

# 웹 페이지 클러스터를 이용한 사용자 분류 기법<sup>†</sup>

백옥현<sup>0</sup> 서성보 이준욱 류근호  
충북대학교 전자계산학과  
{ohpaek, sbseo, junux, khryu}@dmlab.chungbuk.ac.kr

## User Classification Using Cluster of Web Pages

Ok Hyun Paek<sup>0</sup> Sung Bo Seo Jun Wook Lee Keun Ho Ryu  
Dept. of Computer Science, Chungbuk National University

### 요약

변화하는 경쟁적 인터넷 환경에서 E-Business의 성공적인 운영은 웹 사이트를 이용하는 고객들의 행위를 얼마나 잘 이해하느냐에 달려있다. 폭발적으로 늘어나는 웹 사이트 중에서 많은 사용자들을 유지하고 유지하기 위해서는 고객 개인의 특성을 분석해서 특성화된 서비스를 제공하는 것이 중요하다. 이 논문에서는 웹 서버에 의해 수집되는 로그 파일로부터 사용자들이 빈번하게 함께 접근하는 페이지들을 기반으로 웹 페이지에 대한 클러스터링을 수행하고 이러한 웹 페이지의 클러스터를 이용해서 유사한 행동패턴을 가진 사용자들을 분류함으로써 특성화된 서비스를 제공하기 위한 일련의 기법들을 제안한다. 특히 클러스터링을 수행하는 웹 로그에 시간적인 요소를 고려한 제약조건을 줌으로써 좀더 유용한 지식을 찾아낼 수 있는 방안을 제시한다.

### 1. 서론

인터넷을 통한 E-Business가 활성화됨에 따라 고객의 특성을 정확히 파악하고 고객 개인의 관심에 부합되는 개인화된 정보나 상품 서비스 등의 차별화된 서비스를 제공할 것이 요구되고 있다. 현재 고객의 행동 패턴을 분석, 예측하고 개인에게 특성화된 서비스를 제공하기 위해 웹 서버 로그파일을 분석하는 방법이 사용되고 있다. 웹 서버 로그에서 추출해 낼 수 있는 클러스터링은 사용자의 페이지 요청에 따른 데이터가 순서화되어 개인의 행동패턴을 분석하는데 유용하다[서성보2000].

개인화된 웹 서비스를 제공하는 것의 중요성을 인식하여, 이와 관련되어 웹 사용자를 클러스터링하기 위한 연구가 진행되어 왔다 [Moba99]. 그러나, 기존의 웹 로그를 이용한 웹 페이지 클러스터링에 관한 연구에서는 웹 로그에 저장되어 있는 시간 정보의 유용성을 고려하지 않았다. 전체 웹 로그파일에서 동일한 가중치를 주고 클러스터링 하는 것보다 특정 시간간의 로그파일 에 대해 클러스터링 하는 것이 더 의미가 있다. 일례로, 전자상거래 환경에서 추석, 설과 같은 특정 기간 동안의 웹 로그를 분석해서 사용자들을 클러스터링 하는 것은 특정 기간동안의 마케팅에 있어서는 더 높은 부가가치를 지닐 수 있다. 이런 점을 고려해서 이 논문에서 제안하고자 하는 것은 다음과 같다.

- 웹 로그 분석시 시간제약조건을 고려한 사용자 클러스터 분석
- 사이트에 접근하는 사용자들의 분류를 위해서 최소 발생빈도를 가지는 웹 페이지를 주요항목이라 보고, 주요항목에 기반한 페이지 클러스터를 이용해서 그 페이지의 클러스터에 자주 방문하는 사용자들을 분류하는 기법
- 생성된 사용자들의 클러스터에 대해서 속성지향 귀납법(Attribute oriented induction)을 사용한 각각의 클러스터에 속해 있는 사용자 특성 도출

사용자의 클러스터는 E-Business에서 시장 세분화, 타겟 마케팅, 추천 링크, 추천 상품 등과 같이 사용자 개인에 특성화된 웹 콘텐츠를 제공하는데 유용한 지식으로 쓰일 수 있다. 아울러 웹 페이지의 클러스터는 사용자가 식별되지 않은 상태에서도 동적으로 추천 서비스를 제공하는 데 유용하다.

### 2. 관련연구

클러스터링 기법은 모집단 또는 범주에 대한 사전 정보가 없는 경우에 관측값들 사이의 거리(또는 유사성)를 이용하여 전체를 몇 개의 그룹 또는 군집으로 나누는 분석법으로, 통계학, 기계학습, 정보검색 등의 분야에서 연구되어왔다. 대부분의 방법은 행렬, 코사인, Dice and Jaccard Coefficient 등과 같이 두 객체의 쌍 사이의 유사성을 측정하는 것이었다. 그러나, 이런 방법은 트랜잭션들이 아이템들을 공통적으로 갖고 있을 때는 적합하지 않다는 단점이 있다. [Wang99]에서는 아이템들이 조금씩 중복되어 있는 트랜잭션에 적합한 주요항목을 이용한 클러스터링을 수행한다. 이것은, 트랜잭션을 읽고 다음 트랜잭션에서 최상의 클러스터에 할당하는 K-means 알고리즘과 유사하지만, 초기 클러스터의 개수인 K 값을 요구하지 않으며, 거리에 토대를 두지 않고 클러스터의 전체적인 최상의 해결책을 제시한다는 것에서 차이가 있다. [Han97]는 트랜잭션 클러스터링을 위해 하이퍼그래프 분할을 이용하며, 클러스터링의 결과로 트랜잭션 내에서 함께 발생하는 아이템들의 집합을 산출한다.

속성지향귀납법[Han96]은 데이터의 속성에 기반해서 작업에 따라서 데이터를 일반화하는 기법으로 특성화 규칙, 차별화 규칙과 다계층 연관 규칙 등을 발견하는데 사용된다. 찾은 규칙의 종류, 규칙을 찾고자 하는 데이터의 부분집합, 데이터를 일반화하기 위한 개념 계층과 데이터를 어느 정도까지 일반화할 것인지를 정하는 속성 임계치를 입력으로 받아들이 규칙을 생성한다.

### 3. 웹 사용자 클러스터 분석 모형과 처리 과정

웹 로그를 분석하여 웹 페이지를 클러스터링 하고 그것을 이용하여 사용자의 클러스터를 생성하고 분석하는 과정은 그림 1 과 같다. 전처리 과정에서는 필요없는 항목들을 제거하는 데이터 정제과정과, 사용자들 식별하고, 각 사용자의 트랜잭션을 정의하게 된다. 전처리 과정을 거치면 사용자의 트랜잭션 파일이 생성되는데, 여기서 주요항목의 개념을 이용해서 웹 페이지의 클러스터를 생성한 다음, 각 페이지의 클러스터에 빈번하게 접근하는 사용자들이 그룹된다. 이렇게 그룹된 사용자들의 그룹은 웹 서버에 저장되어 있는 사용자들의 등록데이터를 이용해서 각 그룹의 특성규칙을 생성한다.

<sup>†</sup> 이 논문은 한국 전자통신 연구원의 "Cyber-Post BI 기술 개발" 및 한국 과학재단 99년 특정 기초연구 사업의 연구비 지원에 의해 수행됨.

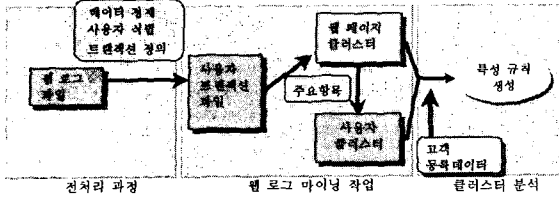


그림 1. 웹 사용자 클러스터 분석 모형과 처리과정

4. 웹 페이지 클러스터를 이용한 사용자 분류

4.1 전처리 과정

웹 로그 파일은 사용자에 의해 접근된 모든 파일의 레코드들로 구성된다. 각 로그 항목은 사용자의 IP주소, 사용자 ID, 접근 시간, 요구방법(GET, POST...), 접근된 페이지의 URL, 프로토콜, 반환코드와 전송된 데이터의 크기 등으로 구성된다. 이 로그 데이터는 불필요한 항목들을 포함하고 있기 때문에 데이터 마이닝을 수행하기 위한 전처리 과정은 필수적이다. 전처리 과정은 데이터 정제와 트랜잭션 정의의 두 가지 단계로 이루어진다. 데이터 정제 단계에서는 에러코드들 포함한 로그 항목과 다른 페이지를 요청할 때 목시적으로 불러지는 이미지 파일들(jpg, jpeg, gif, map 등)을 포함하는 항목과 사운드, 비디오, CGI 실행 파일 등의 이 논문에서 고려대상이 아닌 것은 제거된다.

이 논문에서는 웹 페이지에 접근하는 사용자가 쿠키 정보나 로그인 정보를 통해서 식별되었다고 가정된 상태에서 각각의 사용자에게 대한 클러스터링을 트랜잭션으로 분할한다. 하나의 트랜잭션이 자연스럽게 정의되는 바코드 분석과는 달리, 웹 로그에서 참조된 각각의 페이지에서 의미있는 트랜잭션을 정의하는 것은 간단한 문제가 아니다. 트랜잭션을 정의하는 방법은 [Cool99]에서 여러가지가 제시되고 있지만, 여기서는 사용자에 의해 정해진 최대 시간 간격을 고려해서 같은 사용자에 속하는 로그 엔트리들 기초하여 트랜잭션을 정의하였다. L을 접근 로그 엔트리의 집합이라고 하면, 하나의 로그 엔트리  $l \in L$ 은 IP주소  $l.ip$ , 사용자  $id$   $l.id$ , 사용자가 접근한 URL  $l.url$ , 접근 시간  $l.time$ 으로 구성된다.

정의 1: 하나의 트랜잭션  $t$ 는 다음과 같은 항목으로 이루어지며, 트랜잭션은 최대시간간격으로 구분된다.

$$t = \{ip_i, uid_i, \{l_j.url, - l_n.url\}, l_i.time\}$$

이 때,  $1 \leq k < j \leq n$ ,  $l_k.url \neq l_j.url$ ,  $l_k.uid \neq uid_j$  이고, 최대시간간격을  $\Delta t$ 라 할 때,

$$l_j.time - l_k.time \leq \Delta t$$

하나의 트랜잭션 내 URL 참조의 순서는 고려하지 않는다.

전처리 과정 후 사용자 별, 웹 페이지 별로 각각을 다시 정수형 코드로 1:1 매핑하여 두 개의 매핑 테이블을 형성하고, 생성된 매핑 테이블에 의해 트랜잭션 데이터베이스는 사용자 식별자, 트랜잭션 식별자, 항목, 시간들의 순서로 재구성된다.

표 1. 매핑 테이블과 재구성된 트랜잭션

코드	사용자	코드	웹 페이지 항목	코드	웹 페이지 항목
101	sbseo	1	/product/a.htm	6	/product/f.htm
102	ohp	2	/product/b.htm	7	/product/g.htm
103	junux	3	/product/c.htm	8	/product/h.htm
104	knam	4	/product/d.htm	9	/product/i.htm
105	halee	5	/product/e.htm		

TID	UID	항목	시간
100	101	{1,2,3}	17/Aug/2000:10:22:11
200	101	{1,2,3,4}	19/Aug/2000:12:30:28
300	102	{1,2,3,5}	03/Aug/2000:03:50:09
400	103	{1,2,6}	06/Aug/2000:19:15:21
500	104	{4,7,8}	11/Aug/2000:00:33:18
600	105	{4,7,9}	12/Aug/2000:18:50:25

4.2 시간을 고려한 페이지 클러스터링

이 절에서는 앞 절에서 정의된 트랜잭션에 대해서 주요 항목 개념을 사용하여 트랜잭션의 클러스터를 생성한다.

정의 2: 각 트랜잭션  $t_i$ 가 항목들의 집합  $\{i_1, \dots, i_n\}$ 일 때 트랜잭션의 집합  $\{t_1, \dots, t_n\}$ 을 보자. 클러스터링  $C$ 는  $\{t_1, \dots, t_n\}$ 를 분할한  $\{C_1, \dots, C_k\}$ 이다. 각  $C_i$ 를 클러스터라 한다.

각 트랜잭션을 클러스터링 하기 전에 시간 제약조건을 고려해서 조건에 맞는 트랜잭션만을 추출해야 한다. 사용자가 지정한 시간 제약조건은 클러스터링 하고자 하는 트랜잭션의 시작시간( $t_s$ )과 끝 시간( $t_e$ ) 쌍으로 이루어진 구간의 집합으로 이루어진다.  $\{(t_{s1}, t_{e1}), (t_{s2}, t_{e2}), \dots, (t_{sm}, t_{em})\}$ 으로 표현 될 때  $(t_{sm}, t_{em})$ 은 각각  $m$ 번째 트랜잭션 시작 시간, 끝시간 제약조건을 가리킨다. 트랜잭션  $t_n$ 의 시간  $t_n.time$ 이 이런 시간 제약조건을 만족시키는 트랜잭션의 집합  $T$ 가 추출된다.

$$T = \{t_n | t_{s1} \leq t_n.time < t_{e1} \cup t_{s2} \leq t_n.time < t_{e2} \dots \cup t_{sm} \leq t_n.time < t_{em}\}$$

트랜잭션의 클러스터의 유사도를 측정하기 위해 주요항목을 정의한다. 클러스터  $C_i$ 내에 항목의 지지도는  $C_i$ 에서 그 아이템을 포함하고 있는 트랜잭션의 수이다.  $|S|$ 가 집합  $S$ 의 원소의 개수를 나타내고, 사용자가 지정한 최소 지지도  $\theta$  ( $0 < \theta \leq 1$ )에 대해서  $C_i$ 내에서 항목의 지지도가 최소  $\theta * |C_i|$ 이면 그 항목은 클러스터  $C_i$ 의 주요항목이라 한다.

$Large_i$ 는  $C_i$ 의 주요항목의 집합을 나타내고  $Small_i$ 는  $C_i$ 의 비주요항목의 집합을 나타낸다고 했을 때 클러스터링  $C = \{C_1, \dots, C_k\}$ 을 고려해보자.  $C$ 의 비용은 intra-cluster cost와 inter-cluster cost 두 개의 요소로 계산되며, 이 비용이 최소화되는 클러스터링을 고려해야 한다.

Intra-cluster cost: 이 요소는 클러스터 내부의 비유사성에 가하는 것으로, 비주요 항목의 전체 수를 씬으로써 계산될 수 있다.

$$Intra(C) = \sum_{i=1}^k |Small_i|$$

이것은 너무 많은 비주요 항목을 가진 클러스터가 약하게 결합되는 것을 방지한다.

Inter-cluster cost: 이것은 클러스터 간 유사성에 대해 가해지는 것이다. 주요 항목이 클러스터의 유사성을 결정짓기 때문에, 각 클러스터는 중복되는 주요항목을 가능한 적게 가져야 한다. 이 중복은 다음과 같이 정의될 수 있다.

$$Inter(C) = \sum_{i=1}^k |Large_i| - |\bigcup_{i=1}^k Large_i|$$

Inter(C)는 다른 클러스터내에서 중복되어 있는 주요 항목을 측정하는 것으로서, 비슷한 클러스터가 생성되는 것을 방지한다.

두 가지를 함께 계산하기 위해 어떤 것을 중요하게 여기는 지에 따라 가중치를 줄 수 있다. 클러스터링  $C$ 의 비용함수는 다음과 같다.

$$Cost(C) = w * Intra(C) + Inter(C)$$

가중치  $w > 1$  은 클러스터 내의 유사성을 강조하며, 가중치  $w < 1$  은 클러스터간 비유사성을 강조한다.

트랜잭션의 클러스터링은 트랜잭션의 집합과 사용자가 지정한 최소 지지도가 주어졌을 때, Cost(C)가 최소인 클러스터링  $C$ 를 찾는 것이다.

표 1과 같은 시간제약조건을 고려한 재구성된 트랜잭션이 주어지고 최소 지지도가 60%(i.e.  $6 * 0.6 = 4$  트랜잭션)라고 할 때, 트랜잭션을 클러스터링 하는 예를 보면 다음과 같다.

예 1)  $C_1 = \{100, 200, 300, 400\}$   $C_2 = \{500, 600\}$ 에 대해서  $C_1$ 의 주요항목은 적어도 3 개의 트랜잭션에 포함되어 있어야 하며,  $C_2$ 의 주요항목은 적어도 2 개의 트랜잭션에 포함되어 있어야 한다. 따라서  $Large_1 = \{1, 2, 3\}$ ,  $small_1 = \{4, 5, 6\}$ ,  $arge_2 = \{4, 7\}$ ,  $small_2 = \{8, 9\}$ 이 되고 이에 따른 비용은  $Intra(C) = 5$ ,  $Inter(C) = 0$ ,  $Cost(C) = 5$ 가 된다.

예 2)  $C_1 = \{100, 200\}$   $C_2 = \{300, 400\}$   $C_3 = \{500, 600\}$ 에서, 위와 마찬가지로  $Large_1 = \{1, 2, 3\}$ ,  $small_1 = \{4\}$ ,  $Large_2 = \{1, 2\}$ ,  $small_2 = \{3, 5, 6\}$ ,  $Large_3 = \{4, 7\}$ ,  $small_3 = \{8, 9\}$  이고 비용은  $Intra(C) = 6$ ,  $Inter(C) = 2$ ,  $Cost(C) = 8$ 이다.

예 1의 방법이 Cost(C)가 적으므로 1의 방법으로 클러스터링 된다. 클러스터링 알고리즘은 [Wang99]에서 제시한 것을 사용하였다. 여기서 생성된 각 클러스터는 주요 항목을 공통으로 갖고 있는 트랜잭션들의 그룹이다. 그러나, 각 트랜잭션의 클러스터는 수백의 웹 페이지 참조를 포함하는 수천 사용자의 트랜잭션을 포함하기 때문에 사용자가 공통적으로 갖고 있는 행위 패턴의 통합된 관점이라고는 보기 힘들다. 우리가 알고자 하는 것은 웹 페이지의 클러스터이므로, 각 클러스터에서 비주요 항목을 제외한 주요항목만을 그 클러스터에 남긴다.

즉, 위의 예에서 웹 페이지의 클러스터는  $C_1 = \{1,2,3\}$ ,  $C_2 = \{4,7\}$  이 된다. 결과적으로 생성되는 페이지의 클러스터는 주요 항목이 되는 페이지들의 그룹이 된다.

4.3 사용자의 클러스터 생성

이런 페이지의 클러스터에 속하는 사용자들의 분류는, 각 클러스터에 속해있던 트랜잭션의 사용자 ID를 검색해서 그 사용자의 전체 트랜잭션의 수에 대한 각 클러스터의 주요항목을 포함하고 있는 트랜잭션 수의 비율을 살펴봄으로써 그 클러스터에 속할 것인지를 결정한다. 즉, 다음과 같은 경우에 사용자는 그 클러스터에 속하게 된다.

클러스터의 주요항목을 포함하는 트랜잭션 수 / 각 사용자의 전체 트랜잭션 수  $\geq$  minsup

앞서 기술했던 예에 위의 식을 적용하면  $C_1$ 에는 사용자가 101,102 가 포함되고,  $C_2$ 에는 사용자 104,105 가 포함되는 것을 알 수 있다. 이렇게 생성된 각 클러스터는 비슷한 접근 패턴을 가진 사용자들의 그룹을 나타낸다.

5. 각 클러스터의 특성화 규칙 생성

주요항목 집합에 기반해서 사용자들의 클러스터가 생성되면, 서버에 저장되어 있는 사용자의 등록정보를 이용해서 각 클러스터에 대한 특성 규칙을 생성한다. 특성화 규칙은 데이터의 집합이 주어졌을 때, 그것의 일반적인 특성들을 추출해 내는 것으로, 여기서는 이것을 위해 속성 지향 귀납법을 이용한다.

실험에서는 사용자의 속성 중 나이, 직업, 거주지를 선택하여 그림2와 같은 개념 계층을 명시적으로 정의해서 특성규칙을 찾도록 하였다.

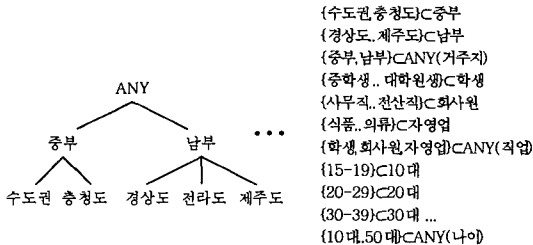


그림 2. 개념 계층

이런 개념 계층과 특성규칙을 생성한 데이터 집합, 데이터를 어느 정도까지 일반화 할 것인가를 나타내는 속성 임계값을 입력으로 받아들이며 특성규칙을 찾는다. 표 2는 데이터베이스에서 사용자 클러스터  $C_1$ 의 특성규칙을 발견하기 위해  $C_1$ 에 속한 사용자의 속성들을 추출해낸 것이다.

표 2. 사용자의 속성

직업	거주지	나이	Count
영업직	충청도	30대	35
전산직	수도권	30대	10
사무직	경상도	30대	30
대학생	수도권	20대	10
대학원생	수도권	20대	15

위와 같은 relation에서 일반화 과정을 더 수행하면 [표 3]과 같은 결과가 나온다.

표 3. 일반화된 relation

직업	거주지	나이	Count
회사원	ANY	30대	75
학생	수도권	20대	25

위의 테이블에서 추출해낸  $C_1$  클러스터의 사용자 특성과 규칙은 다음과 같다.

$$V(x)C_1(x) \rightarrow ((\text{직업}(x) \in \text{회사원} \wedge \text{나이}(x) \in 30\text{대}) [75\%] \vee (\text{거주지}(x) \in \text{수도권} \wedge \text{직업}(x) \in \text{학생} \wedge \text{나이}(x) \in 20\text{대}) [25\%])$$

$C_1$ 에 대한 이런 결과를 봤을 때,  $C_1$ 에 속하는 페이지인 /product/a.htm, /product/b.htm, /product/c.htm 페이지에 접근하는 사용자의 75%가 직업이 회사원이고 나이가 30대, 25%가 수도권 지역에 살고 직업이 학생이며 나이가 20대 라는 특성과 규칙을 유도할 수 있다. 각 클러스터에 대해 이런 규칙을 유도해 냄으로써, 고객의 특성을 명확히 파악할 수 있고, 새로운 고객이 등록했을 때, 그 고객의 속성을 사용하여 개인화된 웹 페이지 구성 및 타겟 마케팅에 이용할 수 있다.

6. 실험 결과

이 논문에서는 웹 페이지 클러스터링을 위해 [Wang99]의 알고리즘을 사용하였으며, 실험데이터로는 인터넷 쇼핑몰 서버의 웹 로그 파일을 적용하였다. 웹 서버와 자바환경은 UNIX SUN OS 5.6 과 JDK 1.2를 사용하였고, 오라클 7.3.4 DBMS를 이용하여 실험하였다.

그림 3은 시간 제약조건으로 연말과 설, 추석에 해당하는 기간을 주고 로그파일을 분석한 결과를 보여주고 있다. 먼저 시간 제약조건을 만족하는 트랜잭션만을 추출한 다음, 주어진 최소 지지도를 만족하는 주요항목 집합을 이용하여 웹페이지 클러스터와 그에 속하는 사용자를 분류한다.

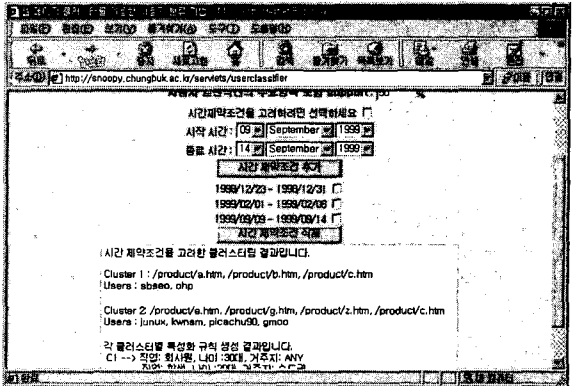


그림 3. 웹 페이지와 사용자 클러스터링 결과

7. 결론 및 향후 연구

이 논문에서는 웹 사이트의 사용자 개인에게 좀더 개인화된 서비스를 제공하고 차별화된 마케팅을 지원할 수 있는 방법으로써, 웹 로그를 분석하여 사용자의 트랜잭션 내에서 최소발생빈도를 가지는 웹 페이지를 주요항목이라는 관점에서 분석했다. 이 주요항목을 이용해서 웹 페이지를 클러스터링 하고 사용자들 그룹핑하는 기법에 대해 기술하였다. 또한 트랜잭션의 클러스터링에 있어 시간 제약조건을 고려함으로써 특정 기간이나 계절별로 효율적인 마케팅 전략을 세울 수 있는 방법을 제시하였다. 향후 웹 페이지 클러스터링을 이용해서 동적인 서비스를 제공할 수 있는 방안과 시간 제약조건에의 효율적인 적용에 대한 연구가 필요하다.

8. 참고문헌

[Cool99] R. Cooley, B. Mobasher and J. Srivastava, Data Preparation for Mining World Wide Web Browsing Patterns, In the Journal of Knowledge and Information Systems, Vol. 1, No. 1, 1999.  
 [Han96] J. Han and Y. Fu, Exploration of the Power of Attribute Oriented Induction in Data Mining, U.M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy (eds.), Advances in Knowledge Discovery and Data Mining, AAAI/MIT Press, 1996.  
 [Han97] E.H. Han, G. Karypis, V. Kumar and B. Mobasher, Clustering based on association rule hypergraphs. SIGMOD workshop on research issues on Data Mining and Knowledge Discovery, 199.  
 [Moba99] B. Mobasher, R. Cooley and J. Srivastava, Creating Adaptive Web Sites Through Usage -Based Clustering of URLs, KDEX'99, November 1999.  
 [Wang99] Ke Wang and Chu Xu, Clustering Transactions Using Large Items, In ACM CIKM-99, 1999.  
 [서성보2000] 서성보, 류근호, 주요 항목 집합을 이용한 문서 클러스터링 및 연관 탐사 기법, 정보과학회 춘계학술대회, 2000.  
 [이강태99] 이강태, 시간연관규칙 탐사 기법, 충북대학교 석사학위 논문, 1999.