Journal of Korea Multimedia Society Vol. 8, No. 9, September 2005(pp. 1163-1176)

빈발 순회패턴 탐사에 기반한 확장된 동적 웹페이지 추천 알고리즘

이근수[†], 이창훈[††], 윤선희[†††], 이상문[††††], 서정민[†††††]

요 약

웹은 가장 커다란 분산 정보저장소로서 빠른 속도로 성장했으나, 웹의 정보를 읽고 이해하는 데는 본질적으로 한계가 있다. 웹의 이러한 환경에서 사용자의 순회패턴(traversal patterns)을 탐사하는 것은 시스템 설계나 정보서비스 제공 측면에서 중요한 문제이다. 본 논문에서는 세션에 나타나는 페이지들간의 연관성 정보를 활용하여 빈발 k-페이지집합을 탐사하여 추천 페이지집합을 생성함으로써 효율적인 웹 정보서비스를 제공할 수 있는 Web Page Recommend(WebPR) 알고리즘[11]을 확장한다. 확장된 내용은 WebPR(A) 알고리즘을 추가하였으며, WebPR(T)에서 윈도우 개념을 도입한 새로운 winWebPR(T) 알고리즘을 제안하고 있다. 두개의 확장된 알고리즘을 포함하여 두개의 실제 웹로그(Weblog) 데이터에 대해 실험 결과에서 알 수 있듯이 윈도우 개념을 도입한 winWebPR(T) 알고리즘이 세션에 나타나는 페이지들간의 모든 연관성 정보를 활용함으로써 가장 우수한 성능을 보였다.

An Extended Dynamic Web Page Recommendation Algorithm Based on Mining Frequent Traversal Patterns

KeunSoo Lee[†], Lee Chang Hoon[††], Sun-Hee Yoon[†††], Sang Moon Lee[††††], Jeong Min Seo[†††††]

ABSTRACT

The Web is the largest distributed information space but, the individual's capacity to read and digest contents is essentially fixed. In these Web environments, mining traversal patterns is an important problem in Web mining with a host of application domains including system design and information services. Conventional traversal pattern mining systems use the *inter-pages association* in sessions with only a very restricted mechanism (based on vector or matrix) for generating frequent k-Pagesets. We extend a family of novel algorithms (termed WebPR - Web Page Recommend) for mining frequent traversal patterns and then pageset to recommend. We add a WebPR(A) algorithm into a family of WebPR algorithms, and propose a new winWebPR(T) algorithm introducing a window concept on WebPR(T). Including two extended algorithms, our experimentation with two real data sets, including LadyAsiana and KBS media server site, clearly validates that our method outperforms conventional methods.

Key words: Traversal Pattern(순회패턴), Frequent k-Pageset(빈발 k-페이지집합), Inter-Pages Association(페이지간 연관성), Recommendation(추천)

※ 교신저자(Corresponding Author) : 서정민, 주소 : 경기도 안성시 석정동(608-743), 전화 : 031)670-5169, FAX : 031)670-5169, E-mail : jmseo@gpl.khnu.ac.kr
접수일 : 2005년 2월 1일, 완료일 : 2005년 4월 21일
[†] 정회원, 한경대학교 컴퓨터공학과 교수
(E-mail : kslee@hknu.ac.kr)
[††] 정회원, 한경대학교 컴퓨터공학과 조교수
(E-mail : chlee@hknu.ac.kr)
[†††] 미림여자정보과학고등학교 교사
(E-mail : sunniyoon@hanmail.net)
[††††] 종신회원, 충주대학교 전자계산학과 교수
(E-mail : smlee@chungju.ac.kr)
[†††††] 한경대학교 대학원 컴퓨터공학과 박사과정

1. 서 론

월드 와이드 웹은 뉴스, 광고, 소비자 정보, 재정관리, 교육, 정부, 전자상거래 등의 많은 정보 서비스를 위해 거대하고 널리 분산된 정보 서비스 센터로서의 역할을 한다[1,2]. 또한 풍부하고 동적인 하이퍼링크 정보와 웹 페이지 접근 정보 등을 포함하고 있어서 데이터 탐사를 수행하는데 요구되는 풍부한 자원을 제공해준다. 그러나 웹 정보의 극히 일부만이 서로 강하게 연관되어 있거나 유용하다.

전형적인 웹 사이트들은 수천 혹은 수백만 명에 의한 페이지 접근정보 시퀀스를 포착할 수 있는 거대한 로그 데이터를 생성한다. 웹 환경에서 사용자의 접근패턴을 포착하는 것을 순회패턴 탐사(mining traversal patterns)라 한다. 웹 서버로그에서 빈발 순회패턴 탐사는 웹 사이트의 설계에 대한 결정을 도와주거나[3,4], 적응적인 웹 사이트를 가능하게 하거나[5], 마케팅 결정을 지원하거나[6], 사용자 인터페이스 테스팅, 보안 목적을 위한 감시 등을 들 수 있다. 특히 더 중요한 응용분야로는 추천시스템[7]이나 목표시장 광고 등과 같은 웹 개인화 분야가 있다.

한편, 시퀀스(즉, 순차 항목집합)를 포함하는 대용량 데이터베이스로부터 순차패턴을 발견하는 것은 지식 발견 및 데이터 탐사 분야에서 중요한 문제이다[8,9]. 즉, 데이터 시퀀스 집합이 주어지면, 목적은 빈발한 서브시퀀스들을 발견하는 것이다. 여기서 빈발하다고 하는 것은 서브시퀀스들을 포함하는 데이터 시퀀스들의 비율이 사용자가 지정한 최소 지지도를 초과한다는 것을 의미한다[8,9]. 빈발 순차패턴 탐사는 수많은 잠재적인 응용분야에 적용될 수 있다: 소매점(즉, market-basket 데이터), 통신, 의학, 웹 등. Market-basket 데이터베이스에서 각 데이터 시퀀스는 일정 기간 동안 개별 고객에 의해 구매된(시간에 따라 순서를 가지는) 항목집합을 말한다. 이때 빈발하게 발생하는 패턴들은 고객의 행위를 예측하는데 매우 유용하다.

최근, 빈발 순차패턴(혹은 순회패턴) 탐사를 위한 효과적인 알고리즘들이 제안되었다[5,8-10]. 이들 알고리즘들은 다양한 휴리스틱(heuristic) 혹은 사이트 정보 등을 이용하여 보다 효과적인 빈발 순회패턴을 탐사하고자 한다. 그러나 이들 방법은 빈발 순회패턴을 탐사함에 있어 순회패턴에 포함되어 있는 정보를 충분히 활용하지 못한다. 예를 들어, [10]에서 제안한 알고리즘은 순회패턴에 포함된 특징을 벡터로 표현하여 빈발 순회패턴을 탐사함으로써 탐사과정에 유용할 수 있는 순회패턴의 순서 특징을 반영하지 못한다. 또한 [5]의 방법은 순회패턴의 특징을 유사도 행렬(similarity matrix)로 표현하고 있다. 그들이 사용한 유사도 행렬은 순회패턴에서 인접한 페이지들간의 특징은 탐사과정에 반영되지만 인접하지 않은 페이지들간의 순서 특징은 표현하지 못한다. 한편, 대표적인 순차패턴 탐사 모델[8,9]로 클러스터링 기법을 이용한 위의 두 방법들과는 전혀 다른 접근방법인 연관 규칙을 생성하는 AprioriAll 알고리즘이 있다. AprioriAll 알고리즘은 한 트랜잭션 안에서 발생하는 항목들간의 연관규칙에 시간의 변이를 추가한 것으로 트랜잭션 상호간의 문제를 다룬다. 즉, AprioriAll에서의 지지도 계산은 몇 명의 고객들이 주어진 후보 시퀀스를 지지하느냐를 측정한다. 본 논문에서는 이러한 순차패턴 탐사 알고리즘인 AprioriAll을 순회패턴 탐사 문제에 변형하여 적용시킨 WebPR(A) 알고리즘을 제안한다. WebPR(A)는 한 세션의 부분 페이지집합이 완전하게 포함되어 있는 클러스터를 생성하게 된다. 따라서 완전한 형태는 아니지만 제한된 순차 부분집합의 빈발 집합(인접하지 않아도 순서만 맞으면 허용)을 생성하는 것으로 볼 수 있다.

본 논문에서는 효율적인 웹 정보서비스를 제공하기 위해 먼저 웹로그로부터 빈발 순회패턴들을 탐사하고, 이를 기반으로 추천 페이지집합을 생성하는 WebPR 알고리즘을 확장한다. 확장된 내용은 대표적인 빈발 순차패턴 알고리즘인 AprioriAll 알고리즘을 변형 적용한 WebPR(A)을 추가하여 폭넓은 실험을 수행하였으며, WebPR(T)에 윈도우 개념을 도입한 winWebPR(T) 알고리즘을 제안한다.

본 논문의 구성은 다음과 같다. 2장에서는 효율적인 웹 정보서비스 제공을 위해 빈발 k-페이지집합 기반 순회패턴 탐사를 수행하는 Web Page Recommend(WebPR) 알고리즘의 확장된 내용을 기술한다. 3장에서는 확장된 두개의 알고리즘을 포함하여 전체 WebPR 알고리즘들을 두개의 실제 사이트의 웹로그에 적용한 실험을 비교 분석한다. 마지막 4장에서는 결론 및 향후 연구과제에 대해 기술한다.

2. 빈발 k-페이지집합 기반 순회패턴 탐사

2.1 개요

이 장에서는 웹로그 데이터를 이용하여 신뢰할 수 있는 빈발 페이지집합 F_k를 탐사하는 다양한 모델들을 제안하고, 이를 웹 사이트에 적용하여 웹 사용자에게 신뢰할 수 있는 정보 서비스를 제공하는 Web Page Recommend (WebPR) 알고리즘을 제안한다.

추천 페이지집합 탐사를 위한 세 개의 WebPR 알고리즘들은 순회패턴들로부터 빈발 k-페이지집합 F_k를 생성하는 방법들에 대한 전체적인 범위를 고려하는 것으로 볼 수 있다. 여기서 F_k는 모든 빈발 k-페이지집합을 의미한다. 기본적으로 네 개의 알고리즘은 자연스러운 확장 개념을 표현한다. 즉, 각 알고리즘은 빈발 k-페이지들의 집합 F_k를 생성함에 있어서 세션에서의 페이지들간의 연관성을 고려하는 정도를 자연스럽게 확장시킨다. 첫 번째 알고리즘인 WebPR(V)는 (여기서 "V"는 Vector를 의미한다) 세션에서의 연관성을 고려하는 정도가 가장 약하다. 이 알고리즘은 단순히 세션 데이터베이스에서 한 페이지의 빈발 횟수만을 고려한다. 따라서 빈발 *1*-세션들의 집합을 생성한다. 두 번째 알고리즘인 WebPR(M)은 (여기서 "M"은 Matrix를 의미한다) 세션에서 인접한 페이지들간의 연관성을 고려한다. 이 알고리즘은 빈발 *2*-페이지들의 집합을 생성한다. 세 번째 알고리즘인 WebPR(A)는 (여기서 "A"는 [8]의 AprioriAll 알고리즘을 의미한다) 세션의 인접한 페이지들뿐만 아니라 멀리 떨어진 페이지들간의 연관성도 고려함으로써 빈발 k-페이지들의 집합을 생성한다. 그러나 이 알고리즘에서는 반드시 세션의 첫 페이지를 포함시켜 연관성을 고려한다. 네 번째 알고리즘인 WebPR(T)는 (여기서 "T"는 Tree를 의미한다) 세션의 모든 서브패턴에 대해 연관성을 고려하여 빈발 k-세션들의 집합을 생성한다. 네 개의 알고리즘에 의해 생성된 빈발 페이지집합의 종속관계는 세션에서 페이지들간의 연관성을 얼마나 고려했느냐에 달려있다; 즉,

$$|F_k^{WebPR(T)}| \le F_{k'}^{WebPR(A)} \le |F_2^{WebPR(M)}| \le |F_1^{WebPR(V)}| \quad (1)$$

WebPR 알고리즘들은 다음과 같은 세부 단계들로 이루어져 있다:

(1) 전처리: 웹로그에 대하여 데이터 클리닝 등의 전처리를 수행하고, IP 주소, 시간 등을 이용하여 세션 데이터베이스 D를 생성한다.
(2) 빈발 페이지집합 탐사: 빈발 k-페이지집합을 생성한다.
(3) 추천 페이지집합 생성: 결속력있는 개념을 표현하는 페이지집합을 생성한다.
(4) 추천 적용: 웹 사이트를 수정하여 직접 동적 추천을 수행하거나 생성된 추천 페이지집합을 웹 마스터에게 제안한다.

전처리(preprocessing)는 제안한 네 개의 WebPR 알고리즘 모두에게 필요하다. 즉, 네 개의 WebPR 알고리즘은 전처리 결과로 얻은 세션 데이터베이스로부터 시작한다. 특히 WebPR(V)와 WebPR(M)은 빈발 k-페이지집합 F_k를 생성하기 위한 전단계로 클러스터링(clustering)을 요구한다. 이와 달리 WebPR(A)와 WebPR(T)는 클러스터링이나 클러스터 모델을 생성하지 않고서도 빈발 k-페이지집합을 생성하고 추천 페이지집합 R을 탐사할 수 있다. 다음 절에서 네 개의 모든 알고리즘에서 요구하는 과정인 전처리(preprocessing) 과정을 설명하고, 이어서 네 개의 제안한 WebPR 알고리즘들의 각각에 대하여 빈발 k-페이지집합을 생성하는 과정과 빈발 k-페이지집

표 1. 세 개의 WebPR 알고리즘 비교

알고리즘	페이지들간의 연관성 활용정도	생성된 빈발 집합
WebPR(V)	1 페이지	빈발 *1*-페이지집합
WebPR(M)	2 페이지	빈발 *2*-페이지집합
WebPR(A)	제한된 순차 부분집합(인접하지 않아도 순서만 맞으면 허용)	빈발 k'-순차 페이지집합
WebPR(T)	모든 k 페이지	모든 빈발 k-페이지집합

합을 이용하여 웹 사이트에서 추천하는 과정을 상세하게 설명한다. 전처리에 관한 자세한 내용은 참고문헌 [11]에 기술되어 있다.

2.2 WebPR(V) 알고리즘

WebPR(V)는 순회패턴 탐사와 추천을 위해 Yan 등[10]의 벡터모델을 단순하게 수정한 것이다. WebPR(V)는 후보 세션 s의 모든 페이지들에 대한 빈발 횟수를 계산한다. 이것은 세션에서 페이지들간의 연관성을 전혀 고려하지 않은 방법으로 빈발 1-세션(즉, 한 페이지)들의 집합을 생성한다. WebPR(V)는 빈발 *1*-페이지집합 F_1의 집합을 생성하기 위해 세션 데이터베이스의 모든 세션들을 클러스터링하는 과정을 요구한다. 생성된 F_1 집합은 최소 지지도($s_{\min}$)에 따라 달라진다. 또한 추천과정에서 추천 페이지집합 R은 F_1 집합으로부터 추천 페이지 개수 제한조건(N_{rp})에 따라 생성된다. WebPR(V) 알고리즘에 관한 자세한 내용은 참고문헌 [11]에 기술되어 있다.

2.3 WebPR(M) 알고리즘

WebPR(M)은 빈발 페이지집합을 생성하기 위해 세션 데이터베이스로부터 빈발 *2*-페이지집합을 생성한다. 이것은 세션의 인접한 페이지들간의 연관성을 고려한 것이다[5,11]. WebPR(M)은 F_2 집합을 생성하기 위해 세션 데이터베이스의 모든 세션들을 각 클러스터별로[1] 인접 행렬(adjacency matrix)을 생성한다. 이러한 인접 행렬을 이용하여 액티브 세션의 현재 페이지와 연관성이 높은 페이지집합을 생성할 수 있다. 즉, WebPR(V)에서 생성한 *CV*를 이용하여 현재 액티브 세션의 클러스터를 찾은 다음 액티브 세션의 현재 페이지와 연관성이 높은(즉, 평균 빈발횟수가 높은) 페이지들의 집합을 생성한다. WebPR(M)의 F_2집합은 최소 지지도($s_{\min}$)에 따라 달라진다. 또한 추천과정에서 추천 페이지집합 R은 F_2집합으로부터 추천 페이지 개수 제한조건(N_{rp})에 따라 생성된다. WebPR(M) 알고리즘에 관한 자세한 내용은 참고문헌[11]에 기술되어 있다.

1) WebPR(V)에서 클러스터링 수행 시 세션 데이터베이스에 각 세션에 대한 클러스터 ID를 저장한다. 따라서 클러스터 ID를 이용하여 클러스터별로 인접 행렬을 생성할 수 있다.

2.4 WebPR(A) 알고리즘

WebPR(A)는 추천 페이지집합 탐사를 위해 AprioriAll 알고리즘[8]을 변형한 것이다. WebPR(A)에서 빈발 k-페이지집합 F_k를 생성하는 방법은 AprioriAll 알고리즘의 방법을 그대로 따른다. 그러나 AprioriAll에서는 생성된 F_k 집합들과 지지도(support)와 신뢰도(confidence)를 기반으로 연관 규칙(association rules)을 생성하지만, WebPR(A)에서는 연관 규칙을 생성하는 대신 지지도만을 이용하여 빈발 페이지집합을 생성하도록 변형시킨다. 생성된 빈발 페이지집합은 지지도에 따라서 달라질 수 있다.

2.4.1 빈발 페이지집합 탐사

그림 1은 WebPR(A)의 빈발 페이지집합 F_k를 생성하는 예를 보여준다. WebPR(A)는 AprioriAll 알고리즘과 동일하게 빈발 k-페이지집합을 생성한다. AprioriAll에서는 각 고객들의 트랜잭션을 시간 순으로 볼 수 있는데 (이것을 소비자 순차집합이라 한다) 그 형태는 <항목집합(T_1) 항목집합(T_2) . . . 항목집합(T_n)>과 같은 시퀀스(sequence)이다. 시퀀스가 특정 고객에 대한 소비자 순차집합에 속해 있다면 그 고객은 이 시퀀스를 지지한다고 말할 수 있다. 이러한 AprioriAll 알고리즘을 웹에서의 사용자 순회패턴을 탐사하는 문제에 적용할 수 있다. 즉, 세션 s=<s_1 s_2… s_n>의 각 페이지 s_i를 소비자 순차집합의 단순 항목집합으로 볼 수 있다. 예를 들어, 세션 s의 첫 번째 페이지 s_1은 소비자가 구매한 첫 번째 단순 항목집합으로 간주할 수 있다.

따라서 WebPR(A)에서는 각 세션에서 빈발 페이지집합의 후보 집합을 구성하고 난 후에 각 후보 페이지집합의 발생 빈도수를 계산하고, 미리 정해진 최소 지지도(minimum support)를 기초로 하여 빈발 페이지집합을 결정한다. 첫 번째 패스에서 WebPR(A)는 각 항목의 발생 빈도수를 카운트하기 위하여 단순히 모든 트랜잭션(혹은 세션)들을 스캔하여 읽는다. 후보 *1*-페이지집합들의 집합 C_1은 그림 1에서와 같이 얻어진다. 최소 지지도가 2라고 가정하면(s_{min} = 2), 요구되는 최소 지지도를 가지는 후보 *1*-페이지집합들로 구성되는 빈발 *1*-페이지집합들의 집합 F_1이

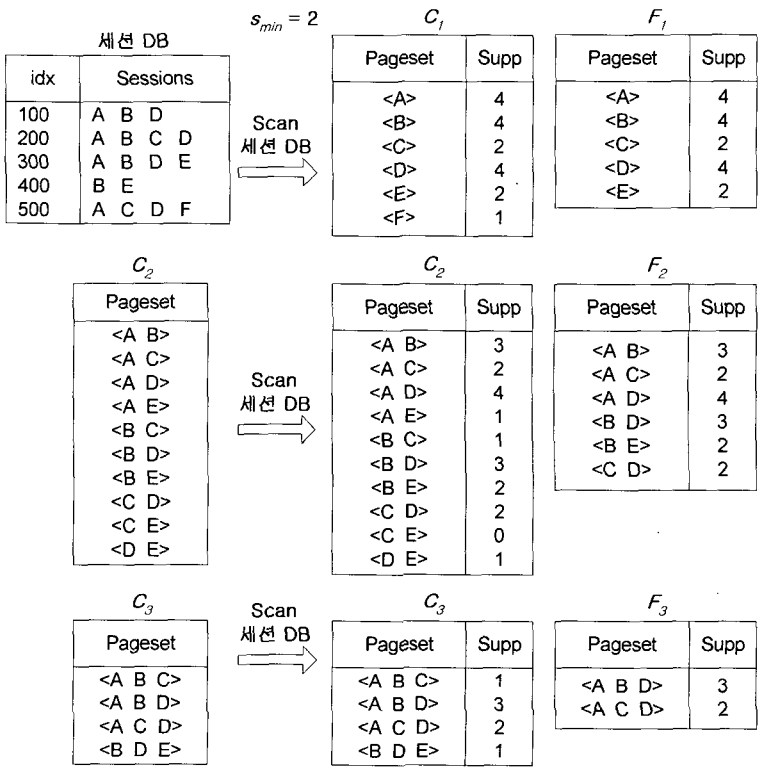


그림 1. AprioriAll의 후보 페이지집합과 빈발 페이지집합 생성 예

결정될 수 있다.

빈발 2-페이지집합들의 집합을 탐사하기 위해서는 하나의 빈발 페이지집합의 한 부분집합이라도 역시 최소 지지도를 가져야 한다는 사실에 입각하여 WebPR(A)는 Apriori-Some 알고리즘을 이용하여 후보 페이지집합들의 집합 C_2를 생성하기 위해 F_1*F_1을 사용한다. 여기서 *는 접합 연산자다. 다음은 세션 데이터베이스를 스캔하여 C_2에 속한 여섯 개의 후보 페이지집합의 지지도를 계산한다. 그림 1의 두 번째 행 가운데 테이블은 C_2에 속한 후보 페이지집합의 지지도 계산결과를 보여준다. 빈발 2-페이지집합들의 집합 F_2는 C_2에 속한 각 후보 2-페이지집합의 지지도에 기초하여 결정된다.

후보 페이지집합들의 집합 C_3는 F_2로부터 다음과 같이 생성된다. F_2에서 첫 번째 항목이 같은 두개의 빈발 2-페이지집합들을 먼저 확인한다. 예를 들어, <A B>와 <A C>에서는 A가 동일한 항목이다. 다음으로 AprioriAll은 <A B>와 <A D>의 두 번째 항목들로 구성된 2-페이지집합 <B D>가 빈발 2-페이지집합들에 속하는지를 검사한다. <B D>가 그 자신이 빈발 집합이므로, <A B D>의 모든 부분 집합들은 빈발하다. 따라서 <A B D>는 후보 3-페이지집합이 된다. WebPR(A)는 세션 DB를 스캔하면서 그림 1에서와 같이 빈발 3-페이지집합을 발견한다. F_3에서 더 이상의 다른 후보 4-페이지집합을 구할 수 없으므로 WebPR(A)는 빈발 페이지집합을 발견하는 과정을 마친다. 여기서 C_k에 속한 각 페이지집합의 지지도를 전체 세션 DB를 스캔하면서 계산해야 하기 때문에 가능한 후보 페이지집합들의 원소들의 개수가 줄여지도록 후보 페이지집합들을 생성하는 것이 중요하다.

2.4.2 추천 페이지집합 생성

WebPR(A)는 F_k 집합을 이용하여 추천 페이지집합 R을 생성한다. WebPR(A)에서 추천 페이지집합 R은 현재 액티브 세션의 길이(l_c) 보다 1이 더 큰 빈발 페이지집합 F_{l_c+1}을 이용하여 생성된다. 예를 들어, 현재 액티브 세션의 길이가 2라고 하면 (l_c=2), F_3을 이용하여 R을 생성한다.

그림 2는 WebPR(A)의 추천 페이지집합 R을 생

idx	Sessions
100	A B D
200	A B C D
300	A B D E
400	B E
500	A C D F

(a)

s_{min} = 2

F_1

Pageset	Supp
<A>	4
<B>	4
<C>	2
<D>	4
<E>	2

F_2

Pageset	Supp
<A B>	3
<A C>	2
<A D>	4
<B D>	3
<B E>	2
<C D>	2

F_3

Pageset	Supp
<A B D>	3
<A C D>	2

(b)

		Test session 1: A -> B -> D	Test session 2: A -> B -> C -> D
PR_A	추천 페이지집합 (R)	{B C D} {D}	{B C D} {D} {}
	Benefit (pages viewed)	1 + 1 = 2	1 + 0 + 0 = 1

(c)

그림 2. PR_A의 추천 페이지집합 생성 예: (a) 세션 DB, (b) PR_A의 반발 세션집합, (c) PR_A의 추천 과정.

성하는 예를 보여준다. 테스트 세션(즉, 액티브 세션)이 A→B→D라고 하면, 먼저 A에서는 F_2로부터 추천 페이지들을 결정한다. 즉, F_2에서 A를 첫 번째 요소로 포함하는 빈발 2-세션들로부터 두 번째 요소들이 추천 페이지들로 결정된다. F_2에는 A를 첫 번째 요소로 포함하는 페이지집합들로 <A B>, <A C>, <A D>가 있다. 따라서 A에서 추천된 페이지집합은 {B C D}가 된다. 마찬가지로 테스트 세션의 A→B에서는 F_3을 이용하여 R을 생성한다. 즉, F_3에서 <A B>를 포함하는 빈발 세션들을 이용하여 R을 생성한다. F_3에는 <A B D>가 있다. 따라서 테스트 세션의 B에서는 {D}가 추천 페이지집합이 된다. 이제 다른 세션 A→B→C→D인 경우를 살펴보자. 먼저 테스트 세션 A에서의 R을 생성하기 위해 F_2를 이용한다. 따라서 위의 예에서와 마찬가지로 A→B까지는 동일하다. 이제 A→B→C를 살펴보자. 그러나 <A B C>를 포함하는 빈발 페이지집합이 존재하지 않기 때문에 추천 페이지집합은 생성되지 않는다.

2.5 WebPR(T) 알고리즘

WebPR(T)는 추천 페이지집합 R을 생성하기 위해 세션의 페이지들 사이에 존재하는 모든 연관성을 이용하는 알고리즘이다. WebPR(T)는 먼저 세션 데이터베이스로부터 메인트리와 서브트리를 생성한다. 즉, 빈발 k-페이지집합 F_k의 집합이 이러한 트리들로 표현된다. 일단 트리들이 생성되면 이러한 트리들을 이용하여 N_{rp}에 따라 추천 페이지집합 R을 생성할 수 있다. WebPR(T) 알고리즘에 관한 자세한 내용은 참고문헌 [11]에 기술되어 있다.

2.6 WebPR(T) 확장

WebPR(T)는 추천 페이지집합 R을 생성하기 위해 세션의 페이지들 사이에 존재하는 모든 연관성을 이용하는 알고리즘이다. 여기서 세션의 처음부터 시작하여 페이지간에 존재하는 모든 연관성을 이용하면 사용자의 성향 파악이 흐려질 수 있다. 따라서 적절한 크기의 윈도우(window) 개념을 적용함으로써 추천 성능을 개선시킴은 물론 수행시간을 크게 단축시킬 수 있도록 WebPR(T) 알고리즘을 확장할 수 있다. 본 논문에서는 WebPR(T) 알고리즘에 윈도우 개념을 적용한 확장된 알고리즘을 winWebPR(T)라

고 표기한다.

winWebPR(T)에서는 적절한 크기의 윈도우를 세션에 적용한다. 여기서 적절한 윈도우 크기를 결정하는 것이 중요하다. 그러나 이것은 적용 도메인(domain)에 따라 다를 수 있기 때문에 시행착오에 의해 결정된다.

그림 3은 winWebPR(T)에서의 윈도우 적용 예를 보여준다. 예를 들어, 윈도우의 크기가 3인 경우 처음 세 페이지까지는 WebPR(T)의 방법과 동일하다. 그러나 세 페이지를 초과하면 윈도우 크기만큼의 페이지집합이 계속 겹치면서 진행된다. WebPR(T)에서는 메인트리와 서브트리를 생성하여 추천에 적용하였지만, winWebPR(T)에서는 메인트리는 생성되지 않고 WebPR(T)의 서브트리와 유사한 하나의 트리만 생성된다.

2.6.1 빈발 페이지집합 탐사

(1) 트리 생성

winWebPR(T)에서의 트리 생성은 윈도우 개념을 사용하여 하나의 트리만 생성한다는 점을 제외하고는 WebPR(T)의 트리 생성방법과 비슷하다. 그림 3에서 보여주는 바와 같이 정해진 윈도우 크기에 따라서 윈도우 크기보다 1만큼 더 깊은 트리를 생성하면 그림 4(b)와 같은 모양의 트리를 생성할 수 있다. 그림 4(b)의 트리는 윈도우 크기가 2인 경우이므로 루트 노드(즉, 더미 노드)를 제외하고 윈도우 크기 2보다 1이 더 큰 깊이 3의 트리가 된다.

(2) 트리 유지관리

winWebPR(T)의 트리 유지관리는 WebPR(T)의 방법과 동일하다. 그림 5는 그림 4(b)에서 생성한 트리의 후보 가지치기를 수행한 결과를 보여준다. 적용 예에서는 최소지지도(S_{min})를 2로 하고 후보 가지치기를 수행한다. 또한 단순성을 위해 적용 예에서는 최소 기간 임계값에 의한 가지치기를 생략한다.

2.6.2 추천 페이지집합 생성

winWebPR(T)에서의 추천은 윈도우 개념을 추가

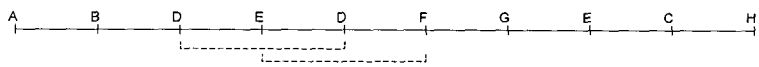
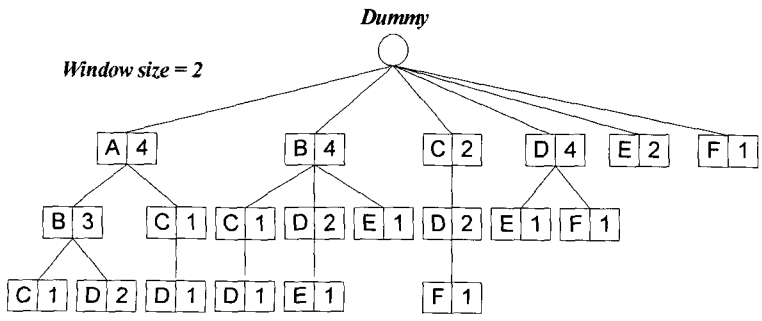


그림 3. winWebPR(T)에서의 예제 세션과 두 개의 윈도우(크기=3)

idx	Sessions
100	A B D
200	A B C D
300	A B D E
400	B E
500	A C D F

(a)



(b)

그림 4. winWebPR(T)의 트리 생성 예: (a) 세션 DB, (b) winWebPR(T) tree.

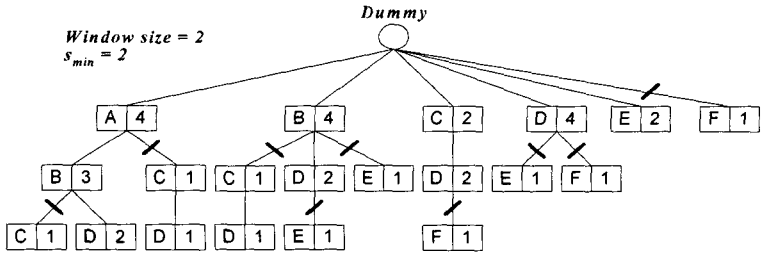
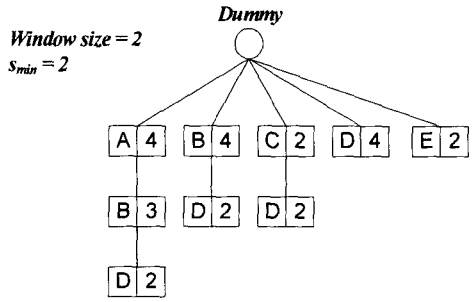


그림 5. winWebPR(T)의 가지치기 수행 예



(a)

구 분		Test session 1: A -> B -> D	Test session 2: A -> B -> C -> D
WinPR_T	추천 페이지집합 (R)	{B} {D}	{B} {D} {}
		{D}	{D} {D}
	Benefit (pages viewed)	1 + 1 = 2	1 + 0 + 1 = 2

(b)

그림 6. winWebPR(T)의 추천 페이지집합 생성 예: (a) winWebPR(T) tree, (b) winWebPR(T)의 추천 과정.

한 것만 제외하고는 WebPR(T)의 추천 방법과 동일하다. 그림 6은 winWebPR(T)의 추천 페이지집합 R을 생성하는 예를 보여준다. 테스트 세션(즉, 액티브 세션)이 A→B→C→D라고 하면, 먼저 A에서는 WebPR(T)에서와 동일하게 A 노드의 자식 노드인 B를 추천한다. 다음은 테스트 세션이 B로 진행하면 역시 마찬가지로 윈도우 크기가 2이기 때문에 A→B의 자식 노드인 D를 추천한다. 테스트 세션이 C까지 진행했을 때 WebPR(T)와 winWebPR(T)의 추천 방법이 달라진다. winWebPR(T)에서는 윈도우 크기가 2이기 때문에 이제 윈도우 크기에 포함되는 B→C 경로만을 이용해서 추천하게 된다. 따라서 트리에서 B→C의 자식 노드들과 C의 자식 노드들을 추천하게 된다.

3. 실험 및 평가

3.1 실험 방법

본 논문에서 수행한 실험 방법은 참고문헌 [11]에서 수행한 방법과 동일하며, 추가된 WebPR(A) 알고리즘과 확장된 winWebPR(T) 알고리즘의 실험결과가 해당 그래프에 포함되어 있는 부분이 다르다.

3.2 실제 데이터

본 논문에서 실험에 적용한 실제 데이터는 참고문헌 [11]에서 적용한 데이터와 동일하다:

LadyAsiana 사이트와 KBSMedia 사이트의 웹로그 데이터.

3.3 실제 데이터를 이용한 WebPR 알고리즘들의 성능 평가

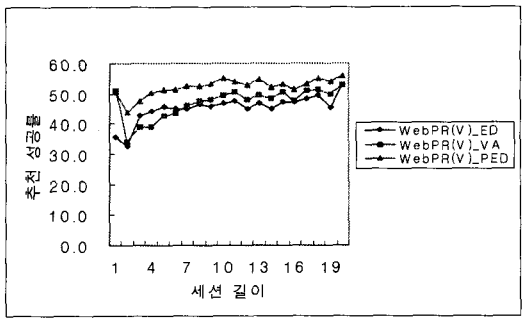
이 절에서는 제안한 PR 알고리즘들을 3.2절에서 기술한 두개의 실제 데이터에 적용한 실험을 수행한다. 실험 방법은 참고문헌 [11]의 3.1절에서 기술한 방법과 동일하다.

3.3.1 클러스터링

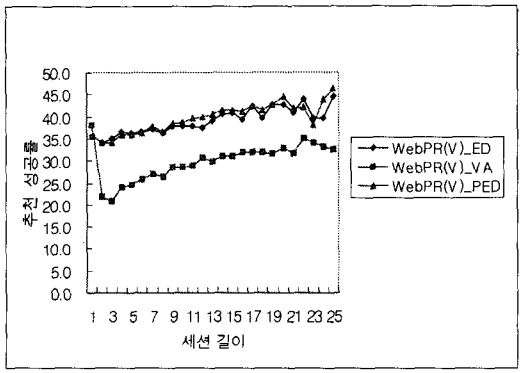
2.1절에서 언급한 바와 같이 WebPR(V)와 WebPR(M)에서는 클러스터링 과정이 요구된다. 본 논문에서는 잘 알려진 K-means 알고리즘을 이용하여 세션 데이터베이스의 세션들을 클러스터링한다. 유사도 ED를 적용한 경우 LadyAsiana는 K=42개, 그리고 KBSMedia는 K=37개의 클러스터들을 생성하였다. 유사도 VA를 적용한 경우 LadyAsiana는 K=66개, 그리고 KBSMedia는 K=39개의 클러스터들을 생성하였다. K-means 알고리즘을 수행할 때 초기 K의 개수는 각 사이트의 전체 페이지 개수로 정하여 시작하였다. 종료조건은 각 클러스터의 세션들이 40개 이하로 변동될 때까지 수행하는 것으로 주었다. 대부분의 경우 6-9 사이클 내에서 수렴하였다.

3.3.2 WebPR(V)_VA, WebPR(V)_ED, WebPR(V)_PED의 성능 평가

WebPR(V)_VA, WebPR(V)_ED, WebPR(V)_PED 중에서 가장 성능이 좋은 모델을 선택하여 WebPR(V)로 표기한다. 가장 좋은 모델 선택을 위해 먼저 각 모델들의 세션 길이별 추천 성공률을 조사하였다. N_{rp}=10인 경우(그림 7)와 N_{rp}=5인 경우에 대해 세가지 모델들의 추천 성공률을 실험하였으나 모든 경우에 N_{rp}=10인 경우의 WebPR(V)_PED 모델이 가장 우수하였다. 실제 N_{rp}=10인 경우의 LadyAsiana에 대하여 평균 추천 성공률은 WebPR(V)_ED가 41.7%, WebPR(V)_VA가 43.3%, 그리고 WebPR(V)_PED가 49.9%를 보여주었다. 반면에 N_{rp}=5인 경우의 평균 추천 성공률은 WebPR(V)_ED가 26.9%, PR(V)_VA가 28.5%, 그리고 WebPR(V)_PED가 41.0%를 보여주었다. KBSMedia의 경우에도 이와 유사한 결과를 보여주었다. 따라서 이후부터 N_{rp}=10인 경우의 WebPR(V)_PED 모델을 PR(V)의 대표 모델로 간주하여 WebPR(V)로 표기한다. 또한 이후 모든 실험 결과는 N_{rp}=10인 경우만을 보여준다.



(a)



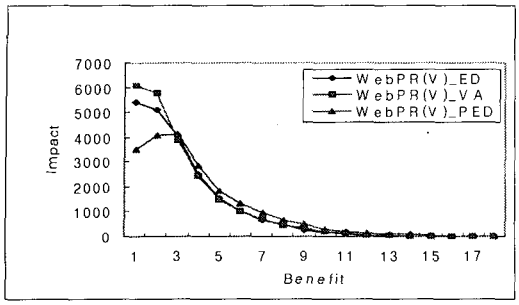
(b)

그림 7. WebPR(V)_ED, WebPR(V)_VA, WebPR(V)_PED의 세션 길이별 추천 성공률: (a) LadyAsiana, (b) KBSMedia.

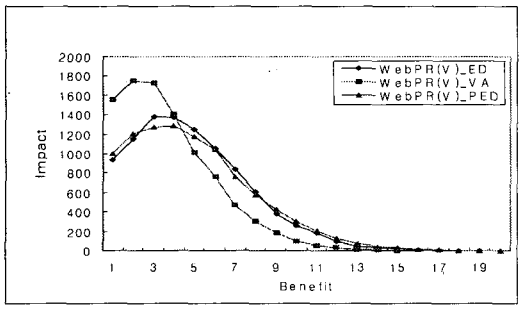
그림 8은 WebPR(V)_VA, WebPR(V)_ED, WebPR(V)_PED의 Impact/Benefit 성능 평가를 보여준다. 그림에서 WebPR(V)_PED의 결과가 가장 우수함을 보여준다.

3.3.3 WebPR(M)_VA, WebPR(M)_ED, WebPR(M)_PED의 성능 평가

WebPR(M)_VA, WebPR(M)_ED, WebPR(M)_PED 능이 좋은 모델을 선택하여 WebPR(M)으로 표기한다. 가장 좋은 모델 선택을 위해 먼저 각 모델들의 세션 길이별 추천 성공률을 조사하였다. 그림 8~9는 N_{rp}=10인 경우의 세 가지 모델들에 대한 추천 성공률을 보여준다. 실제 N_{rp}=10인 경우의 LadyAsiana에 대하여 평균 추천 성공률은 WebPR(M)_ED가 62.2%, WebPR(M)_VA가 63.2%, 그리고 WebPR(M)_PED가 67.2%를 보여주었다. 반면에 N_{rp}=5인 경우의 평균 추천 성공률은 WebPR(M)_ED가 51.0%, WebPR(M)_VA가 53.5%, 그리고 WebPR(M)_PED가 57.0%를

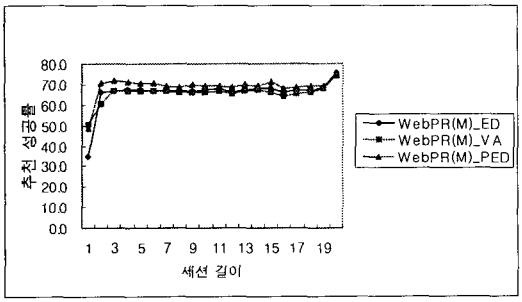


(a)

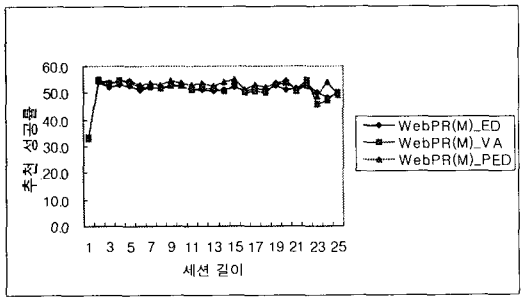


(b)

그림 8. WebPR(V)_ED, WebPR(V)_VA, WebPR(V)_PED의 성능 평가: (a) LadyAsiana, (b) KBSMedia.



(a)

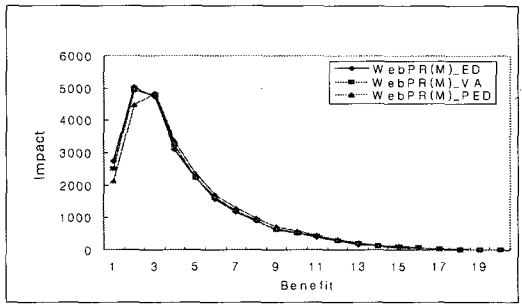


(b)

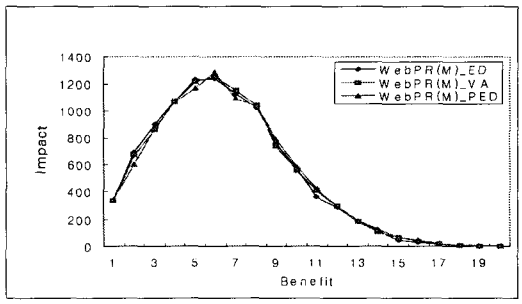
그림 9. WebPR(M)_ED, WebPR(M)_VA, WebPR(M)_PED의 세션 길이별 추천 성공률: (a) LadyAsiana, (b) KBSMedia.

보여주었다. KBSMedia의 경우에도 이와 유사한 결과를 보여주었다. 따라서 이후부터 N_{rp}=10인 경우의 WebPR(M)_PED 모델을 PR(M)의 대표 모델로 간주하여 WebPR(M)으로 표기한다.

그림 10은 WebPR(M)_VA, WebPR(M)_ED, WebPR(M)_PED의 Impact/Benefit 성능평가를 보여준다. 그림에서 WebPR(M)_PED의 결과가 가장 우수함을 보여준다.



(a)



(b)

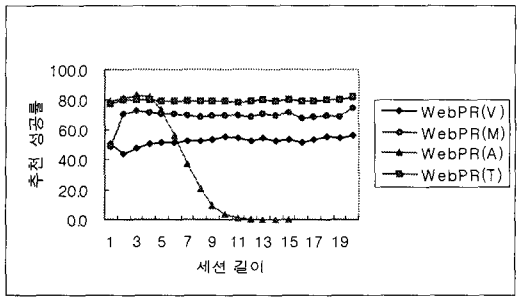
그림 10. WebPR(M)_ED, WebPR(M)_VA, WebPR(M)_PED의 성능 평가: (a) LadyAsiana, (b) KBSMedia.

3.3.4 WebPR(V), WebPR(M), WebPR(A), WebPR(T)의 성능 평가

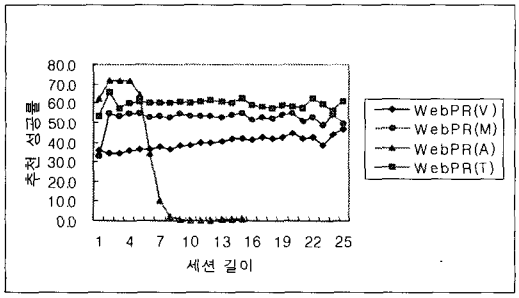
2.4절에서 설명한 바와 같이 WebPR(A)는 AprioriAll 알고리즘을 세션 데이터베이스에 적용시킨 것이다. WebPR(A)에서는 세션에 나타나는 페이지들의 부분집합(즉, k-페이지집합)이 동시에 빈발해야 한다. 따라서 현재 진행되는 액티브 세션이 짧은 경우에는 네 개의 PR 알고리즘 중에서 추천 성능이 가장 좋다. 이러한 특성은 각 사이트의 세션 길이별 분포에 따라서 영향을 받음을 알 수 있다. LadyAsiana 사이트의 경우 세션 길이가 짧은 경우

가 많기 때문에(그림 7 참조), 모든 PR 알고리즘의 추천 성능이 크게 차이가 없다. 그러나 KBSMedia 사이트의 경우와 같이 세션 길이별 분포가 고르게 분산된 모양을 가질 때는 액티브 세션이 짧은 경우에만 WebPR(A)의 Impact/Benefit 성능이 우수하지만, 액티브 세션이 진행될수록(약 6~7 페이지) 급격한 성능 저하를 보인다.

그림 11은 WebPR(V), WebPR(M), WebPR(A), WebPR(T)의 두개의 실제 사이트에 대하여 적용한 세션 길이별 추천 성공률을 보여준다. LadyAsiana의 경우 평균 추천 성공률은 WebPR(V)가 49.9%, WebPR(M)이 67.2%, WebPR(A)가 60.4%, 그리고 WebPR(T)가 79.3%로 WebPR(T)의 결과가 다른 세 개의 모델에 비해 월등한 추천 성능을 보여준다. KBSMedia의 경우 LadyAsiana의 경우보다는 추천 성능이 약간 떨어진다. WebPR(V)가 37.5%, WebPR(M)이 52.0%, WebPR(A)가 30.3%, 그리고 WebPR(T)가 59.9%를 보여준다. 여기서 주목할만한 사실은 네 개의 모델 중에 WebPR(A)의 결과가 가장 저조하다는 것이다. 이것은 LadyAsiana의 경우 세션 길이별 분포가 짧은 쪽에 치우쳐 있으나, KBSMedia의 경우 세션 길이별 분포가 고르게 분포되어 있기 때문에 이러한 결과를 보여주는 것으로 판단된다.



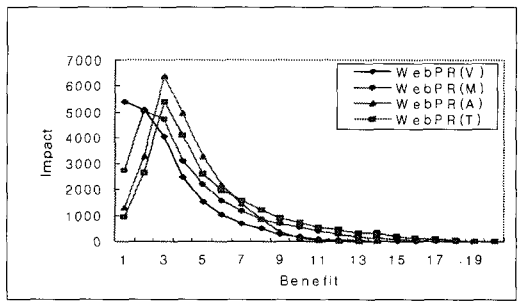
(a)



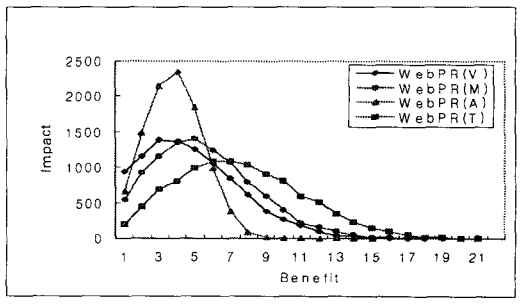
(b)

그림 11. PR 알고리즘들의 세션 길이별 추천 성공률: (a) LadyAsiana, (b) KBSMedia.

그림 12는 WebPR(V), WebPR(M), WebPR(A), WebPR(T)를 두개의 실제 사이트에 적용한 Impact/Benefit 성능평가를 보여준다. 그림에서 알 수 있듯이 두 사이트 모두에서 액티브 세션이 약 6~7 페이지 정도 진행할 때까지는 WebPR(A)의 추천 성능이 가장 우수하지만 그 이후부터는 WebPR(T)의 추천 성능이 가장 우수한 것으로 나타났다. 또한 세션 길이별 분포가 짧은 쪽으로 치우친 LadyAsiana의 경우에는 네 개의 모델들의 추천 성능 그래프가 매우 유사한 모양을 가진다. 그러나 세션 길이별 분포가 골고루 분포된 KBSMedia의 경우는 WebPR(T)의 그래프가 다른 세 개의 그래프와 많은 차이를 보여준다. 이와 같은 결과는 WebPR(T)의 추천 방식이 현재 액티브 세션의 성향을 최대한 활용하여 추천을 수행하기 때문인 것으로 판단된다.



(a)

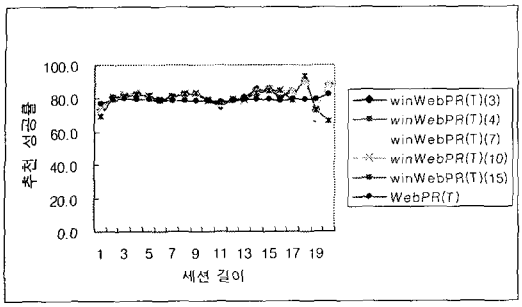


(b)

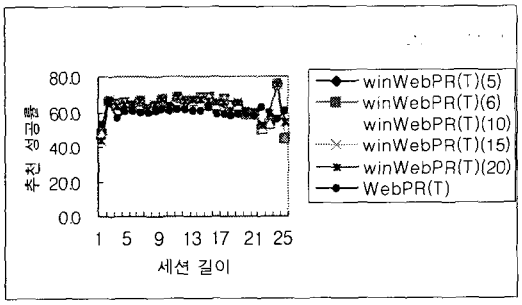
그림 12. WebPR(V), WebPR(M), WebPR(A), WebPR(T)의 성능 평가(N_{rp}=10인 경우): (a) LadyAsiana, (b) KBSMedia.

3.3.5 WebPR(T)와 winWebPR(T)의 성능 평가

그림 13은 WebPR(T)와 winWebPR(T)의 두개의 실제 사이트에 대하여 적용한 세션 길이별 추천 성공



(a)



(b)

그림 13. WebPR(T)와 윈도우 크기에 따른 winWebPR(T)의 성능 평가: (a) LadyAsiana, (b) KBSMedia.

률을 보여준다. LadyAsiana의 경우 평균 추천 성공률은 winWebPR(T)(3)이 80.3%, winWebPR(T)(4)가 80.4%, winWebPR(T)(7)이 80.4%, winWebPR(T)(10)이 80.4%, 그리고 winWebPR(T)(15)가 79.4%로 윈도우 크기가 4인 경우의 winWebPR(T)(4)가 가장 좋은 성능을 보이면서 가장 적은 수행시간을 요구한다. 반면에 KBSMedia의 경우는 winWebPR(T)(6)이 63.3%로 윈도우 크기가 6인 경우에 추천 성능이 가장 좋으면서 가장 작은 윈도우 크기를 가진 것으로 나타났다. 주목할 점은 두개의 실제 사이트 모두에서 윈도우 크기가 커져감에 따라 수행시간이 증가하면서 추천 성능은 오히려 떨어진다는 것이다. 따라서 WebPR(T) 알고리즘에 윈도우 개념을 적용하는 것은 추천 성능과 수행시간 측면에서 매우 효과적임을 알 수 있다.

표 2는 윈도우 크기에 따른 winWebPR(T)와 WebPR(T)의 추천 성능 및 수행시간을 비교한다. 표에서 알 수 있듯이 적절한 윈도우 크기를 결정하는 것은 시행착오에 의해 결정할 수 있으며, 또한 적용 도메인에 따라 달라짐을 알 수 있다.

4. 결 론

본 논문에서는 세션에 나타나는 페이지들간의 연관성 정보를 활용하여 빈발 k-페이지집합을 생성하고, 이를 기반으로 하여 추천 페이지집합을 탐사함으로써 효율적인 웹 정보서비스를 제공할 수 있는 Web Page Recommend (WebPR) 알고리즘[11]을 확장하였다. 확장된 내용으로는 대표적인 빈발 순차패턴 알고리즘인 AprioriAll 알고리즘을 변형 적용한 WebPR(A) 알고리즘을 추가하여 폭넓은 실험을 수행하였으며, WebPR(T)에 윈도우 개념을 도입한 winWebPR(T)을 제안하였다. 실험 결과에서 알 수 있듯이 윈도우 개념을 도입한 winWebPR(T) 알고리즘이 세션에 나타나는 페이지들간의 모든 연관성 정보를 활용함으로써 가장 우수한 성능을 보였다. 이와 같이 추천 페이지집합에 기반하여 웹 사이트를

표 2. 윈도우크기에 따른 winWebPR(T)와 WebPR(T)의 추천성능 및 수행시간 비교

LadyAsiana						
알고리즘	winWebPR(T)					WebPR(T)
윈도우크기	3	4	7	10	15	-
추천성공률	80.3	80.4	80.4	80.4	79.4	79.3
수행시간	3.7	4.3	5.7	6.3	6.7	7.4

KBSMedia						
알고리즘	winWebPR(T)					WebPR(T)
윈도우크기	5	6	10	15	20	-
추천성공률	63.2	63.3	63.3	63.1	63.1	59.9
수행시간	4.4	5.3	6.5	7.3	7.8	8.7

방문한 사용자에게 추천 페이지집합을 포함하는 새로운 페이지뷰(page view)를 제공함으로써 궁극적으로 찾고자하는 목표 페이지에 효과적으로 접근할 수 있도록 한다. 두 개의 실제 웹 사이트로부터 얻은 웹로그 데이터에 적용한 실험 결과에서 알 수 있듯이 페이지간의 연관성 정보를 활용하는 정도가 높을수록 좋은 추천 성능을 보인다.

향후 연구과제로는 페이지간의 연관성 정보를 활용하는 정도를 좀 더 체계적으로 기술하는 것과 다양한 실험을 통하여 WebPR 알고리즘의 특성을 파악하는 것이다. 마지막으로 더 많은 실제 사이트에 적용하는 것과 데모 사이트를 구축하여 WebPR 알고리즘을 적용하는 것이다.

참 고 문 헌

[1] W3C Web Characterization Activity. http://www.w3.org/WCA/, 2003.

[2] J. E. Pitkow, "Summary of WWW characterizations," *Web Journal 2*, pp. 3-13, 1998.

[3] M. Spiliopoulou, "Web usage mining for site evaluation: making a site better fit its users," *Communications of ACM*, 43, pp. 127-134, 2000.

[4] M. C. Drott, "Using web server logs to improve site design," *Proceedings on the Sixteenth Annual International Conference on Computer Documentation*, Quebec, Canada, pp. 43-50, 1998.

[5] M. Perkowitz and O. Etzioni, "Towards adaptive Web sites: Conceptual framework and case study," *Artificial Intelligence*, Vol. 118, pp. 245-275, 2000.

[6] A. Buchner and M. D. Mulvenna, "Discovering internet marketing intelligence through online analytical Web usage mining," *SIGMOD Record*, 27(4), 1999.

[7] B. M. Sarwar, G. Karypis, J. A. Konstan, and J. Riedl, "Analysis of recommender algorithms for e-commerce," *ACM E-Commerce'00 Conference*, Mineapolis, MN, pp. 158-167, 2000.

[8] R. Agrawal and R. Srikant, "Mining Sequential Patterns," *Proc. 11th Int'l Conf. Data Eng.*, Mar. 1995.

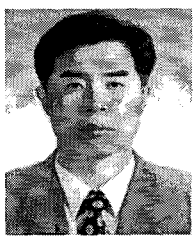
[9] R. Srikant and R. Agrawal, "Mining Sequential Patterns: Generalizations and Performance Improvements," *Proc. Fifth Int'l Conf. Extending Database Technology (EDBT'96)*, Mar. 1996.

[10] T. W. Yan, M. Jacobsen, H. G. Molina, and U. Dayal, "From User Access Patterns to Dynamic Hypertext Linking," The 5th Int'l World Wide Web Conf., Paris, France, May 1996.

[11] 윤선희, 김삼근, 이창훈, "WebPR: 빈발 순회패턴 탐사에 기반한 동적 웹페이지 추천 알고리즘," *한국정보처리학회 논문지*, 제11-B권 제2호, pp. 187-198, 2004.

[12] J. Han and M. Kamber, *Data Mining: Concepts and Techniques*, Morgan Kaufmann publishers, pp. 349-351, 2001.

[13] Z. Chen, W. Liu, F. Zhang, M. Li, and H. J. Zhang, "Web mining for Web image retrieval," *Journal of theAmerican Society for Information Science and Technology*, 52(10), 831-839. 2001.



이 근 수

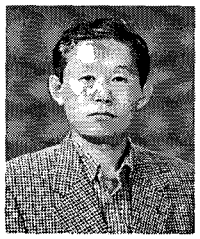
1983년 숭실대학교 전자계산학과 졸업(공학사)
1988년 숭실대학교 대학원 컴퓨터공학과 졸업(공학석사)
1993년 숭실대학교 대학원 컴퓨터공학과 졸업(공학박사)
1992년~1993년 군산대학교 컴퓨터과학과 교환교수
2003년~2004년 미국 George Mason University 전자계산학과 객원교수
1989년~현재 한경대학교 컴퓨터공학과 교수
관심분야 : 패턴인식, 퍼지이론, 컴퓨터비전, 지식기반시스템, 동작이해, 비디오검색

이 창 훈

1987년 광운대학교 전자계산학과 이학사
1989년 중앙대학교 전자계산학과 이학석사
1998년 중앙대학교 컴퓨터공학과 공학박사
1999년~2002년 중앙대학교 정보통신연구소 연구전담교수
2002년~현재 국립 한경대학교 컴퓨터공학과 조교수
관심분야 : 객체지향 소프트웨어공학, 정형화 명세 및 방법, 컴포넌트기반 방법론 등

윤 선 희

1986년 숭실대학교 전자계산학과(공학사)
1988년 숭실대학교 전자계산학과(공학석사)
2003년 숭실대학교 전자계산학과(공학박사)
1992년~현재 미림여자정보과학고등학교 교사
관심분야 : 데이터마이닝, 웹컴퓨팅, 멀티미디어 통신, 멀티미디어 응용 등



이 상 문

1980년 홍익대학교 전산학전공(이학사)
1984년 연세대학교 전산학전공(공학석사)
1993년 홍익대학교 전산학전공(이학박사)
1985년~현재 국립 충주대학교 전자계산학과 교수
2005년~현재 충주대학교 첨단과학기술대 학장
관심분야 : 객체지향 데이터베이스 시스템, 멀티미디어 정보검색시스템

서 정 민

1996년 충주대학교 전자계산학과(공학사)
2000년 충북대학교 전자계산학과(이학석사)
2003년~현재 한경대학교 컴퓨터공학과 박사과정
2001년~2003년 (주)인트컴 연구개발이사
관심분야 : GIS, 웹서비스, 멀티미디어 정보검색, SCM