논문 2005-42CI-6-5

Stochastic 프로세스 모델을 이용한 웹 페이지 추천 기법

(Web Page Recommendation using a Stochastic Process Model)

노 수 호*, 박 병 준**

(Noh Soo Ho and Park Byung Joon)

요 약

다양하고 많은 양의 정보가 존재하는 웹 환경에서 웹 사이트를 방문하는 사용자의 접근패턴도 매우 다양하며, 웹 환경의 변화에 따라서 이러한 접근패턴은 계속 변화한다. 이러한 이유로, 웹 사이트 개발자가 사전에 사용자의 욕구에 완벽하게 부합하는 완벽한 사이트를 개발하기란 사실상 불가능하다. 이에 대한 해결방안으로, 웹 사이트에 대한 사용자 접근 패턴을 학습해서 웹 사이트의 구조나 외형을 자동적으로 개선시켜 나가는 적응형 웹 사이트 (Adaptive Web site)가 제시되었다. 본, 논문에서는 DTMC(Discrete-Time Markov Chain)에 의거한 확률적 모델을 이용하여 적응형 웹 사이트 구축에 필요한 사용자 접근패턴을 학습하고 이를 적용하기 위한 효과적인 방법론을 제시한다.

Abstract

In the Web environment with a huge amount of information, Web page access patterns for the users visiting certain web site can be diverse and change continually in accordance with the change of its environment. Therefore it is almost impossible to develop and design web sites which fit perfectly for every web user's desire. Adaptive web site was proposed as solution to this problem. In this paper, we will present an effective method that uses a probabilistic model of DTMC(Discrete-Time Markov Chain) for learning user's access patterns and applying these patterns to construct an adaptive web site.

Keywords: 웹 마이닝, 적응형 웹 사이트, 패턴발견, 페이지 추천기법

I. 서 론

WWW 환경에서 하나의 웹 사이트는 하이퍼링크로 연결 되어 있는 수많은 HTML 문서들의 집합으로 이루어져 있다. 초기의 웹 사이트는 각 HTML 문서들이 지닌 의미와 문서들 간의 상호관계 등을 고려해 최상의 웹 사이트를 구현하고자 하는 웹 마스터의 의도가 반영된 것이다. 하지만 동적으로 변화해가는 사용자들의 요구를 반영하여 보유정보를 효과적으로 제공할 수 있도록 지속적으로 웹 사이트의 설계를 바꾸어 가는 일은 매우 어려운 일이다. 하나의 웹 사이트에 대한 사용자들의 접근 로그 데이터는 이 사이트에서 일반적으로 사용자들이 보여주는 정보접근 패턴을 알아낼 수 있는 중요한 자료를 제공한다. 그리고 이러한 사용자 접근패턴을 바탕으로 보다 효과적으로 보유정보를 제공할 수 있도록 기존 웹 사이트의 구조와 표현방식을 개선시킬 수 있을 것이다. 적응형 웹 사이트란 이와 같이 사용자 접근패턴을 학습해서 사이트의 조직이나 외형을 자동적으로 개선시켜 나가는 사이트로 기존 연구들은 적응형 웹 사이트에 접근하는 방법을 크게 맞춤화(customization)와 최적화(optimization)로 구분하고 있다. 맞춤화는 개인의 접근 패턴을 학습한 후 개인의 요구와 기호를 고려해서 사이트를 개선시켜 나가는 방식이다. 이러한 서비스를 제공하기 위해서 시스템은 사용자 개개인의 정보를 수집해야 하므로 많은 시간을 허비하게 되고 프라이버시가 침해당할 수 있다는 문제점을 안고 있다. 이와는 달리, 최적화는 특정 개인이 아니라 접근했던 모든 사용자들, 심지어 이후 접근할 가능성을 지닌 사용

* 학생회원, ** 정회원, 광운대학교 컴퓨터과학과
(Dept. of Computer Science, Kwangwoon University)
접수일자: 2005년2월3일, 수정완료일: 2005년11월3일

자들을 위해 사이트 자체를 바꾸는 형태로써 이미 접근한 모든 유저들을 대상으로 접근 패턴들을 학습해 나가는 방식이다. 적응형 웹 사이트를 구축하기 위해서는 사용자들이 사이트와 어떻게 상호작용 하는지를 관찰해야하며 이러한 관찰단계에서 웹 문서 접근 빈도수, 링크 운행 경로, 기타 발생된 문제들이 어떤 것이 있는지 등을 모니터링 하게 된다. 관찰은 일반적으로 웹 서버 로그들을 분석함으로써 이루어지는데, 서버 로그에 들어있는 IP address, 웹 문서에 접근한 날짜와 시간, 요청한 URL 등을 분석함으로써 웹 문서의 접근 빈도수 뿐 만 아니라 시간 의존적인 경향들을 알아낼 수 있다[1]. 본 논문에서는 최적화 기법을 사용하는 적응형 웹 사이트를 구축하기 위한 효과적인 웹 로그 접근 패턴 분석 및 적용 방안을 제시 한다

연관규칙 탐색 기반의 웹 로그 패턴 분석 방식들을 이용한 기존의 웹 페이지 추천방식은 사용자의 접근패턴을 분석하여, 현재 사용자가 참조하고 있는 페이지를 기준으로 관련 있는 페이지 혹은, 한정된 페이지 순회수 이후의 사용자가 참조할 만한 연관성 있는 페이지를 추천함으로써 적응형 웹 사이트가 이 페이지의 링크를 추가 할 수 있게 한다[2]. 그러나 이러한 방식은 사용자가 최종 목적 페이지 참조를 위하여 불필요하게 중간 페이지들을 참조하게 되는 가능성을 여전히 가지고 있으며, 실시간으로 바뀌는 접근패턴을 적용하려면 페이지간의 연관성 및 순차적 패턴을 찾는 알고리즘을 매번 다시 적용하여야 하기 때문에 접근 패턴을 항상 최신의 것으로 유지하기에는 무리가 있다. 또한 방대한 양의 웹 페이지간의 연관성 및 순차적 패턴들을 유지해야 하므로 긴 경로의 순차적 패턴을 유지하기에는 저장 공간의 한계에 부딪히게 된다.

본 논문에서는 DTMC(Discrete-Time Markov Chain)기반의 확률적 모델을 사용, 실시간으로 업데이트되는 패턴을 사용하여 현재 페이지를 기준으로 사이트의 규모에 따라서 결정되는 N 단계 이후 예상되는 참조 웹 페이지를 추천한다. 따라서 최신의 정보를 사용하여 관련성은 높으나 접근경로가 긴 문서들을 효과적으로 추천할 수 있다는 특징을 지닌다.

본 논문은 총 V장으로 구성되어 있다. II장에서는 웹 로그 분석 기법 및 관련 연구를 소개하고, III장에서는 Stochastic 프로세스 모델을 이용한 웹 페이지 추천 기법에 대해 서술한다. IV장에서는 본 연구의 실험결과를 기술하고 V장에서 결론을 맺는다.

II. 웹 로그 분석 기법 및 관련 연구

본 장에서는 웹 로그 분석 기법을 소개하고 기존의 웹 사용자 패턴 발견 기법에 대해 설명한다.

1. 웹 로그 분석 기법

웹 로그 분석 기법은 로그로부터 액세스 패턴이나 의미 있는 행동 등의 유용한 정보를 자동적으로 발견하는 것을 목적으로 한다. 사용자가 어떤 사이트를 방문한 경우 로그파일에 그 흔적이 남게 되는데, 웹 로그 분석 기법은 이러한 로그파일을 기반으로 방문자의 트래픽 정보, 방문 경로 정보, 방문자의 시스템 환경 정보, 사이트 열람 정보 등을 얻을 수 있는 분석 기법이다. 그래서 로그파일의 분석 결과는 웹 사이트 내 가장 자주 접근되는 페이지나 사용자의 접근 패턴 등을 파악하여 웹 사이트의 정보를 효과적으로 전달하기 위한 방법으로 이용자가 최적의 환경에서 사이트를 접근하도록 시스템 성능을 개선하기 위해 활용된다.

웹 로그 분석 기법은 전처리, 패턴 발견, 패턴 분석의 세 가지 과정으로 이루어진다. 그림 1은 웹 로그 분석 기법 과정을 나타낸 것이다.

첫째, 전처리 과정(preprocessing)은 여러 가지 형태의 웹 로그 데이터를 마이닝 기법에 적용하기 위해 적절한 정제, 분류 등의 작업을 실행한다. 둘째, 패턴 발견 과정(pattern discovery)은 변형된 여러 가지 로그 데이터를 가지고 패턴 마이닝 엔진을 실행한다. 발견된 패턴의 종류는 단순한 통계적 분석, 연관 규칙, 군집, 분

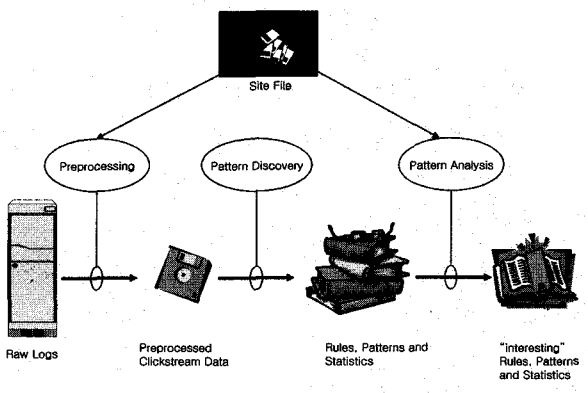

그림 1. 웹 로그 분석 기법
Fig. 1. Web Log Analysis Technique

류, 순차 패턴 등이 있다. 셋째, 패턴 분석과정(pattern analysis)은 발견된 패턴 정보를 사용자가 보다 이해하기 쉽도록 패턴을 표현하는 작업이다. 여기에는 통계, 시각화, 유용성 분석, 데이터베이스 쿼리 등의 기법이 포함된다[3].

2. 패턴발견 관련 연구

기존의 웹 로그 패턴발견 기법에는 연관 규칙(association rules), 순차적 패턴(sequential patterns) 및 경로분석(path analysis) 등이 있다.

가. 연관규칙(Association Rules)

전통적으로 연관 규칙은 항목 집합으로 표현된 트랜잭션에 각 항목간의 연관성을 반영하는 규칙으로 웹 마이닝에서는 방문하는 페이지들 간의 연관성을 발견하는 기법이다. 예를 들어, 방문한 URL의 집합으로 이루어진 트랜잭션들을 분석한 결과, 자주 A페이지를 방문한 고객은 B페이지도 방문한다 라는 연관규칙을 발견한다. Agrawal이 처음 소개한 이후, 데이터베이스 검색 횟수를 줄이거나, 주기억 장치의 한계를 없애는 등의 발전된 알고리즘들이 발표되어 왔다. 연관규칙을 탐사하는 문제는 기본적으로 미리 결정된 최소 지지도 이상의 트랜잭션 지지도를 가지는 항목집합들의 모든 집합들인 빈발항목집합들(large item sets)을 찾아내어 연관규칙을 생성하는 단계로 이루어진다. 대표적인 알고리즘은 Apriori, DHP, Sampling, FUP, DIC 등이 있다[2][3][4][5][6].

나. 순차적 패턴(Sequential Patterns) 및 경로 분석(Path Analysis)

순차적 패턴은 연관 규칙과 유사하나 페이지간의 시간의 순서가 존재한다. 예를 들어, 자주 A페이지를 방문한 이후에 B페이지를 방문한다는 정보이다. Mannila는 웹 로그로부터 페이지 접속을 이용했다. 즉 하나의 페이지 접속을 이벤트로 보고, 접속 순으로 정렬한 이벤트 열을 생성하여 자주 발생하는 에피소드 (이벤트의 조합)를 발견하였다. 기타 대표적인 알고리즘은 AprioriAll, AprioriSome, DynamicSome, GSP등이 있다[7].

그리고 경로 분석은 웹 페이지 상의 페이지들의 클릭 흐름을 나타낸 것이다. Chen[8]은 이용자의 네비게이션 패턴을 분석하기 위한 트랜잭션(transaction)을 얻기 위해 Maximal Forward Reference라는 개념을 처음 소개했다. Maximal Forward Reference는 이용자가 웹 브라우저 상의 백트래킹(backtracking)이 발생하기 전까지의 맨 마지막 페이지까지를 하나의 트랜잭션으로 간주하는 방법이다. 예를 들어 이용자의 페이지 방문이 A-C-A-D-B-D의 순서로 이루어졌을 때 Maximal Forward Reference는 A-C와 A-D-B가 된다.

3. 사용자가 마지막으로 방문할 페이지의 추천

본 논문은 현재 페이지를 기준으로 단순히 관련 있는 페이지를 추천하는 기법에서 한 층 더 나아가서 사용자가 마지막으로 참조할 페이지를 추천하는 기법을 제안한다. 여기서는 사용자의 웹 사이트 사용에 있어서의 특징과 기존의 웹 로그 패턴 발견기법의 한계점, 본 논문의 페이지 추천기법에서 사용한 Stochastic 프로세스 모델에 대해 기술한다.

가. 웹 사이트 사용자의 페이지 참조 형태

일반적으로 어떠한 사이트에 방문한 사용자는 자신이 원하는 정보를 습득하기 위하여 각 페이지의 링크를 따라서 순회하게 되고 특정 페이지에서 사용자가 원하는 정보를 찾으면 그 사이트에서의 웹 서핑을 종료하게 된다. 대표적인 예로 인터넷 상거래 사이트의 경우 방문자는 자신이 원하는 물품이 있는 페이지에 도달하면 이 물품을 구매한 후 그 사이트에서의 웹 서핑을 종료한다. 또한, 입학을 하기위해 학교의 사이트에 접속한 지원자는 입학접수 페이지에 도달하면 자신의 목적을 달성하게 되고 마찬가지로 그 사이트에서의 웹 서핑을 종료한다. 따라서 웹 사용자가 마지막으로 참조할 만한 페이지를 사용자에게 추천하는 것은 대체적으로 웹 사용자의 목적 페이지를 추천하는 것과 같다.

나. 기존의 웹 로그 패턴발견 기법의 한계점

기존의 웹 로그 패턴 발견 기법들을 적용한 페이지 추천방식은 위에서 언급한 사용자의 목적 페이지를 한 번에 추천하기에는 몇 몇 한계점을 가지고 있다. 먼저 연관규칙을 적용한 경우 페이지간의 시간적인 순서를 고려하지 않는 단순히 연관성이 있는 페이지들만을 추천하는 것이 가능하며, 순차적 패턴 및 경로 분석을 적

표 1. 웹 페이지 추천 기법의 비교
Table 1. Comparison of Web Page Recommendation Techniques.

비교항목 / 사용된 기법	최종 페이지 추천	요구되는 저장 공간	지속적인 패턴의 업데이트
연관규칙	고려안함	중	곤란
순차적 패턴	제한적으로 가능	대	곤란
제안하는 추천 기법	가능	소	용이

용한 경우는 페이지들 간의 시간적인 순서를 고려한 페이지 접속의 이벤트 열을 찾아냄으로써 일정 페이지 참조 수 이후의 예상되는 페이지를 추천하는 것이 가능하나, 방대한 이벤트 열을 유지하기에는 막대한 공간적 비용을 요구하므로 이벤트열의 수나 최대 길이는 제한되어 있다고 할 수 있다. 따라서 순차적 패턴 및 경로 분석 기법 역시 사용자의 최종방문 페이지를 추천하기에는 여전히 한계점을 가지고 있다.

본 연구에서 제안하는 페이지 추천기법과 기존의 페이지 추천기법과의 차이점을 표로 나타내면 표 1과 같다.

다. Stochastic 프로세스 모델과 DTMC

사용자의 최종방문 페이지를 추천하기 위하여 본 논문에서는 Stochastic 프로세스 모델의 DTMC를 사용한다.

(1) Stochastic 프로세스 모델

Stochastic 프로세스는 랜덤하게 전개되는 시스템의 확률적 모델이다. 시스템이 구분된 시간점 n(= 0, 1, 2, ...)에서 관찰되어 지며, X_n은 시간점 n에서의 시스템의 상태라고 하면, $\{X_n, n\geq 0\}$은 비연속적 stochastic 프로세스로써 나타내어 질 수 있다. n번째 주말의 Dow-Jones 주가, 자동차 판매회사에서 n번째 주초에 아직 판매 되지 않은 자동차 대수, 금세기에 미대륙에 발생한 n번째 지진의 진도, 어떤 도시에 n번째 날짜에 발생한 범죄량 등이 비연속적 stochastic 프로세스의 예라고 할 수 있다.. 반면에 시스템이 시간 t에서의 상태에 해당하는 X(t)로써 연속적인 시간동안 관찰된다면 이 시스템은 시간-연속적 stochastic 프로세스로써 나타내어 질 수 있다. 예를 들어서 시간-연속적 stochastic 프로세스 X(t)는 시간 t에서 어떠한 가게에서 고장난 기계의 수, 시간 t에서의 지도상의 허리케인 위치, 시간 t에서의 은행 계정에 남아있는 잔고 등을 나타낼 수 있다. 본 논문에서는 사용자의 웹 링크 클릭을 단위로 한 비연속적인 웹 페이지 순회의 특성상, 비연속적 stochastic 프로세스를 사용하게 된다.

(2) DTMC(Discrete-Time Markov Chain)

n을 현재라고 가정하면 X_n은 현재의 시스템의 상태이고, $\{X_{n+1}, X_{n+2}, ...\}$은 미래의 시스템의 상태라고 할 수 있다. 또한 $\{X_0, X_1, ..., X_{n-1}\}$은 과거의 시스템의 상태가 된다. 이 경우 다음의 정의를 만족하는 비연속적 stochastic 프로세스 $\{X_n, n\geq 0\}$를 DTMC라고 말한다.

$$P(Xn+1 = j|Xn = i, Xn-1 = in-1, \dots ,X0 = i0) = P(Xn+1 = j|Xn = i) \quad (1)$$

즉, 시스템의 미래의 상태는 시스템의 과거와는 독립적이며 오직 시스템의 현재 상태와 연관이 있는 stochastic 프로세스가 DTMC이다. $P(X_{n+1} = j|X_n = i)$를 $P_{ij}(n)$으로 표현하면, DTMC가 아래의 성질을 만족할 경우 time-homogeneous하다고 한다.

$$Pij(n) = Pij \text{ for all } n\geq 0 \quad (2)$$

즉, time-homogeneous한 DTMC는 시스템의 상태가 i에서 j로 전이할 확률이 시간 n에 독립적이다. time-homogeneous한 DTMC에서 우리는 상태전이행렬 P를 생성할 수 있다. 시스템의 상태들의 집합 S = {1, 2, ..., m}일 때 상태전이 행렬 P는 다음과 같은 형태를 띈다.

$$\mathbf{P} = \begin{bmatrix} P_{11} & P_{12} & \cdots & P_{1,m-1} & P_{1m} \\ P_{21} & P_{22} & \cdots & P_{2,m-1} & P_{2m} \\ \vdots & & & & \vdots \\ P_{m-1,1} & P_{m-1,1} & \cdots & P_{m-1,m-1} & P_{m-1,m} \\ P_{m1} & P_{m2} & \cdots & P_{m,m-1} & P_{mm} \end{bmatrix} \quad (3)$$

실세계의 많은 시스템들은 time-homogeneous한 DTMC로 모델링 될 수 있다.

(3) DTMC(Discrete-Time Markov Chain)의 주변밀도(Marginal Distribution)

DTMC의 초기상태(initial distribution) a(0) = a, 상태전이행렬(state transition matrix) P가 있을 때, DTMC를 따르는 시스템에서 n번째 step이후의 시스템의 상태는 아래의 식을 통하여 랜덤변수 Xn의 주변밀도를 구함으로써 예측할 수 있다[9].

$$a^{(n)} = a^{(0)}P^{(n)} = aP^n \quad (4)$$

Ⅲ. Stochastic Process 모델을 이용한 웹 페이지 추천

1. 웹 페이지 추천 기법의 개요

본 논문에서는 DTMC에 의거한, 확률적 모델을 적용하여 임의의 n번의 페이지 참조 후에 사용자가 방문할 가능성이 가장 큰 페이지를 추천하는 기법을 제시한다.

이 방식은 기존의 연관규칙 기반의 웹 페이지 추천기법과 달리 서로 연관성이 있는 웹 페이지들의 집합이나, 순차적 패턴들을 찾아내고 이를 사용하여 웹 페이지를 추천하는 것이 아니라, 웹 페이지간의 상호참조 빈도 확률을 행렬에 저장하고 이를 연산하여 웹 페이지를 추천한다. 따라서 처음 제시한 임의의 n번의 페이지 참조 후에 사용자가 방문할 가능성이 가장 큰 페이지를 추천한다는 장점 외에도, 전체적인 웹 페이지간의 연관성에 관한 지식을 저장할 필요가 없이 몇 몇 행렬들만 유지하면 되기 때문에 비교적 작은 저장 공간을 요구한다는 장점을 가진다. 이 행렬을 최신의 정보의 것으로 유지하는 것도 간단하다. 웹 사이트로 사용자의 추가적인 접근이 이루어질 경우 이러한 접근기록이 추가된 웹 로그를 참조하여 별도의 알고리즘을 적용할 필요가 없이 페이지 상호 참조 빈도수를 조정함으로써 행렬을 업데이트하는 것이 가능하다.

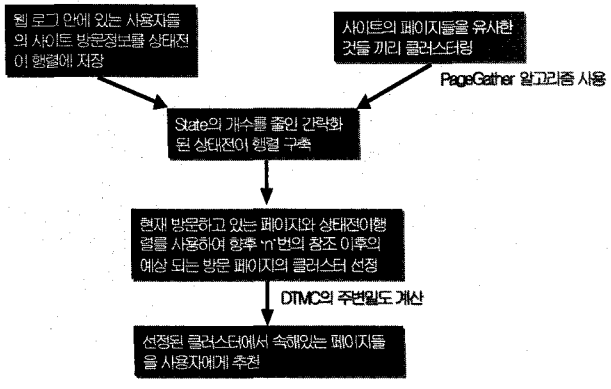

그림 3. 시스템 흐름도
Fig. 3. Flow Chart of System.

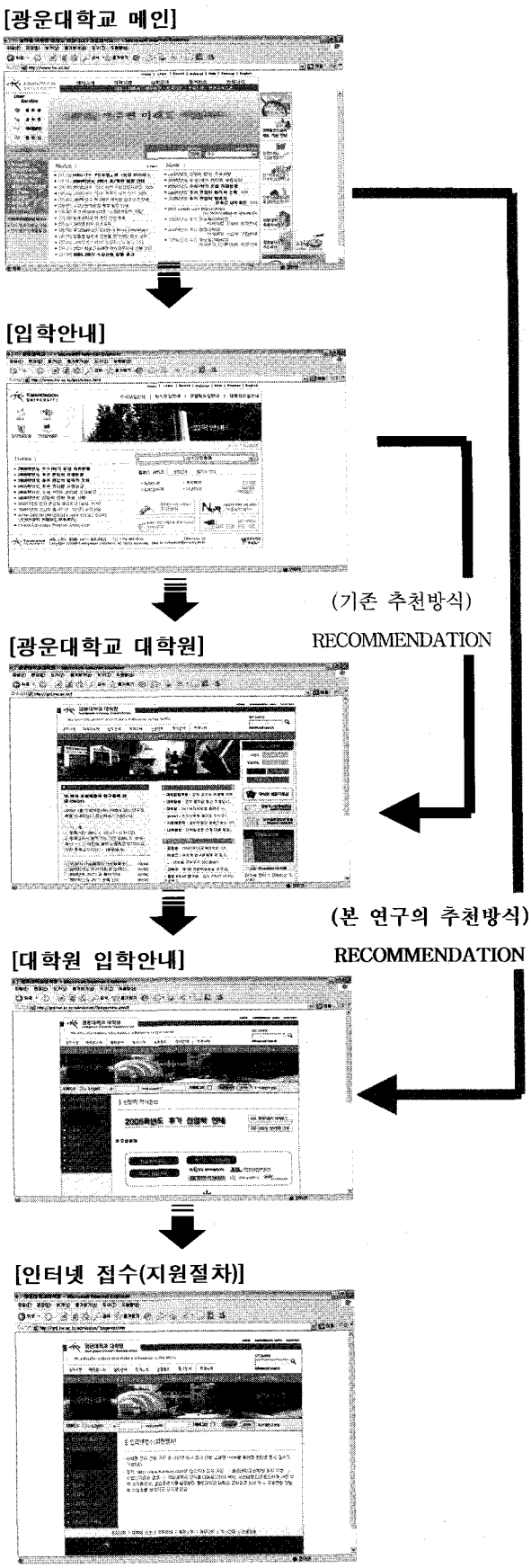

그림 2. 마지막으로 방문되어지는 페이지의 추천
Fig. 2. Recommendation of Page Finally Visited.

그러나 사이트를 구성하고 있는 웹 페이지의 수가 많아질수록 상태전이행렬(state transition matrix)의 지수 연산은 큰 오버헤드를 동반하게 된다. 이러한 오버헤드를 감소시키기 위하여 PageGather 알고리즘을 사용하여 사이트를 구성하고 있는 웹 페이지 간에 유사성이 큰 것들을 클러스터링(clustering)하고, 상태전이행렬을 구성하는 총 state의 수를 감소시켜 그 오버헤드를 줄인다. 이 시스템의 구조는 그림 3과 같다.

2. 웹 페이지 추천 과정

가. 사용자의 웹 페이지 참조 빈도수 조사

웹 로그에 저장되어있는 클라이언트 IP, 시간, 요구되어진 페이지를 분석하여 페이지들 간의 상호 참조 빈도수를 구한다. 그림 4는 웹 로그를 참조하여 같은 세션 안에서 엑세스한 웹 페이지들의 상호참조 빈도수를 구하는 예를 보여준다.

다음 이 상호참조 빈도수를 사용하여 현재 참조하는 페이지를 상태로 하는 상태전이행렬을 구축한다. 그림 5는 5개의 페이지를 보유하고 있는 웹 사이트에서의 상태전이행렬을 구축한 예이다.

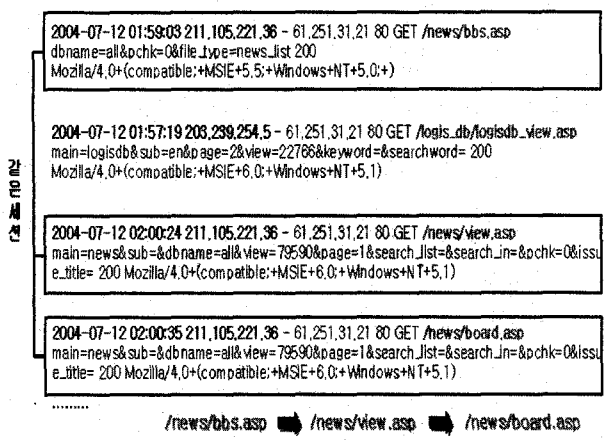

그림 4. 웹 로그 파일의 분석
Fig. 4. Analysis of Web Log File/

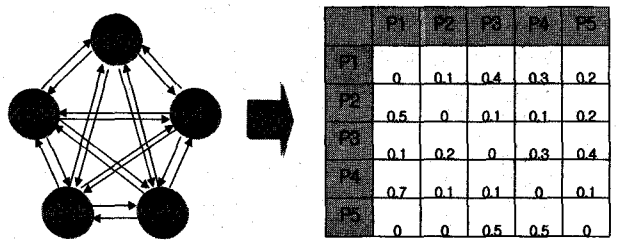

그림 5. 상태전이행렬 구축
Fig. 5. Construction of State Transition Matrix.

나. 페이지들의 클러스터링

다음단계에 수행할 행렬의 지수계산을 위한 오버헤드를 줄이기 위하여 상태전이행렬을 간소화 시켜야 한다. 이를 위하여, 각 상태에 해당하는 페이지들을 서로 관련 있는 것끼리 클러스터링하여 상태수를 감소시킨다. 본 논문에서는 클러스터링 기법으로써 PageGather 알고리즘을 채택하였다. PageGather 알고리즘은 페이지들의 상호참조 빈도수를 기반으로 하여, 유사한 페이지들을 클러스터링 하는 기법이며, 이 알고리즘은 아래의 세 가지 기본 단계를 거친다.

① 방문한 사용자의 접근 로그를 처리한다.

② 페이지들 간의 상호참조 빈도수를 계산하고 이를 기반으로 상태전이행렬을 생성한다.

③ 위의 행렬과 관련 있는 그래프를 생성하고 이 그래프에서 clique(cluster)를 찾아낸다.

PageGather 알고리즘은 웹 문서를 클러스터링 할 경우 K-Means 알고리즘이나 HAC 알고리즘에 비해서,

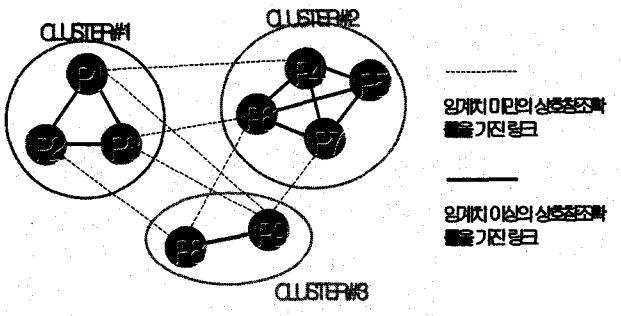

그림 6. PageGather 알고리즘을 이용한 웹 페이지 클러스터링
Fig. 6. Web Page Clustering Using PageGather Algorithm.

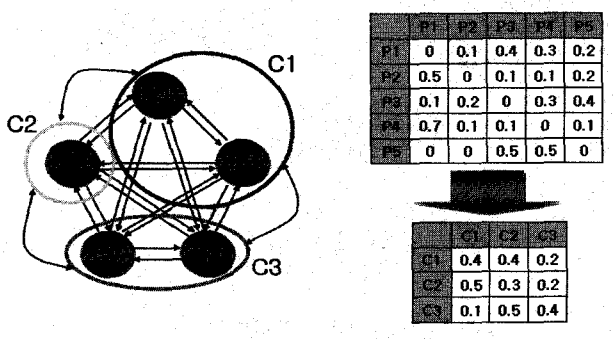

그림 7. 상태전이행렬의 상태의 개수 감소
Fig. 7. Reduction of The Number of Transition Matrix's States.

매우 빠른 수행속도를 보인다. 따라서 웹 구조상에 존재하는 많은 수의 문서들을 클러스터링 할 경우 사용할 수 있는 적절한 알고리즘이라고 할 수 있겠다[10]. 그림 6은 PageGather 알고리즘의 적용 예이다.

다. 상태전이 행렬의 간소화

모든 페이지들 대신에 나 항에서 구성한 클러스터들을 상태로 하는 간략화 된 상태전이행렬을 구축 한다. 그림 7은 상태의 개수가 감소된 상태전이행렬을 생성한 예를 보여준다. 클러스터링을 하기 전에는 5개 페이지 모두에 대한 상태를 상태전이행렬에 반영하기 위해서 5×5행렬이 필요했으나 클러스터링 후에는 3개의 클러스터에 대한 상태만을 상태전이행렬에 반영하면 되므로 3×3행렬만으로도 충분히 상태전이행렬을 구축할 수 있게 된다.

라. 방문할 페이지의 클러스터 선정

현재 방문하고 있는 페이지와 상태전이행렬을 사용하여 향후 'n' 단계 이후 방문할 것으로 예상되어지는 클러스터를 선정한다. 식 (4) ($a^{(n)} = a^{(0)}P^{(n)} = aP^{(n)}$)를 사용하여 현재 사용자가 참조하는 페이지가 속해있는 클러스터 a(처음상태)를 기준으로 n-step의 페이지 참조 이후 참조할 가능성이 가장 큰 페이지의 클러스터 (X_n의 확률밀도)를 예측한다. 현재 3번째 클러스터 안에 포함되는 페이지를 참조할 경우 처음상태 'a'는 벡터 (0, 0, 1) 이며, n-step의 페이지 참조 후에 사용자가 방문할 것으로 예상되는 페이지의 클러스터는 아래 식의 결과로 나온 벡터의 element중에서 가장 큰 값을 가진 것의 index이다.

$$(0,\ 0,\ 1) \quad \times \quad \begin{bmatrix} 0.4 & 0.4 & 0.2 \\ 0.5 & 0.3 & 0.2 \\ 0.1 & 0.5 & 0.4 \end{bmatrix}^{n} \qquad (5)$$

마. 사용자에게 선정된 클러스터를 추천

라 항에서 선정한 클러스터에 속해있는 페이지들의 링크를 현재 사용자가 참조하고 있는 페이지에 동적으로 첨부함으로써 본 시스템이 추천하는 사용자의 최종 목표 페이지들을 보여준다.

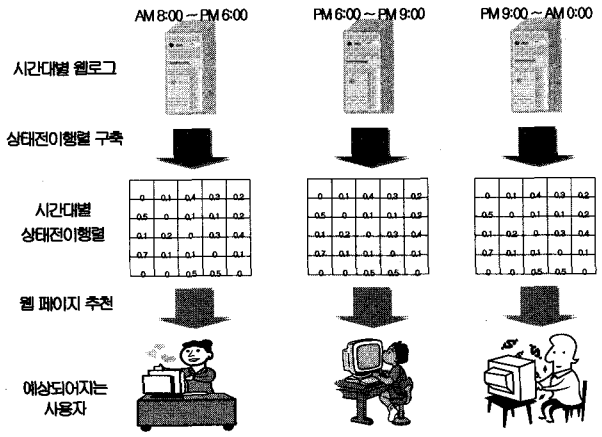

그림 8. 시간별 상태전이행렬을 이용한 페이지 추천
Fig. 8. Recommendation of Web Page with Each Time's State Transition Matrix.

바. 시간대별 상태전이행렬 구축

웹 사용자의 웹 페이지 접근 패턴은 시간대별로 상이하다. 쇼핑몰 사이트의 예를 들어보면, 주간에는 주부들이 선호하는 물품들에 대한 빈번한 웹 페이지 접근 패턴을 보일 것이며, 야간에는 회사원이나 학생이 선호하는 물품들에 대한 빈번한 웹 페이지 접근 패턴을 보일 것이다. 본 논문에서 사용한 DTMC는 시간대에 상관없이 언제나 상태전이확률은 일정하다는 전제조건을 가지고 있다. 이러한 모순점을 해결하기 위하여 본 논문에서는 시간대 별로 여러 개의 상태전이행렬을 구축하여 실제 웹 페이지 추천 시 각 시간대 별로 적절한 상태전이행렬을 활용할 것을 제안한다.

Ⅳ. 실험결과

본 논문의 실험에서 사용한 웹 로그는 가격비교사이트 (www.bb.co.kr)에서 추출한 것이다. 이 사이트는 시간당 약 2만8천건의 페이지 요청을 받는다. 처음 5시간 동안의 웹 로그를 training 데이터로 사용하였으며 이후 5시간동안의 웹 로그를 test데이터로 사용하였다. 이 웹 로그 분석시 총 2032개의 페이지가 발견되었다. 실험에서는, 2032개의 페이지를 사용해서 상태전이행렬을 구성하였으며, 2032개 페이지 모두를 현재 방문하고 있는 페이지로 하여 실험을 할 경우 시간이 너무 많이 소요되는 관계로 임의로 선택된 200개의 페이지만을 사용자가 현재 방문하고 있는 페이지 a로써 사용되었다. 사용자가 페이지 a를 참조하고 있을 때 본 논문에서 제시한

표 2. Step별 추천한 페이지의 accuracy
Table 2. Each Step's Accuracy of Web Page Recommendation

Step	1-step	5-step	10-step
Accuracy	33.8%	40.5%	21.2%

표 3. 클러스터링 이전/이후에 추천한 페이지에 대한 성능 비교
Table 3. Comparison of Page Recommendation's Performance before Clustering and after Clustering.

	클러스터링이전 (2032 state)	클러스터링이후 (619 state)
RunTime	15분32초	6초
Accuracy	48.9%	43.2%

표 4. 시간대별 상태전이행렬을 이용한 페이지 추천의 accuracy
Table 4. Accuracy of Page Recommendation with Each Time's State Transition Matrix.

Time	13시~16시	17시~20시	21시~0시
Accuracy	47.1%	45.4%	47.3%

시스템이 n-step이후에 사용자가 참조할 클러스터로써 c를 추천하였을 경우, test데이터의 로그들 중에서 페이지 a를 참조한 세션이 최종 목표 페이지로 클러스터 c 안에 있는 페이지를 참조할 확률을 accuracy로써 측정하였다 아래의 표 2는 본 실험에서 측정된 각 step별 accuracy를 측정한 결과이다.

위의 표 2를 살펴보면, 단순히 바로 다음 페이지를 볼 확률(1-step)만 적용한 페이지 추천의 accuracy 보다는 일정 step 특히, 5-step 이후에 참조할 페이지를 추천한 accuracy가 더 높은 것을 알 수 있다. 또한, 불필요하게 많은 step 이후에 참조할 페이지를 추천한 결과는 오히려 accuracy가 감소하는 것을 확인할 수 있다.

다음, 앞의 실험에서 사용한 200개의 페이지 중에서 accuracy가 높은 상위 10%의 페이지들과 accuracy가 낮은 하위 10%의 페이지들을 대상으로, 원래의 상태전이행렬과 클러스터링을 하여 state의 개수를 감소시킨 상태전이행렬로 5-step이후의 페이지를 예측하였을 경우, 걸리는 시간 및 accuracy의 차이를 측정하였다. 표 3은 Pentium4-2Ghz CPU를 탑재한 시스템에서 측정한 실험 결과를 보여준다.

측정결과, 하나의 페이지를 추천하는데 걸리는 시간은 클러스터링 이후 비약적으로 감소하였으며, Accuracy의 감소는 상대적으로 적었던 것을 알 수 있다.

다음은 Ⅲ장 2절 바 항에서 제안한 시간대별로 다른 상태전이행렬을 이용한 페이지 추천에 대한 실험 결과이다. 시간대를 주부들이 접근할 가능성이 큰 13시~16시, 중 고등학생이 접근할 가능성이 큰 17시~20시, 직장인이 접근할 가능성이 큰 21시~0시로 분류하였으며 각각의 시간대의 로그를 기반으로 3개의 상태전이행렬을 구축한 후 다음날 해당 시간대의 로그들을 테스트 데이터로 하여 시간대별 상태전이행렬들이 추천한 페이지의 accuracy를 측정하였다. 표 4는 그 실험 결과를 보여준다.

실험결과, 시간대를 구분하여 여러 개의 상태전이행렬을 유지하고 시간대별로 적절한 상태전이행렬을 활용하여 페이지를 추천하는 방식의 accuracy가 표 2에 나타나 있는 그렇지 않은 페이지 추천방식의 accuracy보다 5 ~ 7% 정도 더 우수한 것을 알 수 있었다. 이러한 방식을 확장하여 저장 공간이 허락하는 한도 내에서 한 달의 시간대별 혹은 한 해의 시간대 별로 상태전이행렬의 개수를 증가시켜 accuracy를 더욱 향상시키는 것이 가능할 것으로 기대된다.

Ⅳ. 결 론

본 논문에서는, 기존의 연관규칙 기반의 웹 페이지 추천의 몇 몇 한계점을 극복하기 위하여 DTMC에 의거한 확률적 모델을 적용하여 임의의 n번째 페이지 참조 후에 사용자가 방문할 가능성이 가장 큰 페이지들을 추천하는 형태의 적응형 웹 사이트 구축을 위한 방법론을 제시하였다. 또한, 표 2에서 보여주는 실험결과에서, 누구나 흔히 생각할 수 있는 페이지들끼리의 상호참조 확률을 이용해서 바로 다음 페이지를 추천하는 방법보다는 본 논문에서 제안한 DTMC 모델의 확률밀도 계산을 응용하여 일정 step이후에 예상되어지는 페이지를 추천하는 방식이 사용자가 방문할 최종페이지를 예측하는 데 있어서 더 향상된 결과를 가져온 것을 알 수 있었다. 다음, 상태전이행렬의 연산에 의한 오버헤드를 줄이기 위하여 PageGather 알고리즘을 사용하여 사이트를 구성하고 있는 웹 페이지간의 유사성이 큰 것들을

클러스터링(clustering)하였고 실제로 이를 통해 비약적인 실행시간 감소의 효과를 얻을 수 있었다. 마지막으로 웹 사이트는 시간대별로 다른 특성의 사용자가 접근하므로 접근패턴이 달라질 수밖에 없다는 문제점을 해결하기 위하여 시간대 별로 다른 상태전이행렬을 구축하여 페이지 추천의 정확도를 향상시켰다.

그러나, 링크구조가 단순하거나, 계층적 웹 구조상에서 다수의 terminal 노드에 해당하는 페이지들을 가지고 있는 사이트에서는 이러한 PageGather 알고리즘을 적용했을 경우, 클러스터의 크기가 작아져서, state의 수를 효과적으로 감소시키지 못하였다. 다른 기존의 클러스터링 기법을 사용할 경우, 다수의 페이지들을 클러스터링 하는데 있어서 적지 않은 오버헤드가 예상된다. 따라서, 다양한 구조의 웹 사이트에 적용 가능하면서도, 오버헤드가 적은 효과적인 클러스터링 기법의 개발은 본 논문이 제시하는 시스템을 광범위한 사이트에 적용하기 위한 우선적인 필요조건이 될 것이다.

참 고 문 헌

[1] Mike Perkowitz and Oren Etzioni, "Adaptive Web Sites: an AI Challenge", In Proc of the 15th International Joint Conference on Artificial Intelligence, pp. 16-21, 1997.

[2] R. Agrawal, R. Srikant, "Fast Algorithms for Mining Association Rules", Proc. of the 20th VLDB Conference, Santiago, Chile, Sept. 1994.

[3] Jaideep Srivastava, R. Cooley, M. Deshpande, P-T. Tan, "Web Usage Mining: Discovery and Applications of Usage Patterns from Web Data", SIGKDD Explorations, Jan. 2000.

[4] Jung Soo Park, Ming-Syan Chen, Philip. S. Yu, "An Effective Hash-based Algorithm for Mining Association Rules", Proc. of ACM SIGMOD Conference on Management of Data, San Jose, California, pp.175-186, May 1995.

[5] Hannu Toivonen, "Sampling Large Database for Association Rules", Proc. of the 22nd VLDB Conference, Mumbai(Bombay), India, 1996.

[6] D. W. Cheung, J. Han, V. Ng, C. Y. Wong, "Maintenance of Discovered Association Rules in Large Databases: An Incremental Updating Technique", Int'l Conference on Data Engineering, New Orleans, Louisiana, Feb. 1996.

[7] Sergey Brin, Rajeev Motwani, Jeffrey D. Ulman, Shalom Tsur. "Dynamic Itemset Counting and Implication Rules for Market Data.", Proc. of ACM SIGMOD Conference on Management of Data, 1997.

[8] M. S. Chen, J. S. Park, P. S. Yu, "Data Mining for Path Traveral Patterns in a Web Environment", Proc. of the 16th Inernational Conference on Distributed Computhing Systems, pp.385-392, 1996.

[9] Vidyadhar G. Kulkarni, Modeling and Analysis of Stochastic Systems, Chapman & Hall, London, UK 1995.

[10] Mike Perkowitz, Oren Etzioni, "Adaptive Web Sites: Automatically Synthesizing Web Pages", In Proc of the 15th national/10th conference on Artificial intelligence/Innovative applications of artificial intelligence, pp.727-732, 1998.

저 자 소 개

노 수 호(정회원)

2003년 광운대학교 컴퓨터과학과 학사 졸업

2005년 광운대학교 컴퓨터과학과 석사 졸업(공학석사)

2005년 삼성전자 정보통신총괄 IT연구소 연구원

<주관심분야 : Artificial Intelligence, Data/web Mining, Agent>

박 병 준(정회원)

1984년 서울대학교 컴퓨터공학과 학사 졸업(공학사).

1988년 University of Minnesota, Computer Science 석사 졸업(공학석사)

1997년 University of Illinois at Urbana-Champaign, Computer Science 박사 졸업(공학박사)

2000년~현재 광운대학교 컴퓨터소프트웨어학과 교수.

<주관심분야 : Artificial Intelligence, Data/web Mining, Knowledge-Based Systems>