2-계층 클러스터링을 사용한 웹 사용자 그룹의 행동규칙추출방법에 관한 연구

황 준 원[†] · 송 두 헌[††] · 이 창 훈[†††]

요 약

유용한 웹 사용자 그룹을 파악하고 이들의 행동패턴을 찾는 것은 eCRM에서 매우 중요하다. 그러나 온라인 사용자 데이터에는 불확실한 정보가 많이 포함되어 있어 이를 바탕으로 유사한 성향을 가진 사용자 그룹을 생성하는 경우 신뢰성이 떨어지는 문제점이 있다. 본 논문에서는 불확실성이 포함된 사용자와 페이지의 서로 다른 두 데이터 계층의 상호작용을 통해 좀 더 신뢰성 있는 사용자 그룹을 생성하고 데이터에 내재된 이들의 행동패턴을 추출하는 방법을 제시하였다. 그리고 C4.5를 사용하여 생성된 행동규칙과의 비교를 통해 본 논문에서 제시하는 방법과의 비교분석을 실시하였다.

키워드 : 웹 마이닝, 클러스터링, eCRM, C4.5

A Study on Behavior Rule Induction Method of Web User Group using 2-tier Clustering

JunWon Hwang[†] · Doo Heon Song[††] · ChangHoon Lee[†††]

ABSTRACT

It is very important to identify useful web user group and induce their behavior pattern in eCRM domain. Inducing user group with a similar inclination, a reliability of user group decreases because there is an uncertainty in online user data. In this paper, we have applied the 2-tier clustering, which uses the outcome of interaction with data from other tiers. Also we propose a method which induces user behavior pattern from a cluster and compare C4.5 with our method.

Key Words : Web Mining, Clustering, eCRM, C4.5

1. 서 론

온라인에서 수집한 사용자 데이터를 분석하여 유사한 성질을 가진 사용자 그룹을 파악하고 그들의 행동특성을 분석하는 일은 eCRM 뿐만 아니라 온라인 사이트의 운영을 위해서도 중요한 일이다[1]. 일반적으로 사용자가 직접 입력한 온라인 등록 정보 및 기타 다른 경로를 통해 확보한 데이터에 대해 데이터마이닝 기법을 사용하여 사용자 그룹을 생성한다[2, 3]. 그리고 웹 페이지를 대상으로 유사도가 높은 웹 페이지 그룹을 생성하고 웹 로그 분석을 통해 생성된 그룹이 온라인 상에서 어떤 행동특성을 갖는지 파악한다[4, 5]. 그러나 사용자가 입력한 인구통계학적 정보들은 부정확하거나 입력되지 않은 경우가 많아 사용자 데이터에는 많은 불확실성이 내포되어 있다. 또한 웹 로그에 기록된 사용자의 페이지 방문정보, 구매정보 등과 같은 행동정보량이 부족하여 이를 바탕으로 유사한 행동패턴을 갖는 사용자 그룹을 파악하는 것은 쉽지 않은 일이다. 이러한 문제점을 해결하기 위해 페이지 클러스터를 생성하고 웹 로그 분석을 통해 사용자와 페이지 클러스터의 연관정보를 구축한 후 사용자 데이터를 대상으로 클러스터링 할 때 연관정보를 함께 이용함으로써 사용자 그룹의 정확성을 높여주는 방식이 제시되었다[6].

행동규칙은 데이터에 내재된 사용자의 행동패턴을 규칙의 형태로 표현한 것으로 본 논문에서는 생성된 사용자 그룹에서 행동규칙을 추출하는 방법을 제시하였다. 또한 규칙 생성을 위해 많이 사용되는 C4.5[7]를 사용자 데이터에 적용하여 행동규칙을 생성하고 2-계층 클러스터링 방법과의 비교분석을 실시하여 제안하는 방법의 타당성을 검증하였다. 그리고 서로 다른 데이터 계층에 여러 클러스터링 알고리즘을 적용하고 사용자 데이터의 불확실성의 정도와 페이지 계층에 대한 연관정보의 양에 대한 다양한 셋팅을 구성하였다. 이를 통해 여러 조건에서 2-계층 클러스터링과 단일계층 클

† 준 회 원 : 건국대학교 컴퓨터공학과 박사과정수료
†† 정 회 원 : 용인송담대학 컴퓨터게임정보과 교수
††† 종신회원 : 건국대학교 컴퓨터공학과 교수
논문접수 : 2007년 8월 31일, 심사완료 : 2007년 10월 29일

DOI: 10.3745/KIPSTD.2008.15-D.1.139

러스터링과의 품질분석을 하였다.

본 논문의 구성은 다음과 같다. 2장에서는 2-계층 클러스터링의 개념과 이를 웹에 적용하는 방법을 설명하였고, 3장에서는 생성된 클러스터에서 행동규칙을 추출하고 평가하는 방법을 설명하였고, 4장에서는 단일계층 및 2-계층 클러스터링에서 생성된 클러스터의 품질분석을 실시하고 C4.5와 행동규칙을 비교하였으며 5장에서는 결론을 기술하였다.

2. 관련연구

2.1 클러스터링 알고리즘

클러스터링 알고리즘은 크게 분할적 클러스터링(Partitional Clustering)과 계층적 클러스터링(Hierarchical Clustering) 알고리즘으로 분류할 수 있다. 분할적 클러스터링은 클러스터의 계층을 고려하지 않고 평면적으로 클러스터링 하는 방법으로 일반적으로 미리 몇 개의 클러스터로 나뉘어 질 것인지를 예상하고 클러스터의 개수를 정한다. K-means는 분할적 클러스터링 방법으로 전체 데이터 집합에서 임의로 k개의 seed point를 선정하고 이를 클러스터의 중점으로 시작한다. 이후 데이터 개체에 대한 소속 클러스터의 재할당 과정과 목적함수의 평가를 반복적으로 수행하여 목적함수를 최적화 한다.

계층적 클러스터링은 가장 유사한 두 개체들 선택하여 병합해 가는 병합적 계층 클러스터링(Agglomerative Hierarchical Clustering) 방법과 가장 먼 개체들 선택하여 나누어 나가는 분할적 계층 클러스터링(Divisive Hierarchical Clustering) 방법이 있다. 두 클러스터의 유사도를 측정하는 방법에는 최단연결법, 최장연결법, 중심연결법, 평균연결법이 있다[8].

2.2 2-계층 클러스터링 개념

2-계층 클러스터링 기법은 한 계층의 클러스터링 결과를 다른 계층의 클러스터링에 이용하는 방법이다. 이는 클러스터링을 할 때 자신의 컨텐츠 정보 뿐만 아니라 다른 계층의 데이터에 대한 링크 정보를 함께 이용함으로써 컨텐츠의 부정확성에서 오는 오차를 줄일 수 있는 장점이 있다[9]. (그림 1)은 2-계층 클러스터링의 개념을 나타낸 것으로 다른 계층에 대한 링크정보를 이용하여 클러스터링을 수행하는 과정을 보여주고 있다.

상위계층과 하위 계층에 6개의 노드로 이루어진 2개의 데이터 집합이 있고 두 데이터 사이에는 링크 정보가 존재한다. 상위 계층의 노드 1의 하위 계층에 대한 링크정보는 [1, 0, 0, 0, 0, 0]이고 노드 5의 링크정보는 [0, 0, 1, 0, 1, 1] 이다(단계1). 두 계층 가운데 한 계층에서 컨텐츠 정보에 의해 클러스터링이 실시되는데 (그림 1)에서는 하위계층에서 먼저 실시되었다(단계 2). 하위 계층의 노드들이 클러스터화 되면서 상위 계층의 노드에 대한 링크정보는 클러스터에 대한 링크정보로 갱신된다(단계 3). 상위 계층의 노드 1의 링크정보는 [1, 0]로 노드 5의 링크정보는 [1, 1]로 갱신된다. 노드 5의 경우 1번 클러스터에 대한 링크정보값이 2번 클러스터에 대한 것보다 더 크다. 따라서 상위 계층의 클러스터링을 할 때 다른 계층에 대한 링크 정보를 이용한다면 노드 5는 1번 클러스터에 속하게 된다.

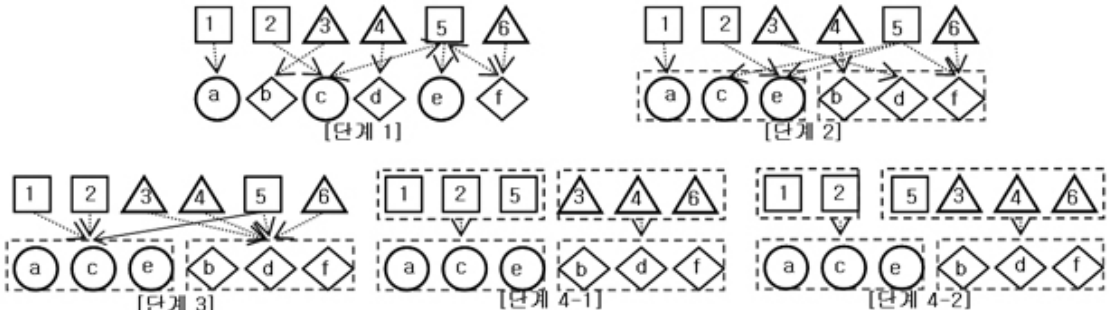
2-계층 클러스터링을 위해서는 두 계층을 서로 연관시켜 줄 수 있는 방법이 있어야 하는데 식 (1)은 2-계층 클러스터링의 거리공식이다.

$$D = \alpha D_c + (1-\alpha) D_l,\ 0 \le \alpha \le 1 \qquad (1)$$

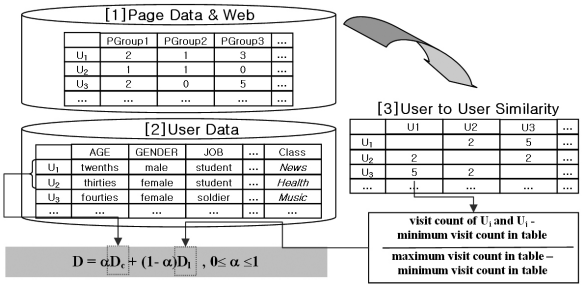
D_c는 컨텐츠 클러스터링에서의 거리값이고 D_l는 다른 계층과의 링크정보값이다. α는 상수값으로 어느 계층에 더 가중치를 부여할지를 설정해준다. (단계2)에서의 하위 계층의 클러스터링의 경우 α는 1이므로 컨텐츠 정보만을 사용하였다. (단계 4-1)에서의 클러스터링에서 α는 0으로 링크정보만을 사용하였다. 만일 컨텐츠 정보가 링크정보보다 더 높은 비중으로 사용되었더라면 노드 5가 1번 클러스터에 할당되지 않았을 수도 있고 이는 두 계층의 정보를 모두 이용했기 때문인데 어느 계층에 더욱 가중치를 두었느냐에 따라 달라질 수 있다(단계 4-2). 따라서 컨텐츠 정보와 링크정보를 모두 사용하는 0<α<1인 경우만이 2-계층 클러스터링이고 α=0 혹은 α=1 인 경우는 단일 계층 클러스터링을 나타낸다.

2.3 2-계층 클러스터링의 웹 적용

2-계층 클러스터링은 서로 다른 두 데이터 계층이 있고 그 사이에는 두 계층을 연관시켜 주는 링크정보가 존재해야 적용이 가능하다. 웹 데이터는 크게 사용자 데이터와 페이지 데이터로 구분 할 수 있고 이 둘간의 링크정보는 웹 로그의 사용자의 페이지 방문정보를 분석함으로써 얻을 수 있



(그림 1) 2-계층 클러스터링의 개념



(그림 2) 웹 에서의 2-계층 클러스터링

다[10]. (그림 2)는 2-계층 클러스터링 공식을 사용하여 사용자 1과 사용자 2간의 거리를 구하는 것을 보여주고 있다.

먼저 D_c를 구하기 위해 사용자 1과 사용자 2간의 컨텐츠 정보에 대해 범주형(nominal) 데이터의 경우는 VDM[11] 기법을 사용한 후 유클리디안 거리를 적용한다. 그리고 페이지 데이터의 컨텐츠 정보를 이용하여 페이지 클러스터를 구하고 웹 로그를 분석하여 사용자와 페이지그룹간의 방문정보 테이블을 생성한다. 그리고 이것으로부터 사용자 간의 유사도 테이블을 생성하고 D_l을 구한 후 D_c와 더한다. 여기서 D_l, D_c은 똑같은 범위의 거리값을 갖는데 $0 \le D_l$, $D_{cl} \le 1$ 이다.

사용자 클러스터를 생성 하기 위한 2-계층 클러스터링 알고리즘은 다음과 같다.

단계 1 : 페이지 계층에서 페이지 컨텐츠를 이용하여 페이지 클러스터를 생성한다. 웹 로그에 있는 사용자의 페이지 방문정보와 생성된 페이지 클러스터를 이용해 <사용자 - 페이지 클러스터> 간의 유사도 테이블을 작성한다.

단계 2 : 사용자 계층에서 사용자 컨텐츠와 전단계에서 만들어진 <사용자 - 페이지 클러스터> 간의 유사도 테이블의 링크 정보를 이용하여 사용자 클러스터를 생성한다. 웹 로그에 있는 사용자의 페이지 방문정보와 생성된 사용자 클러스터를 이용하여 <사용자 클러스터 - 페이지> 간의 유사도 테이블을 작성 또는 업데이트 한다.

단계 3 : 페이지 계층의 컨텐츠와 <사용자 클러스터 - 페이지> 간의 유사도 테이블의 링크 정보 값을 이용하여 클러스터링 한다. 생성된 클러스터를 이용해 <사용자 - 페이지 클러스터> 간의 유사도 테이블을 업데이트 한다.

단계 4 : 생성된 전체 클러스터들의 밀도의 합이 전 단계보다 낮으면 단계 2로 가고 높으면 종료한다.

3. 행동 규칙 추출

본 논문에서의 실험은 실제 데이터와 유사하게 가상데이터를 생성하여 실시되었다. 이를 위해 필요한 데이터 생성 방법과 생성된 클러스터로부터 행동규칙을 추출하고 평가하는 방법을 제시하였다.

3.1 행동규칙 설정방법

일반적으로 사용자들은 고유한 행동패턴을 갖는다. 이를 위해 사용자들의 특정한 행동패턴을 데이터 속에 내재시키기 위하여 데이터 생성 규칙을 적용하였는데 이것을 통해 각 사용자 데이터와 웹 로그의 내용이 결정된다. 데이터 생성 규칙은 내재규칙과 분포결정규칙으로 구성된다. 내재규칙은 데이터 내의 행동패턴을 규칙으로 정의한 것이고 분포결정규칙은 각 내재규칙이 어느 정도의 경향을 가지고 그 행동패턴들을 결정하는지를 정의한 것으로 생성규칙은 다음과 같은 내용을 갖는다.

(속성, 속성 값, 분포도)$^+$ → (방문자 비율, 페이지 클래스, 방문률) (2)

예를 들어 "나이가 10대이고 학문이 취미인 학생들은 과학 관련 페이지를 자주 방문한다" 라는 행동패턴에 대한 생성규칙은 "(AGE 10대 20) ∧ (JOB 학생 50) ∧ (HOBBY 과학 50) → (60 과학 40)" 로 표현 할 수 있다. 이 생성규칙의 내용은 전체 데이터 중에서 10대가 차지하는 비율을 20%로 하고 그 중에서 직업이 학생인 비율이 50%, 취미가 과학인 비율이 50%로 사용자데이터를 생성한다. 그리고 이러한 데이터에서 60%의 사용자에 대해 사이트 내의 "과학" 관련 페이지 방문률을 40%로 하고 나머지 60%는 임의로 다른 페이지를 방문하도록 생성함을 의미한다. 그리고 생성규칙에서 지정되지 않는 내용은 모두 임의로 설정되었다.

3.2 행동 규칙 추출 및 평가방법

생성된 클러스터의 품질(Quality)을 평가하기 위해 정보검색 분야에서 널리 사용되고 있는 재현률(Recall)과 정확도(Precision)를 사용하였다[12]. 전체 클러스터의 크기가 $|C|$ 이고 i번째 클러스터의 개체 크기가 $|C_i|$이며 클러스터 내에서 이 행동패턴을 따르는 개체 수를 $|C_i'|$, 전체에서의 수를 $|P_j|$ 라고 하면 재현률과 정확도는 다음과 같다.

$$\mathrm{Re}call = \left(\sum_{i=1}^{|C|} |C_i'|\right) / |P_j| \quad (3)$$

$$\mathrm{Pr}ecision = \sum_{i=1}^{|C|} \frac{|C_i'|}{|C_i|}. \quad (4)$$

사용자 계층의 데이터를 클러스터링하여 생성된 클러스터의 대표값은 그 사용자그룹의 대표값으로 이는 행동규칙에서 조건부의 내용을 의미한다. 클러스터 내의 각 속성의 최대 빈발값을 대표값으로 선정하면 개체가 n개의 속성으로 이루어진 경우에 클러스터는 n개의 값으로 표현된다. 이는 항상 n개의 조건값을 갖는 규칙을 생성하기 때문에 필요 없는 조건이 제거된 최적화된 규칙을 표현할 수 없다. 따라서 최대 빈발값이 그 속성값들을 대표할 수 있는지의 여부를 판별하여 그 비율이 유의수준을 넘을 경우 조건으로 채택하고 그렇지 못한 경우는 기각함으로써 클러스터가 갖는 규칙의 유연성을 부여하였다. 본 논문에서는 추출된 최대빈발값이 95%의 신뢰성을 갖도록 유의수준의 설정을 5%로 하

였다. 그리고 생성된 각 클러스터가 갖는 개체의 크기 $|C_i|$에 따라 $|C_i| \le 8$개인 경우는 이항분포, $8 < |C_i| < 30$인 경우는 카이제곱분포, $30 \le |C_i|$인 경우는 표준정규분포를 각각 적용하였다.

추출된 행동규칙이 데이터에 내재규칙을 어느 정도 포함하고 있는지 분석하여 추출된 규칙의 유용성을 측정하였다. 속성 집합 $A = \{a_1, a_2, ..., a_l\}$이고, 각 속성이 갖는 범주형 데이터 값의 크기집합 $S = \{s_1, s_2, ..., s_m \mid 1 \le m \le |A|\}$ 일 때, i번째 속성 값의 집합 $V_i = \{v_{i1}, v_{i2}, ..., v_{ij} \mid 1 \le i \le |A|, 1 \le j \le s_j\}$ 이다. $v_i \in V$ 에 대해 데이터에 내재시킨 규칙 $r = "v_1 \wedge v_2 \wedge ..... \wedge v_m \rightarrow page\ class"$ 이고 찾은 규칙은 $r' = "v_1' \wedge v_2' \wedge ..... \wedge v_n' \rightarrow page\ class"$ 이다. $v_i \neq v_i'$인 속성의 수가 k개 라면 추출규칙 r'를 평가하는 공식은 다음과 같다.

- $if\ k = 0$: 추출된 모든 규칙 값이 내재규칙에 존재하는 경우

$$Eval(r') = n/m$$

- $if\ k \ge 1$: 추출된 규칙 값이 내재규칙과 다른 것이 존재하는 경우

$$Eval(r') = 1 - \frac{k}{|A|}$$

생성 가능한 모든 페이지 클래스 수가 N이고 발견된 페이지 클래스 수는 N'이다. 행동규칙의 목적패턴에서 같은 페이지 클래스를 갖는 클러스터의 집합을 $W = \{w_1, w_2, ..., w_i \mid 1 \le i \le N\}$라 하면

$$Inst(w) = \sum_{i=1}^{|w_i|} |C_i|, \quad Inst(w') = \sum_{i=1}^{|w_i|} |C_i'|$$

페이지 클래스 단위별 규칙 평가도 $Eval(R_i)$와 추출된 행동 패턴의 다양성을 반영하는 전체 규칙 평가도 $Eval(R)$은 다음과 같다.

$$Eval(R_i) = \sum_{i=1}^{i=N'} \frac{Inst(w_i')}{Inst(w_i)} \times Eval(r'_i) \quad (5)$$

$$Eval(R) = Eval(R_i) \times \frac{N'}{N} \quad (6)$$

4. 실 험

성질이 다른 두 계층의 데이터에 여러 클러스터링 알고리즘을 사용하는 메타모델을 구성하여 클러스터링 알고리즘의 조합이 어떤 성능을 내는지를 다양한 데이터 셋팅을 통해 살펴보았다. 그리고 생성된 클러스터의 품질을 분석하기 위해 정확도와 재현률을 적용하였고 클러스터에서 추출한 행동규칙의 품질을 분석하기 위해 규칙평가도를 사용하여 C4.5와 비교하였다.

4.1 실험환경 및 데이터

사용자 계층의 데이터는 연령(10대, 20대 등 5개), 성별(남, 여), 결혼여부(기혼, 미혼), 직업(학생, 군인 등 20개), 취미(컴퓨터, 건강 등 10개)의 5개 속성을 가지며 1000 명으로 구성되었다. 웹 문서 계층은 Open Directory Project[13]에 있는 카테고리로부터 10개를 선정하여 각각 100개의 페이지를 임의 선택함으로써 10개 클래스와 1000개의 페이지로 구성되었다. 10개 페이지 클래스에 대해 10개의 행동규칙을 설정하였고 모든 사용자는 각자 20개의 페이지를 방문하도록 웹 로그 데이터가 생성되었다. 사용자 데이터의 경우 3.1 절에서 언급한 식 (2)에서 방문자 비율을 30%와 90%로 방문율을 40%와 80%의 2가지로 설정하였는데 이 수치값은 단순히 비율의 높고 낮은 정도를 설정한 것이다. 이러한 설정은 2-계층 클러스터링의 거리공식에서 D_c와 D_l에 영향을 주는 것으로 D_c와 D_l의 성능이 전체에 미치는 영향을 분석하기 위한 것이다. 클러스터링 알고리즘은 크게 분할적 클러스터링 알고리즘과 계층적 클러스터링 알고리즘으로 분류할 수 있는데 본 논문에서는 각 계열에서 많이 이용되는 K-means와 병합적 계층 알고리즘을 사용하여 각 계층에 적용하였다. 이를 통해 서로 다른 특성을 갖는 클러스터링 알고리즘의 조합이 전체 성능에 어떤 영향을 미치는지 살펴보았다.

4.2 클러스터의 품질 분석

생성된 클러스터의 품질을 분석하기 위해 일반적으로 널리 사용되는 정확도와 재현률을 측정하였는데 실험결과에서 K-means 알고리즘은 K로 병합적 계층 클러스터링 알고리즘은 H로 표기하였다. 모든 표들은 식 (1)에서 가중치 α의 값을 0부터 1까지 증가시켜 10회 실험하여 구한 평균값을 나타낸 것이다. 2-계층 클러스터링 구간은 0<α<1으로 표의 내용은 이 구간의 평균값이다.

정확도 분석에서 사용자 데이터 계층만을 이용한 단일계층 클러스터링(α=1)은 30.1%이지만 2-계층 클러스터링의 경우는 64.3%로 매우 좋은 결과를 보여주고 있다. 그리고 링크정보만을 이용한 단일계층 클러스터링(α=0)의 경우도 58.7%의 높은 정확도를 보여주고 있다. 링크정보는 페이지 데이터 계층과 사용자의 웹 로그 분석을 통해 생성되는데 방문자비율과 페이지 방문률이 높을수록 좋은 결과를 보인다. 이는 같은 행동유형을 갖는 사용자 비율이 높을수록 2-계층 클러스터링에서 컨텐츠 클러스터링의 정확도가 높아지기 때문이다. 또한 페이지 방문률이 높을수록 페이지 방문기록에 대한 정보량이 증가함으로써 사용자 컨텐츠의 부정확성에서 오는 오차를 링크정보를 통해 줄일 수 있는 확률이 높아지기 때문이다. 재현률의 경우도 정확도와 유사한 패턴을 가져 α=0과 α=1일 때 46.4%와 71.1%이나 2-계층 클러스터링의 경우 75.5%의 좋은 결과를 보여주고 있다.

〈표 1〉 데이터 구성에 따른 정확도

알고리즘		방문자 비율 - 방문률	α=0	α=1	2-tier
Page	User				
H	K	30-40	0.34	0.14	0.34
H	K	30-80	0.56	0.14	0.62
H	K	90-40	0.48	0.48	0.55
H	K	90-80	0.75	0.45	0.80
H	H	30-40	0.34	0.18	0.35
H	H	30-80	0.77	0.15	0.85
H	H	90-40	0.41	0.44	0.54
H	H	90-80	0.82	0.46	0.89
K	K	30-40	0.34	0.13	0.38
K	K	30-80	0.62	0.13	0.65
K	K	90-40	0.54	0.44	0.59
K	K	90-80	0.80	0.45	0.85
K	H	30-40	0.37	0.18	0.44
K	H	30-80	0.85	0.15	0.9
K	H	90-40	0.52	0.44	0.6
K	H	90-80	0.88	0.46	0.93
평균			0.587	0.301	0.643

〈표 2〉 데이터 구성에 따른 재현률

알고리즘		방문자 비율 - 방문률	α=0	α=1	2-tier
Page	User				
H	K	30-40	0.41	0.42	0.46
H	K	30-80	0.74	0.35	0.78
H	K	90-40	0.61	0.64	0.67
H	K	90-80	0.89	0.63	0.92
H	H	30-40	0.43	0.27	0.47
H	H	30-80	0.87	0.25	0.92
H	H	90-40	0.51	0.61	0.65
H	H	90-80	0.90	0.62	0.94
K	K	30-40	0.53	0.37	0.53
K	K	30-80	0.80	0.31	0.81
K	K	90-40	0.72	0.59	0.75
K	K	90-80	0.90	0.61	0.94
K	H	30-40	0.54	0.27	0.57
K	H	30-80	0.93	0.25	0.96
K	H	90-40	0.66	0.61	0.73
K	H	90-80	0.94	0.62	0.98
평균			0.711	0.464	0.755

서로 다른 성질을 갖는 사용자 데이터와 페이지 데이터 계층에 K-means와 병합적 계층 클러스터링 기법을 적용하여 어느 경우에 좋은 성능을 보이는 지를 살펴보았다. 알고리즘에 관계없이 정확도와 재현률 모두 항상 2-계층 클러스터링이 좋은 결과를 보여주었다. 정확도의 경우 2-계층 클러스터링이 단일계층 클러스터링 α=0, α=1일 때보다 5.6% 35.3%가 향상되었고 재현률의 경우는 4.4%, 29.1%의 향상을 보여주었다. 특히 페이지 계층에 관계없이 사용자 계층에 병합적 계층 클러스터링 기법을 사용하였을 때 더 좋은 성능을 보여주었다.

4.3 클러스터의 규칙평가도 분석

클러스터에서 추출된 규칙과 C4.5의 규칙에 대해 식 (6)의 규칙평가도를 적용하여 데이터에 내재된 행동패턴을 얼마나 잘 나타내었는지 분석하였는데 일반적으로 클러스터의 정확도와 재현률이 높은 2-계층 구간에서 규칙평가도의 값도 비례하여 높게 나타남을 알 수 있었다.

먼저 C4.5는 α(가중치)의 영향을 받지 않으므로 동일한 결과를 나타내지만 2-계층 클러스터링의 경우 2-계층 클러스터링 구간(0<α<1)에서 우수한 결과를 보이고 있다. 특히 페이지 계층의 알고리즘에는 관계없이 사용자 계층에 병합적 계층 클러스터링 알고리즘을 사용했을 때 우수한 결과를 나타내고 있다. 방문자 비율이 30%인 경우에는 비교적 2-계층 클러스터링이 C4.5 보다 우수하고 90%인 경우에는

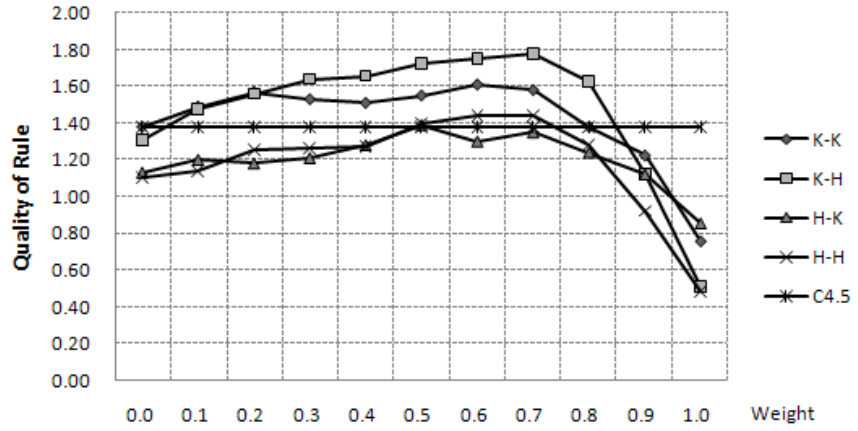
〈표 3〉 알고리즘에 따른 정확도

알고리즘		α=0	α=1	2-tier
Page	User			
H	K	0.53	0.30	0.58
H	H	0.59	0.26	0.66
K	K	0.58	0.29	0.62
K	H	0.66	0.31	0.72
평균		0.587	0.290	0.643

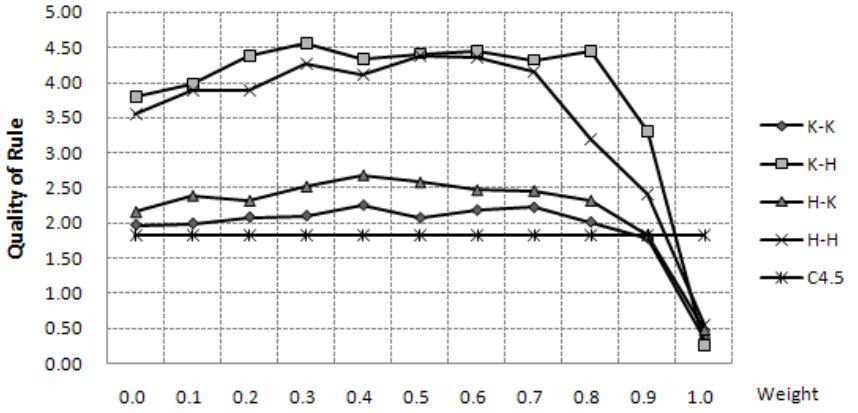
〈표 4〉 알고리즘에 따른 재현률

알고리즘		α=0	α=1	2-tier
Page	User			
H	K	0.66	0.51	0.71
H	H	0.68	0.44	0.75
K	K	0.74	0.47	0.76
K	H	0.77	0.44	0.81
평균		0.711	0.464	0.755

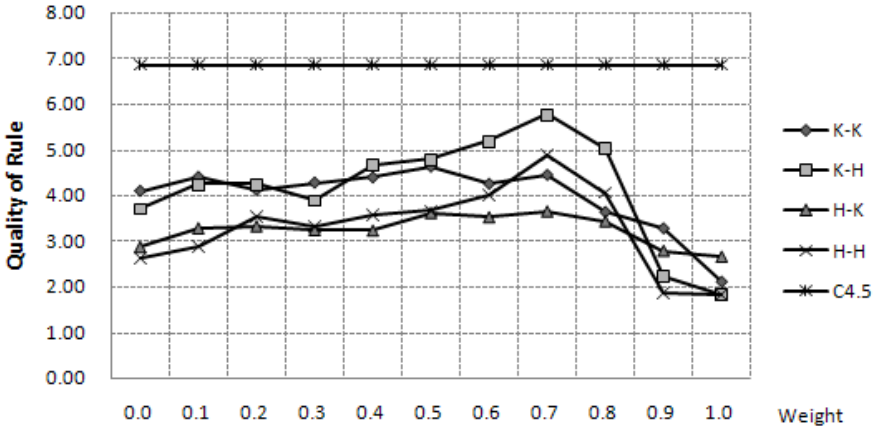
C4.5가 2-계층 클러스터링보다 좋은 결과를 나타내고 있다. 방문자 비율이 높다는 것은 전체 사용자 데이터에서 같은 행동패턴을 갖는 사용자들이 많다는 것을 나타내며 이들은



(그림 3) 규칙평가도 (30%, 40%)

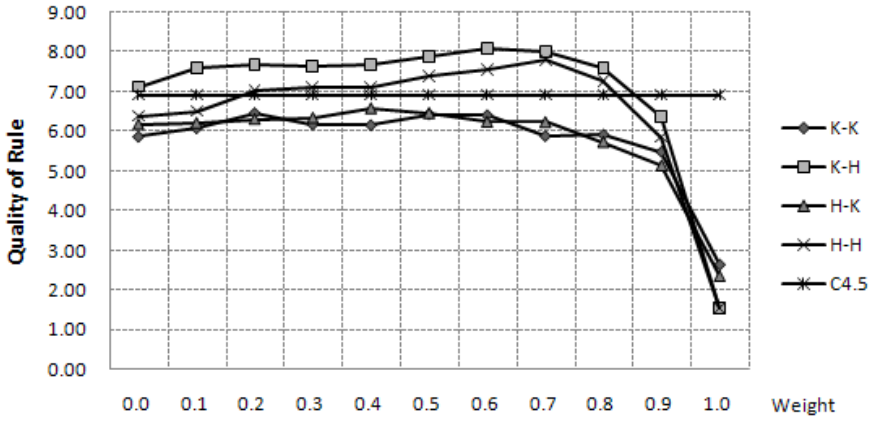


(그림 4) 규칙평가도 (30%, 80%)



(그림 5) 규칙 평가도 (90%, 40%)

서로 유사한 컨텐츠 정보를 갖는다. C4.5는 같은 행동패턴을 갖는 사용자들을 분류할 때 가장 영향을 미치는 속성 값을 선정하고 이것들은 곧 사용자 그룹이 갖는 행동규칙의 조건부의 내용을 구성하게 된다. 방문자 비율이 높으면 속성값을 선택할 때 오차를 줄이게 되어 정확도를 높일 수 있기 때문에 방문자 비율이 높을수록 월등히 좋은 성능을 보인다. 그러나 방문률이 40%에서 80%로 높아지는 경우에는 별 차이를 보이지 않는다. 방문률은 사용자의 컨텐츠 정보와는 관계없이 페이지 계층에 대한 연관성을 높이는 효과를 주는 것으로 C4.5에서는 이 정보를 이용하지 않기 때문이다. 그러나 2-계층 클러스터링의 경우는 컨텐츠 정보와 링크 정보를 모두 사용하기 때문에 방문자 비율과 방문률이 증가함에 따



(그림 6) 규칙평가도 (90%, 80%)

라 규칙평가도 값이 커지는 것을 알 수 있다. 특히 방문자 비율 보다 방문률이 더욱 규칙평가도에 영향을 주고 있다. 이는 웹 로그를 분석하여 얻는 사용자의 페이지 방문정보가 행동규칙을 설정하는데 사용자의 컨텐츠 정보보다 더 유익하다는 것을 보여준다.

5. 결 론

일반적으로 웹 사용자들의 정보량은 부족하고 불확실성을 갖고 있어 사용자 데이터만을 갖고 사용자 그룹을 형성하고 웹 로그 분석을 통해 행동규칙을 추출하는 방식은 정확도가 떨어질 수 있다. 2-계층 클러스터링은 서로 다른 계층의 상호작용에 의한 잇점을 꾀하기 때문에 웹 사이트에서 유사 사용자의 그룹형성과 행동패턴을 추출하는데 효과적이다.

본 논문에서는 웹 사용자 그룹을 형성하고 이들의 행동패턴을 효율적으로 추출하기 위해 사용자 데이터 계층과 페이지 데이터 계층 간의 연관관계를 고려한 2-계층 클러스터링을 실시하였다. 그리고 각 계층에 서로 다른 클러스터링 알고리즘을 적용하고 분석함으로써 특정 알고리즘에 관계 없이 2-계층 클러스터링이 단일계층 클러스터링 보다 우수하다는 것을 알 수 있었다. 또한 성질이 다른 데이터 계층에 서로 다른 알고리즘을 적용하는 경우 좀더 좋은 성능을 가질 수 있음을 보여주었다.

그리고 본 논문에서는 생성된 사용자 클러스터로부터 사용자의 행동규칙을 추출하는 방법을 제시하였고 추출된 행동규칙이 내재된 행동패턴을 얼마나 잘 나타내는지를 C4.5와의 비교를 통해 살펴보았다. C4.5는 같은 행동패턴을 가지는 사용자들의 컨텐츠 정보가 유사할 때 좋은 결과를 보여주었지만 컨텐츠 정보가 어느 정도 다른 경우에는 나쁜 결과를 나타내었다. 2-계층 클러스터링은 사용자의 컨텐츠 정보 뿐만 아니라 페이지 데이터와의 연관성 정보를 함께 이용하기 때문에 컨텐츠 정보가 어느 정도 다르더라도 연관정보를 같이 사용하므로 좋은 결과를 보여주었다. 그리고 클러스터의 정확도가 높을수록 행동규칙의 유용성이 높음을 실험을 통해 알 수 있었다.

참 고 문 헌

[1] Vincent S. Tseng, Jeng-Chuan Chang, and Kawuu W. Lin, “Electronic commerce technologies(ECT): Mining and Prediction of Temporal Navigation patterns for personalized services in e-commerce,” Proceedings of the 2006 ACM symposium on Applied computing, pp.867-871, 2006.

[2] B. Mobasher, H. Dai, T. Luo, and M. Nakagawa, “Web Data Mining: Effective personalization based on association rule discovery from web usage data,” Proceedings of the 3rd international workshop on Web information and data management, pp.9-15, 2001.

[3] R. Kosala, and H. Blockeel, “Web Mining Research: A Survey,” ACM SIGKDD Explorations, Vol.2 No.1, pp.1-15, 2000.

[4] Y. Xie, and V. V. Phoha, “Web user clustering from access log using belief function,” Proceedings of the 1st international conference on Knowledge capture, pp.202-208, 2001.

[5] M. Gery, and H. Haddad, “Web clustering and usage mining: Evaluation of Web Usage Mining Approacvhes for User’s Next Request Prediction,” Proceedings of the 5th ACM international workshop on Web information and data management, pp.74-81, 2003.

[6] J. W. Hwang, D. H. Song, and C. H. Lee, “Performance Analysis of 2-tier Clustering,” 2006 International Conference Hybrid Information Technology, vol.2, pp.542-547, 2006

[7] J. R. Quinlan, “C4.5: Programs for Machine Learning,” Morgan Kaufmann Publishers, 1993.

[8] Brian S Everitt, “Hierarchical Clustering Techniques”, Cluster Analysis, pp.55-90, 1993

[9] H.J. Zeng, Z. Checn, and W. Y. Ma, “A Unified Framework for Clustering Heterogeneous Web Objects,” Proceedings of the 3rd International Conference on Web Information Systems Engineering, pp.161-172, 2002.

[10] J. Srivastava, R. Cooley, M. Deshpande, and P. N. Tan, “Web Usage Mining: Discovery and Applications of Usage Patterns from Web Data,” ACM SIGKDD Explorations, Vol.1 No.2, pp.12-23, 2000.

[11] C. Stanfill, and D. Waltz, “Toward memory-based reasoning,” Communications of the ACM, Vol.29, No.12, 1986.
[12] M. Steinbach, G. Karypis, and V. Kumar, “A Comparison of document clustering techniques,” KDD Workshop on Text Mining, 2000.
[13] Open Directory Project, http://dmoz.org/, 2007.

황 준 원

e-mail : jwhwang@konkuk.ac.kr
1998년 건국대학교 컴퓨터공학과(학사)
2000년 건국대학교 컴퓨터공학과 (공학석사)
2002년~현 재 건국대학교 컴퓨터공학과 박사과정수료
관심분야 : eCRM, 웹 마이닝, 데이터마이닝 등

송 두 헌

e-mail : dsong@ysc.ac.kr
1981년 서울대학교 계산통계학과(학사)
1983년 한국과학기술원 전산학과(석사)
1994년 캘리포니아대학교 전산학과 박사과정수료
1983년~1986년 KIST 연구원
1997년~현 재 용인송담대학 컴퓨터게임정보과 교수
관심분야 : 기계학습, 데이터마이닝, 데이터베이스, 보안 등

이 창 훈

e-mail : chlee@konkuk.ac.kr
1980년 연세대학교 수학과(학사)
1977년 한국과학기술원 전산학과(석사)
1993년 한국과학기술원 전산학과(박사)
1996년~2000년 건국대학교 서울캠퍼스 정보통신원 원장
2000년~2002년 건국대학교 정보통신대학원 원장
2001년~2002년 건국대학교 정보통신대학 학장
1980년~현 재 건국대학교 컴퓨터공학과 교수
관심분야 : 지능시스템, 운영체제, 보안, 전자상거래 등