0.224	0.482	0.642	list_dbno_104, inv_info, investment, newlist,
0.72	0.27	0.45	0.63	bsnewlist_BA, inv_info, list_dbno_88, 3market
0.78	0.36	0.42	0.58	list_dbno_37, list_dbno_126, bsnewlist_BF
⋮	⋮	⋮	⋮	⋮
0.65	0.54	0.11	0.45	invetment, list_dbno_33, list_dbno_41, newlist

〈표 6〉 구조 정보 증거와 병합의 임계값 0.4인 검증

구조	병합	interes-tingness	지지도	빈 발 항 목 집 합
0.651	0.09	0.561	0.74	investment, list_dbno_41, 3market, list_dbno_89
0.677	0.112	0.545	0.713	investment, list_dbno_44, inv_info
0.65	0.12	0.53	0.692	3market, list_dbno_104, inv_info, list_dbno_118
0.63	0.14	0.49	0.65	unstock, list_dbno_113, list_dbno_BK, newlist
0.696	0.224	0.472	0.642	list_dbno_104, inv_info, investment, newlist,
0.712	0.27	0.442	0.63	bsnewlist_BA, inv_info, list_dbno_88, 3market
0.74	0.36	0.38	0.58	list_dbno_37, list_dbno_126, bsnewlist_BF
⋮	⋮	⋮	⋮	⋮
0.61	0.54	0.07	0.45	invetment, list_dbno_33, list_dbno_41, newlist



(그림 3) Dempster-Shafer 이론을 적용한 추천 정확성

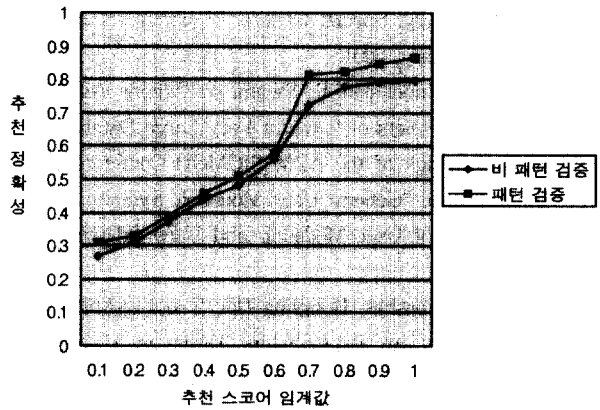
이처럼 Dempster-Shafer 적용을 통해 ARHP에 대한 입력 데이터 수는 전체 빈발항목집합의 16%가 감소하였으며 클러스터의 수 역시 3~4 정도 감소하였다. 이는 빈발항목집합의 interestingness 기준으로 지지도에 기반하지 않고 구조 및 사용 정보 증거를 바탕으로 통계적인 추론을 수행하여 신뢰도(degree of belief)를 계산하였기 때문이다. 특히 Dempster-Shafer 증거 이론은 다양한 증거 공간으로부터 얻어진 증거를 병합할 수 있어 불확실성을 줄일 수 있음을 실험을 통해 알 수 있었다.

마지막으로 검증되어진 빈발항목집합을 적용하여 생성된 11개의 클러스터를 이용하여 추천 리스트를 생성하였다. <표 4>에 적용된 같은 조건의 사용자 세션을 적용한 결과 세 번째 페이지(list_dbno_117)와 다섯 번째 페이지(list_dbno_89)가 높은 추천 스코어 값을 가지며 다른 페이지로 대치되었다. 또한 검증되기 전 단계에서는 추천 스코어 임계값(0.5) 보다 작은 페이지들이 추천 리스트에 추가되었다. 이는 기존의 빈발항목집합 중에서 유용성이 작은 항목이 사라지고 따라서 클러스터 내의 가중치 벡터가 변함에 따라 발생한 것이다.

〈표 7〉 패턴 검증이 적용된 추천 리스트

세션 윈도우	추천 페이지	추천 스코어
investment -> list_dbno_33	list_dbno_88	0.814
	inv_info	0.793
	bsnew_list_BA	0.706
	investment	0.671
	3market	0.646
	list_dbno_114	0.55

<표 7>의 현 사용자 세션에 대한 추천 정확성 평가 기준인 평가 스코어는 평균값 0.842이였다. 전체 트랜잭션에 대한 추천 정확성에 대한 평가 스코어는 0.817였다. 따라서 추천의 정확성은 검증 전에 비해 14.5% 증가하였음 실험을 통해 알 수 있었으며 (그림 4)을 통해 확인할 수 있다.



(그림 4) 패턴 검증의 효율성

6. 결 론

본 논문에서는 웹 사용 패턴의 검증을 통한 동적인 사용자 프로파일의 생성에 관한 방법에 대해 제안하였다. 웹 로그 데이터로부터 익명 사용자들의 웹 사용 패턴을 찾아 클러스터를 생성하기 위해 연관 규칙 생성 알고리즘과 ARHP 알고리즘을 적용하였다. 그러나 연관 규칙 생성 알고리즘은 너무나 많은 수의 패턴을 찾고 서로 관련이 없거나 유용하지 못한 패턴을 찾는 문제점이 있다. 이 문제점을 해결하기 위해 먼저 찾아진 패턴의 유용성 여부를 판단하게 하는 기준으로 interestingness 적용하였으며 패턴의 interestingness를 계산하기 위해 Dempster-Shafer 이론을 적용하였다. Dempster-Shafer 이론은 다양한 증거 공간으로부터 얻은 증거들을 병합하여 불확실한 패턴에 대한 신뢰도를 향상시키는 이론으로서 본 논문에서는 다양한 증거 공간으로 웹 구조 정보와 웹 사용 정보를 적용하였다. 검증된 패턴은 ARHP 알고리즘의 입력 데이터로 사용되고, 이는 유용한 패턴으로 이루어진 클러스터를 생성하였다. 생성되어진 클러스터는 현재 사용자 세션에 따라 동적으로 추천 페이지를 생성하였고 생성된 추천 페이지에 대한 추천 정확성 실험을 통해 본 논문이 제시한 방법이 효율적임을 알 수 있었다.

Dempster-Shafer 이론에서 적용한 다양한 증거 공간을 웹 구조 정보와 웹 사용 정보에 국한하지 않고 웹 컨텐츠 정보를 하나의 증거 공간으로 한 연구가 진행되어야 하며 본 논문의 실험에서 고려되지 않은 plausibility 값을 적용한 실험이 진행되어야 한다. 또한 동적 사용자 프로파일을 생성하기 위해 서버측에 별도의 장치를 마련하여 생성 속도를 향상시켜야 할 것이다.

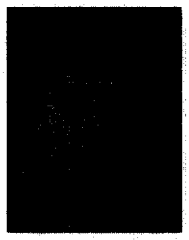
참 고 문 헌

[1] Adomavicius, G., and Tuzhilin, A. "Expert-Driven Validation of Rule-Based User Models in Personalization Applications," International Journal on Data Mining and Know-

ledge Discovery. Special Issue on E-commerce and Data Mining, January, 2001.
[2] Agrawal R., Imielinski T., Swami A. "Mining Association Rules betweenSets of Items in Very Large Databases," In the Proceedings of the ACM SIGMOD Conference, 1993.
[3] Alfred Kobsa. "Generic User Modeling Systems," In the Proceedings of User Modeling and User-Adapted Interaction, 2000.
[4] Bamshad Mobasher, Honghua Dai, Tao Luo,Miki Nakagawa, Yuqing Sun, Jim Wiltshire "Discovery of Aggregate Usage Profiles for Web Personalization," WEBKDD2000.
[5] C. Shahabi, A. Zarkesh, J. Adibi, and V. Shah,"Knowledge discovery from users Web-page navigatioin," In Proceedings of Workshop on Research Issues in Data Engineering, 1997, Birmingham, England.
[6] E. H. Han, et al., "Clustering Based On Association Rule Hypergraphs," Proc. of SIGMOD '97 Workshop on Research Issues in Data Mining and Knowledge Discovery (DMKD), May, 1997.
[7] G. Karypis and V. Kumar, "Multilevel k-way Hypergraph Partitioning," DAC, pp.343-348, 1999.
[8] G. Shafer, "A Mathematical Theory of Evidence," Princeton University Press, 1976.
[9] Goldberg, D. Nichols, "Using Collaborative Filtering to Weave an Information Tapestry," Comm. of the ACM35 (12), pp.61-70, 1992.
[10] M. Spiliopoulou and L. C. Faulstich, "WUM : A Web Utilization Miner," In Proceedings of EDBT Workshop WebDB98, LNCS 1590, Springer Verlag, Balencia, Spain, 1999.
[11] O. Nasraoui, H. Frigui, A. Joshi, R. Krishnapuram, "Mining Web access logs using relational competitive fuzzy clusering," In Proceedings of the Eight International Fuzzy Systems Association World Congress, August, 1999.
[12] O. R. Zaiane, M. Xin, and J. Han, "Discovering weeb access patterns and trends by applying olap and data mining technology on web logs," In Advances in Digital Libraries, Santa Barbara, CA, pp.19-29, 1998.
[13] R. Cooley, et al., "Data Preparation for Mining World Wide Web Browsing Patterns," Knowledge and Information Systems, Vol.1-1, 1999.
[14] R. Cooley, et al., "Discovery of Interesting Usage Patterns from Web Data," WEBKDD, 1999.

안 계 순

e-mail : kyesun@im.inha.ac.kr
2001년 시립인천대학교 전자계산공학과 졸업
2003년 인하대학교 전자계산공학과 대학원 석사과정 졸업예정
관심분야 : Web Usage Mining, 지능형 추천 시스템, Machine Learning



고 세 진

e-mail : simpler01@hotmail.com
1999년 시립인천대학교 전자계산공학과 졸업
2002년 인하대학교 전자계산공학과 대학원 석사과정 졸업
관심분야 : Web Usage Mining, 지능형 추천 시스템, Machine Learning

정 준

e-mail : jjeong@im.inha.ac.kr
1999년 인하대학교 전자계산공학과 졸업
2001년 인하대학교 전자계산공학과 대학원 석사과정 졸업
2001년~현재 인하대학교 전자계산공학과 대학원 박사과정 재학중
관심분야 : Collaborative Filtering, Content-based Filtering, Machine Learning

이 필 규

e-mail : pkrhee@inha.ac.kr
1975년~1982년 서울대학교 전기공학 박사
1982년~1985년 KIST 시스템구조데이타 통신실 연구원
1985년~1986년 East Texas State University 전산학 박사
1987년~1990년 University of SW Louisiana 전산학 박사
1991년~1992년 한국전자통신연구소 컴퓨터연구단 선임연구원
1993년~1994년 IBM T. J. Watson Research Center 객원연구원
1992년~2001년 인하대학교 전자계산공학과 부교수
2001년~현재 인하대학교 전자계산공학과 정교수
관심분야 : 이미지프로세싱, Agent, Machine Learning