Journal of Korea Multimedia Society Vol. 10, No. 9, September 2007(pp. 1106-1116)

스퀀스 연관규칙을 이용한 개인화 웹 마이닝 설계

윤종찬[†], 윤성대[††]

요 약

최근 들어 웹을 이용한 e-Commerce의 거래는 그 크기나 복잡도면에서 급속도로 확산되고 있다. 그러므로 웹 사이트의 설계나 웹 서버 설계 등이 복잡해지고 있다. 또한 웹 사용자가 많은 웹 이동경로를 이용하기 때문에 웹 사용자에 대한 데이터를 분석하는 일이 쉽지 않다. 기존 논문에서는 연관 규칙 탐사는 항목들간의 상관성을 찾아내는 것으로 기존의 연관 규칙 탐사 알고리즘들은 상관성이 높은 모든 항목들을 찾아낸다. 그러나 사용자들은 종종 자신이 관심있는 연관 규칙들만을 찾길 원한다. 하지만 기존의 알고리즘을 그대로 사용하여 찾아낸 모든 연관 규칙들 중에서 원하는 규칙들만을 찾아내는 것은 매우 비효율적이다. 본 논문에서는 웹 사용자의 이동경로의 사용자 패턴을 데이터마이닝 기법 중 하나인 연관규칙을 이용하여 사용자에게 맞는 이동경로를 구한 후 모든 경로를 이어주기 위해 시차 연관규칙을 이용하여 각 노드들을 이어주는 시스템을 제안한다. 제안한 시스템은 시차 연관규칙 기법을 통해 웹 사용자의 이동 경로를 사용자의 특성에 맞는 개인화 또는 고객 세분화된 사이트를 구축가능하게 제안한다.

Design of a Personalized Web Mining System Using a Sequence Association Rule

Jong-Chan Yun[†], Sung-Dae Youn[††]

ABSTRACT

Recently e-commerce trade on the web has grown rapidly in scale and complexity, just as web site designs and web servers have become more complicated. In view of these complexities, it is obviously difficult to analyze web user's data since they web users employ so many different web paths. The existing association rule investigation algorithms identify all items with a high correlation. However even though users often only want to find items in which they have interest, it is still difficult to find the rules they want out of all of the many association rules found by existing algorithms. In this paper, we propose a system linking each node with the sequence association rule, linking all routes after finding a path corresponding to a user with the association rule-one of the data mining techniques which identify user patterns in web user paths. The suggested system helps us construct individualized or customer-subdivided sites using the sequence association rule in order to harmonize the paths of web users with user characters.

Key words: Association Rule(연관규칙), Web Mining(웹 마이닝), Web-Personalization(웹 개인화)

1. 서 론

최근 인터넷환경의 급격한 발전으로 인터넷에서의 e-Commerce가 더욱 활발하게 진행되고 있으며 인터넷을 통해 웹 쇼핑을 하는 것뿐만 아니라, 다양한 서비스를 받을 수 있게 되었다. 그러나 인터넷의

※ 교신저자(Corresponding Author) : 윤성대, 주소 : 부산광역시 남구 대연3동 599-1(608-737), 전화 : 051)620-6398, FAX : 051)620-6450, E-mail : sdyoun@pknu.ac.kr
접수일 : 2007년 3월 26일, 완료일 : 2007년 9월 3일

[†] 정회원, 부경대학교 전자상거래시스템(협동)과정 박사수료
(E-mail : yjc313@hanmail.net)
[††] 정회원, 부경대학교 전자컴퓨터정보통신공학부
(E-mail : sdyoun@pknu.ac.kr)

방대한 정보로 인해 정보 홍수에 빠진 사람들은 정보에 대해 까다로운 요구를 하고 있다. 이를 테면 개인화 또는 맞춤화된 정보를 제공 받기를 원하고 있다[1,2].

인터넷환경에서 목표를 갖는 마케팅 전략을 세우기 위해서는 사용자 개개인의 취향, 접근패턴에 관한 정보가 필요하다. 이와 같은 정보를 기반으로 사용자 개개인의 특성에 맞는 동적인 웹 페이지구성이나 링크 정보를 제공할 수 있다. 특히, 사용자의 접근패턴을 알 수 있으면 웹 공간구성, 제품들 간의 상관관계를 고려한 마케팅 전략수립 및 사용자가 원하는 부가적인 정보를 미리 제공하여 효과적인 사용자 관리를 할 수 있다[3]. 효과적인 사용자 관리를 해결하기 위해 데이터 마이닝을 사용한다. 데이터 마이닝은 대규모 데이터로부터 조직이 필요로 하는 정보를 추출해 내기 위한 다양한 분석기법을 선정, 모형을 구축, 평가, 결과를 적용하는 일련의 과정이다.

본 논문에서는 데이터마이닝기법 중 연관 규칙 기법을 적용하여 웹 공간 구성이나 웹 이동경로 간의 링크 정보의 패턴을 분석하여 웹 사용자에게 필요한 정보를 주고자 한다. 연관 규칙은 항목들 간의 상관성을 찾아내는 것으로 기존의 연관 규칙 알고리즘들은 상관성이 높은 모든 항목들을 찾아낸다. 그러나 사용자들은 종종 자신이 관심 있는 연관 규칙들만을 찾길 원한다. 하지만 기존의 알고리즘을 그대로 사용하여 찾아낸 모든 연관 규칙들 중에서 원하는 규칙들만을 찾아내는 것은 매우 비효율적이다. 따라서 사용자들의 요구에 맞도록 항목 제한 조건을 주고 그에 맞는 방법을 사용한다면 보다 효율적으로 관심 있는 연관 규칙들만을 찾아낼 수 있다. 또한, 최근 몇 년 동안 폭발적으로 발전한 WWW은 접근 가능한 온라인상에서의 웹 데이터의 거대한 자원이 되었고, 특히 웹 서버의 로그 파일을 이용한 분석이 많이 이용되고 있다. 하지만 기존의 웹 이동경로에서는 사용자가 자주 찾는 웹 페이지를 찾아주는 것에만 관심을 두고 그 웹 페이지에 대해 사용자가 어느 정도의 관심(머무는 시간 정도)을 주는 지에 대해서는 고려하지 않았다. 시차(sequence)를 고려하여 머무는 정도에 따라 가중치를 부여함으로써 사용자가 원하는 정보가 담긴 패턴을 우선적으로 찾는 것이 본 연구의 목적이다. 사용자 개개인의 취향과 접근 패턴에 관한 정보를 기반으로 사용자 개개인의 특성에 맞는 동적인 웹 페이지 구성이나 링크패턴 정보를 제공할 수 있는 개인화 웹 마이닝이 필요한 것이다. 이러한 결과를 가지고 웹 마케팅(1:1마케팅)도 가능하고 효과적인 eCRM이 가능해 지는 것이다.

본 논문에서는 대량의 빈발항목은 아니지만, 실제로 의미 있는 유용한 항목에 대한 연관 규칙을 탐사할 수 있는 방법을 제안한다. 제안하는 방법은 데이터들 간의 상대적인 발생 빈도를 고려하는 척도인 최소 지지도를 이용하여 각 경로 계층이 존재하는 데이터들에 대하여 상대적으로 최소 지지도(minSup)보다 작은 경로는 제거시킨다. 또한, 각 경로를 순환하기 위해서 빈발항목 중에서 시차 연관 규칙을 이용하여 지지도가 가장 높은(maxSup) 경로간을 연결해서 각 개인에게 맞는 최단의 웹 경로를 순환하는 시스템을 설계하고자 한다.

본 논문은 2장에서 관련연구, 3장에서는 시스템 설계와 4장에서는 결론 및 향후연구과제에 대해서 서술한다.

2. 관련연구

2.1 연관 규칙(Association Rule)

연관 규칙이란 동시에 발생하는 사건들을 규칙의 형태로 표현한 것으로 특정 사건이 발생하면 동시에 혹은 일정한 시간 간격 사이에 다른 사건이 일어나는 관계를 의미한다.

데이터베이스에서 알려져 있지 않은 숨겨진 패턴을 탐사하는 연구 중에 연관 규칙에 대해 가장 많은 연구가 이루어지고 있다. 연관 규칙은 한 항목 그룹과 다른 항목 그룹 사이에 존재하는 강한 연관성을 찾아내어 그룹화 하는 클러스터링의 일종이다. 대규모 비즈니스 트랜잭션 데이터들 사이에서 흥미 있는 연관 관계의 발견은 카탈로그 디자인, 교차(cross) 마케팅, 손실 원인 분석 등의 비즈니스 의사결정 프로세스에 많은 도움이 된다. 또한, 동시에 구매될 가능성이 큰 상품들을 찾아냄으로써 시장바구니 분석(Market Basket Analysis)에서 다루는 문제들에 적용할 수 있다.

연관 규칙 발견 알고리즘으로는 Apriori, OCD, SETM, DHP 알고리즘 등이 있다[4]. 이 알고리즘 중에 Apriori방법은 이진 연관 규칙에 대한 빈발 항목집합을 찾아내는데 유용한 알고리즘이다. Apriori

는 k번째의 항목집합이 k+1번째 항목집합을 발견하기 위해 사용되는 레벨단위로 진행되는 반복 접근법을 사용한다.

연관 규칙기법에 적용되는 데이터는 판매 시점에서 기록된 거래와 품목에 관한 정보를 담고 있다. 연관 규칙 탐사과정은 크게 두 단계로 진행된다. 첫 번째는 높은 지지도(Support)를 갖는 아이템의 집합을 식별하는 작업이고, 두 번째 단계는 높은 신뢰도(Confidence)를 갖는 연관 규칙을 도출하는 작업이다. 여기서 지지도와 신뢰도의 개념은 아주 중요한 개념으로 빈발 항목 집합을 찾아내는데 있어 큰 역할을 한다[5].

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

$$\text{support}(A \Rightarrow B) = P(A \cap B) = \frac{\text{품목 } A\text{와 } B\text{를 동시에 포함하는 거래수}}{\text{전체 거래수}} \quad (1)$$

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

$$\text{confidence}(A \Rightarrow B) = P(B|A) = \frac{\text{품목 } A\text{와 } B\text{를 동시에 포함하는 거래수}}{\text{품목 } A\text{를 포함하는 거래수}} \quad (2)$$

2.2 웹 마이닝(Web-Mining)

인터넷 사이트는 기존의 물리적인 시장보다 마케팅 규칙을 생성하기 위한 자료를 쉽게 얻을 수 있으므로 일대일 마케팅 활동을 적은 비용과 노력으로 할 수 있다. 따라서 인터넷의 활용과 전자상거래의 확산에 따라 더 많은 데이터의 기록이 가능해졌고, 웹 사이트의 효과적인 관리에 관심을 갖게 되었으며 그 중 웹 로그 데이터에 대한 분석이 가장 먼저 관심의 대상이 되었다. 그러나 CRM(Customer Relationship Management)을 위해 접속자 개개인에게 맞는 서비스를 제공하기 위해서는 접속자의 특성과 탐색 특성을 알아야 하지만 일반적인 로그 분석 툴에 의한 결과는 한계점을 지니고 있기에 데이터 마이닝 방법론의 웹 데이터에의 적용이 연구되고 있다[2,6-8].

웹 마이닝이란 웹상에 존재하는 대량의 데이터 속에서 의미 있고 유용한 정보를 발견하고 추출하는 일련의 프로세스라고 정의할 수 있다. 웹 마이닝에서 이용되는 데이터는 일반적인 웹 로그 분석에서 사용되는 로그 데이터뿐만 아니라 고객에 관한 정보 및 웹 사이트의 컨텐츠와 관련한 웹 데이터를 함께 포함하고 있다[9].

웹 마이닝은 분석의 원천이 웹이라는 점에서 일반적인 데이터 마이닝과 구별되며, 이와 같은 특성은 여러 가지 의미를 내포하고 있다. 우선 웹은 Hyper Link된 웹 문서의 도식화 형태로 구성되고, 웹 문서는 구조화되거나 그렇지 않은 자료를 포함한다. 따라서 웹 로그 파일이라는 독특한 데이터의 특성상 정제 과정에서 특별한 주의를 요하게 된다[2,10].

데이터 마이닝을 통해서 기업은 웹 사이트상의 패턴을 의미 있는 정보로 분석하고 인터넷상의 고객들의 예상치를 이해하고 연관시킬 수 있게 된다. 데이터와 웹이 제공하는 방대한 사업지식의 흐름에 근거한 웹 마이닝은 웹 이용자와의 관계를 생성하고 유지시키며 생산성 있는 웹 사이트를 구축하는 데 있어 결정적 열쇠가 되는 것이다.

R. Kosala와 H. Blockeel은 웹 마이닝을 웹 컨텐츠 마이닝, 웹 구조 마이닝, 웹 이용 마이닝의 세 가지로 구분하였다. 이러한 웹 마이닝의 분류를 도식화 한 것이 그림 1이다[2].

웹 컨텐츠 마이닝은 웹 사이트와 관련한 자료, 문서 등으로부터 유용한 정보를 추출하는 것이고, 웹 구조 마이닝은 웹 페이지 내의 하이퍼링크 구조에 초점을 두고 웹에 링크된 하위 구조의 모형을 발견하고자 하는 것으로 유사한 웹 페이지를 군집화하거나 서로 다른 사이트 간의 관계에 관한 정보를 얻는데 유용하다. 또한 웹 이용 마이닝은 웹 서버에 저장된 웹 로그 데이터를 통해 웹 서버 접속자의 접속패턴을 발견하고 분석하는 프로세스를 말한다. 이러한 웹 이용 마이닝에 사용되는 데이터는 웹 서버 로그 파일,

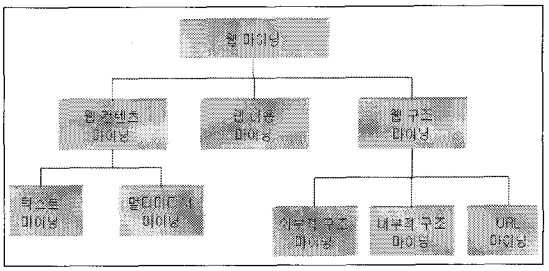


그림 1. 웹 마이닝의 분류

프락시 서버 로그 파일 등 웹 로그 파일 뿐만 아니라 접속자의 프로파일, 로그인 데이터, 사용자의 세션이나 거래(Transaction), 쿠키, 사용자의 질의, 북마크 데이터, 마우스 클릭이나 스크롤 등 사용자와 웹 사이트 사이에 발생하는 모든 데이터를 사용한다. 웹 서버의 로그 데이터를 분석하는 웹 로그 분석도 웹 이용 마이닝의 한 분야라고 할 수 있다[11].

2.3 웹 개인화(Web-Personalization)

Kravatz(1999)에 의하면 개인화는 초기 웹 페이지의 내용과 화면구성을 웹 이용자에게 맞도록 맞춤 제작하는 것으로 인터넷 경험과 검색 능력이 떨어지는 이용자라도 개인화 된 웹 페이지에 오래 머물게 하여 일대 일 마케팅 기회로서 활용하게 되는 것이다[6]. 웹 사이트에서 일어나는 모든 작업은 정보시스템을 이용하여 수행된다. 따라서 웹 사이트를 방문하는 모든 고객에 대한 정보를 수집하여 체계화하기가 현실 세계에 비해 용이하다. 또한 웹 사이트의 모든 방문자에게 개별적으로 가장 적합한 서비스를 전개하는 데 상대적으로 적은 비용이 소요된다. 인터넷 비즈니스가 이러한 기회 요인을 충분히 활용하려는 노력의 일환으로 외국의 많은 사이트들은 개인화된 서비스를 제공하고 있다[12,13].

웹 개인화는 두 단계로 나눌 수 있다. 첫 번째 단계는 사전처리와 데이터 준비 단계이며 여기서는 데이터 클리닝, 사용자 구체화, 세션 구체화가 포함된다. 두 번째 단계는 마이닝 단계이며 여기서는 사용 패턴이 연관규칙이나 클러스터링과 같은 방법을 통해 발견된다.

그림 2는 방문 빈도와 방문 시간만으로 고객들을 분류할 때 사용하는 매트릭스이다. 상위 20%인 고객들만을 Web Analyzer의 ID매칭을 이용하여 선별, 집중 타켓 마케팅을 펼침으로써 고객들의 충성도와 웹에 머문 정도로 웹 마케팅에서 사용하는 방법 중의 하나이다[13].

처음 인터넷이 나왔을 때는 많은 정보를 찾아주고 관리하는 서비스였으나 정보의 양이 많아지자 대중적 가치가 높은 것을 찾아주고 관리해 주는 서비스가 각광을 받기 시작했다. 그러나 결국은 개개인이 원하는 것을 찾아주고 관리해 주는 서비스로 나갈 수밖에 없다. 요즘 강화하고 있는 개인화 서비스가 이런 추세를 반영하고 있다.

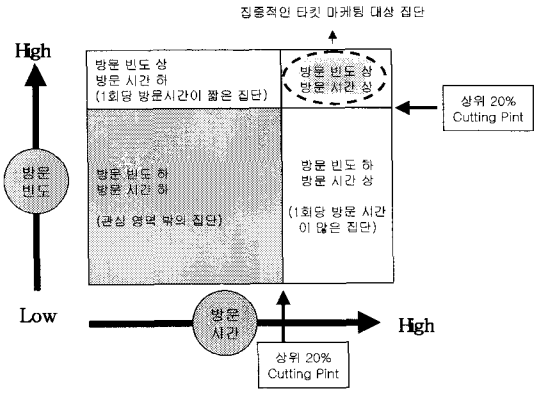


그림 2. 방문 빈도와 방문 시간 매트릭스

현재 개인화 서비스는 다양한 분야에서 이루어지고 있으며, 특징은 차세대 웹인 시맨틱 웹 기술이 보급되면서 더욱 편하고 강력한 개인화 기능이 나타나고 있는 것이다[6,14]. 이러한 개인화 서비스는 최근에 등장한 이야기가 아니고 웹 초창기부터 언제나 가장 큰 목표점의 하나였다. 그런데 요즘 들어 새삼스럽게 '개인화'라는 말이 다시 떠오르는 이유는 웹기술의 발달에 따라 초창기 웹에서 구현하지 못했던 개인화 서비스가 좀더 구체적으로 구현되기 시작했기 때문이다. 초창기인 1세대 웹은 대부분 사람의 손으로 이루어지는 수작업 시대라 할 수 있다. 반면, 요즘 떠오르는 2세대 웹인 시맨틱 웹(또는 웹2.0)은 대부분의 일을 컴퓨터끼리 처리하는 자동화 시대를 열고 있다. 이에 따라 이전에는 개인의 노동력에 의해 이루어지던 개인화 서비스의 상당 부분이 프로그램끼리 알아서 자동으로 처리하는 형태로 바뀌고 있다[14,15].

예를 들면, RSS(Really Simple Syndication)구독기를 들 수 있다. 예전에는 일일이 방문해 홈페이지에 새로 올라온 글을 확인하느라 많은 시간을 보냈지만, 이제는 RSS구독기가 알아서 수 백 군데의 홈페이지를 방문하고 새로 올라온 글만 모아서 보여준다. 시맨틱웹의 기술 중 하나인 RSS를 통해 사용자 개인이 원하는 사이트의 새 글만 모아서 볼 수 있는 개인화가 가능해 진 것이다.

3. 시스템 설계

3.1 시스템의 자료구조

웹 로그 마이닝이란 웹 로그 파일을 분석하여 사

용자의 문서 방문 패턴을 찾아내어 기존의 웹 문서 구조를 개선하는 데 적용하는 기법이다. 웹 로그 파일에는 IP Address, 방문시간, 방문한 페이지, 방문 순서 등에 대한 정보가 기록된다. 이를 분석하면 사용자가 관심을 두고 있는 웹 페이지가 어떤 것인지 알 수 있고 이를 이용하여 웹 문서의 구조를 개선할 수 있다.

표 1에서 T_ID는 트랜잭션의 횟수를 의미하고, A, B, C, D, E, F는 웹 이동경로의 node(*.html)이다. 괄호 안의 첫 번째 숫자는 방문한 페이지에서 단위(머문 시간)이고, 두 번째 괄호안의 숫자는 웹 이동경로 순위이다.

표 2의 단위는 웹 페이지에서 머문 시간을 나타낸다. 예를 들어, 0은 최소 시간 즉, 잠시 지나가는 웹 페이지를 말하고 1은 20초에서 3분 정도를 머문 것을 나타낸다.

웹 사용 패턴 마이닝을 위한 절차는 그림 3과 같이 3단계로 이루어진다.

표 1. 접근된 로그 itemsets

T_ID	이동 itemsets					
	A	B	C	D	E	F
1	1(1)	1(2)	0	2(3)	0	0
2	1(1)1(4)	2(2)1(5)	2(3)	0	2(6)	0
3	0	2(1)	0	0	1(2)	2(3)
4	1(1)	1(4)	1(2)	0	3(5)	1(3)
5	0	1(3)	1(1)	0	3(4)	1(2)

표 2. 단위당 시간 가중치

단위(머문 시간)	가중치(W)
0(0≤t≤19초)	0
1(20초≤t≤3분)	0.2
2(3분<t≤5분)	0.3
3(5분<t≤10분)	0.5

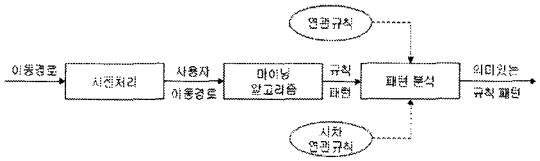


그림 3. 웹 사용 마이닝 과정

웹 마이닝을 위한 첫 번째 단계인 그림 3에서 이동경로의 데이터의 사전처리는 그림 4에서 보는 것과 같이 그 데이터를 정제한다. 이 단계에서는 사용자의 이동경로를 적절한 정제, 즉 경로에 머문 시간에 적당한 가중치를 부여하여 정제한다. 또한 연관규칙을 이용하여 최소 지지도를 구하여 최소 지지도보다 낮은 이동경로의 노드를 제거하여 사용자 이동 패턴을 식별한다. 그림 4의 이동경로 뷰에서는 각 minSup(최소 지지도)를 가지고 식별된 이동경로 패턴을 보고, 모든 이동경로를 순환하기 위해서 시차 연관규칙을 이용하여 노드간의 최고 지지도(maxSup)를 구하여 최종 이동경로를 완성한다.

그림 3의 두 번째 단계인 마이닝 알고리즘 단계는 사용자들의 이동경로 규칙이나 패턴을 찾아내기 위하여 웹 마이닝 알고리즘을 적용하는 단계이다. 마지막 단계인 패턴 분석은 연관성 규칙기법, 서로 관련성이 높은 페이지들을 발견하는 데 사용된다. 웹 사이트 설계를 할 때, 연관성 규칙 결과를 이용하여 효과적으로 사이트 구조를 설계할 수 있다. 시차 연관규칙은 하나의 노드에서 순차적으로 처리되는 페이지들의 특징을 밝히는 데 사용된다. 본 논문에서는 웹 구조를 최소 한 번 순환하기 위한 경로 설정을 위해 시차 연관규칙을 이용하여 최대 지지도를 구하여 각 노드를 연결하도록 한다.

그림 5는 각 빈발항목에서 최소 지지도(minSup)를 구해서 최소 지지도보다 낮은 항목을 제거하는 과정이다.

표 3은 기존 연관규칙을 이용한 minSup(최소 지지도)과 시차 연관규칙을 이용한 maxSup(최대 지지도)를 구하는 알고리즘에서 사용하는 항목에 대한 설명이다.

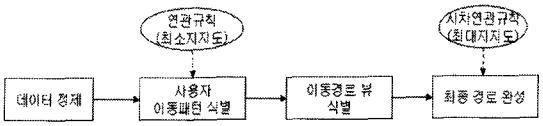


그림 4. 이동 경로 데이터 사전처리 과정

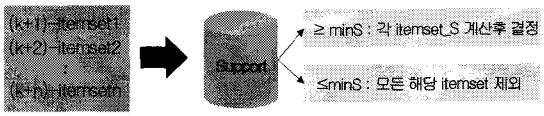


그림 5. 빈발 항목에 대한 최소 지지도(minSup)

표 3. 그림 7에서 사용하는 항목에 대한 설명

- L_k (k=1,2,3, ····, m) : k 단계에서 발견된 모든 빈발 항목 집합들의 집합
- C_k (k=1,2,3, ····, m) : L_k를 위한 후보 빈발 항목집합들의 집합
- D : 거래들이 들어있는 DB
- C_1을 생성 각 항목 하나하나를 원소로 하는 후보 빈발 항목집합들을 구성
- C_1에 속한 각 후보 빈발 항목집합. 즉, 각 원소 하나하나 단위로 지지도를 계산
- L_k에서 웹 페이지가 모두 선택되기 위해서 빈발항목이 2개인 것 중에서 지지도 계산
- 빈발 2항목인 것 중에서 스퀀스하게 연결된 것을 구함
- 지지도(시차 연관규칙의 지지도를 이용 : SL_k)가 가장 큰 웹 페이지 노드를 연결

그림 6은 연관규칙에서 빈발항목을 마이닝하기 위한 단계이다. 이 단계에서는 예외처리단계, 각 결과에 대한 데이터 조인단계 그리고 마지막 단계로 필요 없는 항목 즉, 최소 지지도보다 낮은 지지도에 해당하는 항목들을 삭제하는 단계로 수행되어진다.

그림 7의 알고리즘들을 이용해서 표 1의 트랜잭션에서 발생한 웹 이동경로를 개인에게 맞는 웹 개인화를 작성하면 다음과 같이 나타나는 것을 볼 수 있다. 표 1에 대한 전체 트랜잭션은 5개이고, 6개의 itemsets이 데이터베이스 D에 있다.

데이터베이스 D에 있는 웹 로그 데이터에 연관규칙 Apriori 알고리즘을 적용한 결과는 다음과 같다. 빈발 1항목의 C_1 : [A], [B], [C], [E], [F]들의 itemsets의 계산은 3/5(0.6), 5/5(1), 3/5(0.6), 4/5(0.8), 3/5(0.6)의 각각 지지도가 나온다. 괄호안의 수치가 각 itemsets의 지지도이며, 최소 지지도(minSup : minimum_Support)는 0.4(minSup = 0.4)이다. itemset1에서 node [D]는 트랜잭션 5개 중에서 한 번만 이동하였기 때문에 빈발 항목의 최소 지지도를 구하는 데 적합하지 않기 때문에 제거하기로 한다. 따라서, L_1 : [A], [B], [C], [E], [F]항목들만 획득하게 된다. L_1을 가지고 C_2의 빈발 2항목에 대한 지지도를 구하면 C_2 : [A][B](3/5=0.6), [A][C](2/5=0.4), [A][E](2/5=0.4), [B][E](4/5=0.8), [C][B](3/5=0.6), [F][B](2/5=0.4), [C][E](3/5=0.6), [C][F](2/5=0.4), [F][E](3/5=0.6)의 결과가 나온다. minSup(0.4)를 제거한 L_2의 결과는 L_2 : [A][B], [B][E], [C][B], [C][E]이다. 빈발 2항목에서 제거되는 항목은 한 번 발생하거나 빈발 1항목에서 제거된 [D]를 포함한 항목은 제거하도록 한다. 다시 L_2를 가지고 C_3의 빈발 3항목에 대한 지지도를 구하면 C_3 : [A][B][E] (2/5=0.4), [A][C][E](2/5=0.4), [C][B][E] (3/5=0.6) [C][F][B](2/5=0.4), [C][F][E](2/5=0.4), [F][B][E] (2/5=0.4)의 결과가 나온다. minSup(0.4)를 제거한 L_3의 결과는 L_3 : [C][B][E]이다. 더 이상 itemsets이 없기 때문에 연관규칙의 지지도는 구할 필요가 없다.

식 3과 식 4는 기존 연관규칙을 이용한 minSup과 시차 연관규칙을 이용한 maxSup을 구하는 알고리즘에서 각 빈발항목에 대한 가중치들의 합과 각 빈발

1) exception_process step(예외처리 단계)
첫 번째 최소 지지도를 구하는 단계에서 itemset1에서 한 번만 이동된 node는 최소 지지도에서 제외시킨다.

2) join step
(표 1의 데이터 처리 및 결과 데이터 조인 단계)
빈발 항목 집합 L_k를 찾기 위해 후보 C_k의 집합은 L_{k-1}과 L_{k-1}의 조인으로 생성된다.

3) Prune step(데이터 삭제 단계)
후보 C_k의 k-1항목 부분집합이 L_{k-1}에 속하지 않을 때 이를 모두 제거한다. L_k는 C_k에서 최소 지지도를 만족하지 못하는 항목들을 제거하여 L_k를 생성한다.

그림 6. 연관규칙을 위한 빈발항목 마이닝 단계

```
Procedure Sequence Association Rule algorithm
Input : frequent 1-itemsets, W1, Wi
Output : Lk, Answer, X
  L1={frequent 1-itemsets };
  for (k = 2 until Lk-1 ≠ 0 step 1) do begin
     Ck = apriori_gen(Lk-1);        // new candidate
     forall records r ∈ D do begin
           Cr = subset(Ck, r);      // candidates in r
           forall candidates c ∈ Cr  do begin
                 c.count = c,count + 1;
           end
      end
      Lk = { c ∈ Ck | c,count ≥ minsup };
  end
  Answer = UkLk
  for (i=1; i<n; i++) do begin
  W1=0
  Wi = i번째 가중치 값
  S=itemset/T_ID
  W1=W1+Wi
  X=W1 × S
  end
```

그림 7. 기존 연관규칙을 이용한 minSup과 시차 연관규칙을 이용한 maxSup구하는 알고리즘

항목들의 지지도를 구하는 식이다.

식 3에 의해 구해진 각 노드간의 가중치의 결과는 표 4와 같다. 식 4는 각 빈발항목의 지지도를 구하는 식이다. 식 3을 이용한 [C][E]간의 가중치[(0.3+0.3+0.2+0.5+0.2+0.5)×3/5 = 1.2]는 페이지 [C][B]의 가중치[(0.3+0.2+0.2+0.2+0.2+0.2)×3/5 = 0.78]보다 높다는 것이다. 이 결과로 빈발 1항목의 결과에서 보듯이 페이지 [E]의 자체 가중치가 높게 나오는 결과가 된다. 웹 구조 개선에서 볼 때 [E]를 포함한 웹 페이지가 가장 관심 있는 빈발항목이 된다.

그림 8은 그림 7의 기존 연관규칙을 이용한 minSup(최소 지지도)과 시차 연관 규칙을 이용한 maxSup(최대 지지도)을 구하는 알고리즘에 대한 후보 항목집합과 빈발 항목집합을 생성하는 흐름도이다.

$$\sum_{i=1}^{n} Wi \times S \qquad (3)$$

Wi : 각 빈발항목의 가중치의 합
S : 각 빈발항목의 지지도

$$S = \text{Support}(A \rightarrow B) = \frac{Node\,A\text{와 } Node\,B\text{를 포함한 이동 경로수}}{\text{전체 거래수}(T\_ID)} \qquad (4)$$

이 결과로 표 1의 기본 웹 구조가 그림 9처럼 변형된다. 웹 구조는 계층적 구조를 가지고 있기 때문에 [A][C][F] 페이지에서 사용자가 가장 관심 있는 페이지(높은 지지도가 나온 페이지) [E] 페이지로 바로 갈 수 있는 구조로 개선된다.

웹 사이트를 사용자의 특성에 맞는 개인화 또는 고객 세분화된 사이트로 구축 가능하다.

표 4. 그림 7에 대한 항목 집합 결과

빈발항목 집 합	Apriori알고리즘(MinSup=0.4)
빈발 1항목	[A](3/5=0.6), [B](5/5=1), [C](3/5=0.6), [E](4/5=0.8), [F](3/5=0.6)
빈발 2항목	[A][B](3/5=0.6), [A][C](2/5=0.4), [A][E](2/5=0.4), [B][E](4/5=0.8), [C][B](3/5=0.6), [F][B](2/5=0.4), [C][E](3/5=0.6), [C][F](2/5=0.4), [F][E](3/5=0.6)
빈발 3항목	[A][B][E](2/5=0.4), [A][C][E](2/5=0.4), [C][B][E](3/5=0.6) [C][F][B](2/5=0.4), [C][F][E](2/5=0.4), [F][B][E](2/5=0.4)
빈발 4항목	없음

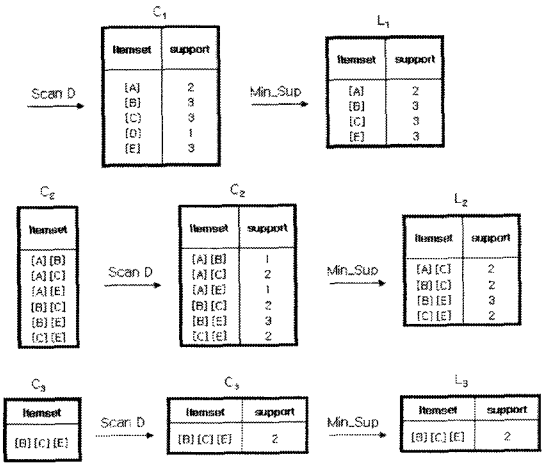


그림 8. 그림 7에 대한 후보 항목집합과 빈발 항목집합 생성도

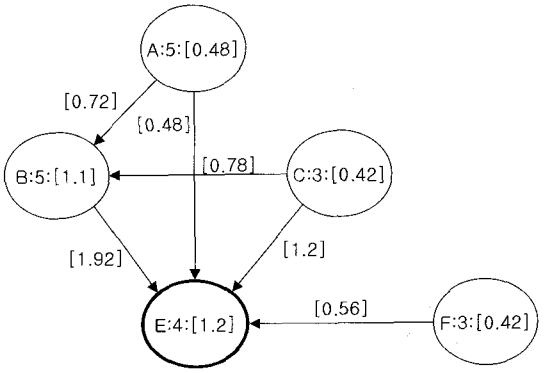


그림 9. 가중치에 의해 발견된 웹 구조

그림 10은 itemsets이 최소 지지도 이상인 항목집합에서 maxSup(최대 지지도)을 가진 itemsets을 발견하는 알고리즘이다. 그림 11은 그림 10에 대한 블록도를 나타낸 것이다.

```
Procedure frequent Path algorithm
Input : i-itemsets
Output : result
 SL : Sequence Support
 SLi = {large i-itemsets};
 for (k=2;SLk-1≠∅;k++) do begin
        Ck=apriori_gen(SLk-1);
                //new candidate itemset Create
        get_count(Ck)
        SLk={c∈Ck | c.count ≥maxSup}
 end
 result =∪ SLk
```

그림 10. 시차 연관규칙을 이용한 maxSup발견하는 알고리즘

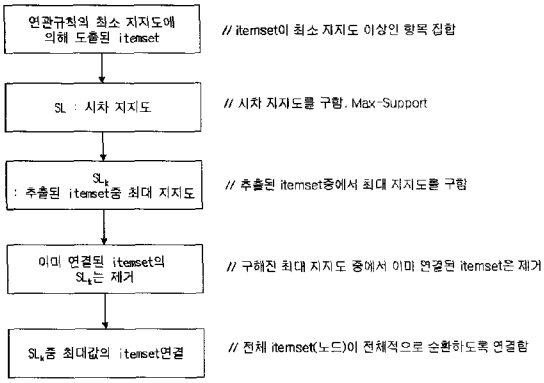


그림 11. 그림 10에 대한 블록도

각 노드간의 시차 연관규칙에 의해 구한 지지도는 다음과 같다. [A][B]=(2/3)=67%, [A][C]=(1/2)=50%, [A][E]=0%, [B][E]=(4/4)=100%, [C][B]=0%, [F][B]=(2/2)=100%, [C][E]=0%, [C][F]=(2/2)=100%, [F][E]=0%로 구해진다. 따라서 최대 지지도로 구해진 노드는 [B][E], [F][B] 그리고 [C][F] 중에서 이미 연결된 [B][E], [F][B]는 제거된다. 연결되지 않은 노드 중 시차 연관 규칙의 지지도에 의해 구해진 결과(최대 지지도)에서 [A]와 [C]를 연결하는 것보다는(시차 연관 규칙의 지지도는 50%가 나옴) [C]와 [F]를 연결하는 것이 시차 연관 규칙의 지지도를 구한 결과를 봤을 때 전체 노드에 대한 각 노드에 대한 이동경로 순환과정에 대한 노드 연결이 효과적이다.

표 5는 이미 구해진 빈발 2항목 중에서 시차 연관규칙을 이용한 maxSup을 구한 결과 중 이미 연결된 빈발 2항목은 제외시키고 연결되지 않은 노드를 최대 지지도를 가진 [C][F]를 연결하는 itemsets의 결과를 나타내고 있다.

표 5. 그림 11을 이용한 maxSup을 가진 itemsets 결과

항목 2의 집합	시차연관규칙을 이용한 MaxSup 결과	비고
[A][B]	= 2/3 = 67%	
[A][C]	= 1/2 = 50%	
[A][E]	= 0%	
[B][E]	= 4/4 = 100%	MaxSup(이미 연결)
[C][B]	= 0%	
[F][B]	= 2/2 = 100%	MaxSup(이미 연결)
[C][E]	= 0%	
[C][F]	= 2/2 = 100%	MaxSup → 선택
[F][E]	= 0%	

그림 12는 빈발항목을 구하는 알고리즘과 시차연관규칙의 지지도를 이용하여 각 노드 간 지지도를 구하여 최소 이동 경로를 순환하도록 작성된 웹 구조이다.

표 6은 웹 이동경로를 기존의 일반 웹 이동경로, 연관규칙의 최소 지지도를 이용한 웹 이동 경로 그리고 기존의 연관규칙에 시차를 적용한 최대 지지도를 이용한 경로 지정을 비교한 결과이다. 표 5의 이동경로(T_ID : 1, 2, 3, 4, 5)의 각 경로를 보면 시차 연관규칙에서 웹 이동 경로가 순환되는 것을 볼 수 있다.

또한 표 6에서 보듯이 일반 연관 규칙보다도 시차를 적용한 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. 그리고 그림 13의 결과에서 보듯이 웹 사용자가 웹을 이용하는 데 걸리는 시간에서 시차 연관규칙이 일반경로나 일반 연관규칙보다 좀 더 짧은 것으로 나타났다. 그림 13은 세 가지 웹 경로 연결 상태에서 찾은 빈번한 항목 집합의 평균수를 보여준다. 시차연관규칙에서는 115개의 다른 빈번한 k항목 집합을 찾고 Apriori연관규칙에서는 120개를 찾으며 일반경로에서는 135개를 찾는다. 시차연관규칙과 Apriori연관규칙의 차이는 일반 연관규칙의 결과에 maxSup(최대 지지도)값을 적용하였기 때문이다. 그림 14는 표 6의 각 연결 상태에서 웹 경로 수를 나타낸 것이다. Apriori 연관규칙과 시차연관규칙에서는 큰 차이를 보이지는 않았다. 그러나 일반 웹 경로보다는 간편해진 것으로 보여진다.

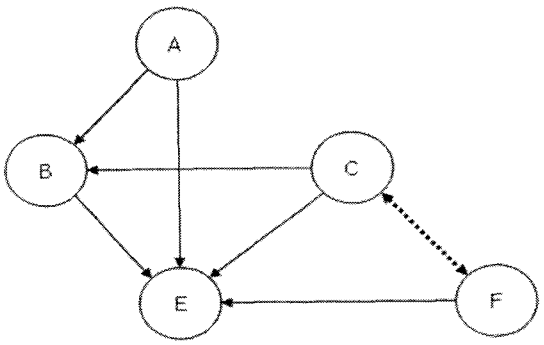


그림 12. 모든 노드를 연결한 웹 구조

표 6. 각 연결 상태 비교

	일반경로	연관규칙 (minSup)	시차 연관규칙 (maxSup)
이동 경로 (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→B→E→F	1. A→B 2. A→B→E 3. B→E→F 4. A→E 5. C→F→E
사용자 Click수	많음	보통	보통(짧음)
사용자 이용시간	많음	보통	보통(짧음)
웹 디자인 설계	복잡	간편	간편

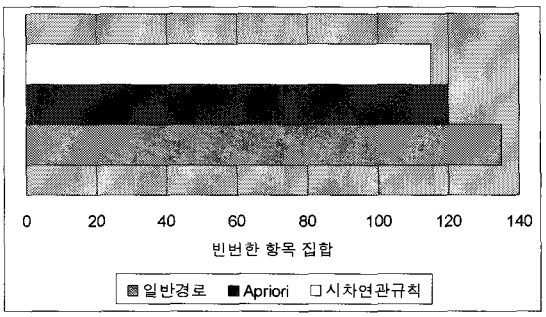


그림 13. 웹 경로 연결상태 별 빈번한 항목 집합 검색 결과

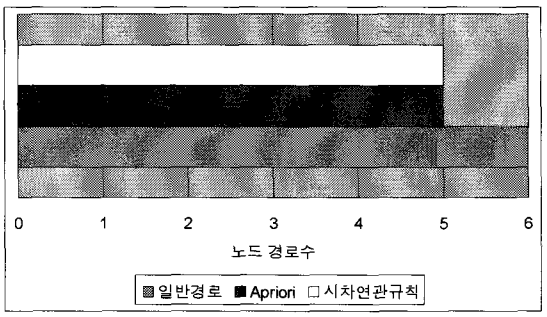


그림 14. 웹 경로 연결상태 별 노드 경로수

그림 15에서 보듯이 각 웹 경로에서의 클릭 수는 큰 변화가 없지만 제안한 시차 연관규칙에서는 Apriori연관규칙을 이용한 웹 경로보다는 짧아진 것을 볼 수가 있다. 트랜잭션의 마지막 5단계인 모든 웹 경로를 지원하는 단계에서는 다른 경로를 지정한 것보다 개인화에 맞게 웹 경로를 지정할 수가 있고 또한 전체 웹 경로 지정이 짧아진 것을 알 수가 있다. 따라서 고객의 사용 만족도가 시차연관규칙을 이용해 작성된 웹 경로가 순환하는 것이다.

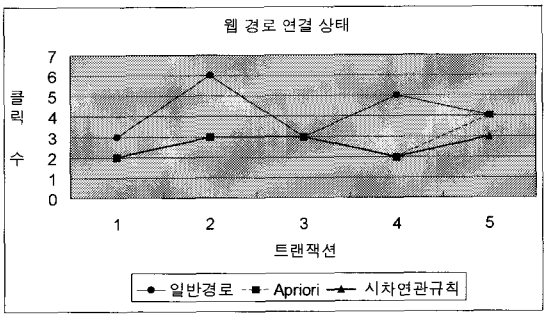


그림 15. 데이터셋에 따른 웹 경로에 따른 클릭 수

4. 결론 및 향후 연구과제

국내 웹 사용자가 급속도로 증가하고 쏟아지는 대량의 정보와 다양해지는 웹 사용자의 요구를 개인별로 처리할 수밖에 없는 실정이다. 웹 사용자들의 웹 사이트상의 패턴을 의미 있는 정보로 분석하고 인터넷상의 사용자들과 예상경로를 이해하고 연관 시킬 수 있게 되었다. 그리고 웹 데이터에 대한 분석은 웹 사이트 운영 전략을 수립하는데 있어서 가장 중요한 요소로 작용하는데, 웹 사이트를 방문한 사용자의 웹 이동경로와 이동패턴 등을 종합적으로 분석할 수 있는 웹 마이닝을 이용하여 웹 이용자의 이동경로를 순환시키고자 한다.

웹 구조를 개선한다는 것은 웹 문서간의 관계가 깊은 문서끼리 링크를 연결시켜주는 것이라 할 수 있다. 실험결과에 의하면 시간 탐색과 최대 지지도를 이용한 항목집합들의 연결이 일반 경로와 Apriori 알고리즘에 의한 최소 지지도에 의한 경로보다는 웹 경로 설정에 있어 모든 이동경로를 순환하는 것으로 보여진다.

본 논문에서는 웹 사용자의 웹 이동경로의 사용자 패턴을 데이터마이닝 기법 중 하나인 연관규칙을 이용하여 웹 사용자의 웹 이동 패턴을 지지도를 이용하여 구한다. 웹 사용자에 맞는 웹 이동경로를 구한 후 모든 경로를 순환하기 위해 시차 연관규칙을 이용하여 각 노드들을 이어주는 지지도를 구하여 연결되지 않은 노드들은 구한 지지도 중 최고 지지도 값을 가진 노드들을 연결해 주는 시스템을 제안하고자 한다.

실험결과 제안된 연관규칙 알고리즘은 웹 경로를 찾아내는데 있어서 기존의 알고리즘보다 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 효율적으로 적용할 수 있다.

제안한 시스템은 시차 연관규칙 기법을 통해 웹 사용자의 이동 경로를 사용자의 특성에 맞는 개인화 또는 고객 세분화된 사이트를 구축가능하게 해 준다. 따라서, 웹 이용자가 원하는 웹 경로를 설정 해 줄 수 있으며, 웹 이용자는 불필요한 클릭 수를 줄이고 원하는 웹 페이지를 빠르게 이동할 수가 있어 시간상의 절약도 가능하다. 시차 관계를 고려한 연관규칙은 웹 경로를 순환할 수 있고, 경로간 연관규칙 마이닝 시 높은 신뢰도를 보이므로 개인화 웹 경로지정시 활용도를 높일 수 있을 것으로 본다.

향후 연구 과제로는 많은 이동 경로를 가진 웹에서 웹 이용자의 이동 패턴을 분석하여 지식 습득 모델과 대화식 모델을 구현하여 좀 더 빠른 웹 이동경로를 개인화 시키고 좀 더 정확한 사용자 패턴 분석을 위해 다른 가중치들을 연결시켜 보도록 하는 것이다.

참 고 문 헌

[1] DE-XING WANG, XUE-GANG HU, and 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, pp. 134-135, 4-5 November 2002.

[2] 배희호, "Web mining을 이용한 개인화시스템 설계에 관한 연구," 경복논총 No.6, pp. 8-14, 2002.

[3] 김일, 박규석, "웹 컨텐츠 마이닝을 이용한 개인화 시스템의 설계 및 구현," THESES COLLECTION, Vol.21, No.1, pp. 6-7, 2003.

[4] DE-XING WANG, XUE-GANG HU, XIAO PING LIU, and HAO WANG, "Association Rules Mining on Concept Lattice using Domain Knowledge," *Proceedings of the First International Conference on Machine Learning and Cybernetics, Guangzhou*, pp. 2152-2154, 18-21 August 2005.

[5] 이태림, 구자용, 박헌진, 이긍희, 최대우 공저, 데이터마이닝, 한국방송통신대학교출판부, 서울, pp. 159-164, 2005.

[6] Miguel Gomes da Costa Junior and Zhiguo Gong, "Web Structure Mining : An Introduction," *IEEE International Conference on Information Acquisition*, pp. 591-593, 2005.

[7] C Oosthuizen, J Wesson, and C Cilliers, "Visual Web Mining of Organizational Web Sites," *Proceedings of the Infomation Visualization (IV'06)*, pp. 395-396, 2006.

[8] Nivedita Roy and Tapas Mahapaatra, "Web Mining : A Key Enabler in E-Business," *IEEE Proceedings of ICSSSM' 05'*, pp. 1122-1124, 2005.

[9] Yew-Kwong Woon, Wee-Keong Ng, Xiang Li, and Wen-Feng Lu, "Efficient Web Log Mining for Product Development," *Proceedings of the 2003 International Conference on Cyberworlds (CW'03)*, pp. 295-300, 2003.

[10] 채승경, 사용무, "데이터마이닝을 이용한 웹 데이터 분석," 한국데이터베이스학회 국제학술대회 논문지, Vol.4 No. 1, pp. 352-354, 2001.

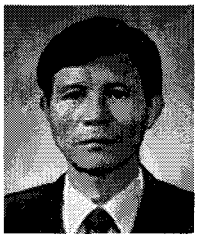
[11] 김양희, 조성의, "웹 사용 패턴 분석을 위한 데이터마이닝," Joural of Research Institute of Applied Science, Mokpo National University, Vol.1, pp. 266-268, 2001.

[12] 김삼근, 김병천, "웹 마이닝 시스템의 효율성 개선을 위한 확장된 로그 획득 시스템," 韓京大學校 論文集, 제33호, pp. 213-215, 2002.

[13] 김광용, 방명하, 강성범, 정수용 공저, "Web Analyzer을 이용한 로그 분석과 eCRM," 시대의 창(出), 서울, pp. 55-66, 2002.
[14] 김해룡, 이문규, "인터넷 개인화 서비스의 유형별 효과," 연세경영연구 제39권 제2호(통권 제75호), pp. 155-163, 2002.
[15] 정유진, 웹 2.0 기획론, 한빛미디어(出), 서울, 2006.

윤 종 찬

2003년 동명정보대학교 경영정보학과 졸업 (경영학사)
2005년 부경대학교 대학원 전산정보학과 졸업 (공학석사)
2006년 부경대학교 대학원 전자상거래시스템학과 박사과정수료
관심분야 : 전자상거래, 데이타마이닝, 유비쿼터스, e-CRM 등



윤 성 대

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