Journal of The Korea Society of Computer and Information
Vol. 18, No. 9, September 2013
www.ksci.re.kr
http://dx.doi.org/10.9708/jksci.2013.18.9.177

유용한 연관 규칙 추출을 위한 시각적 탐색 기반 접근법

김 준 우*, 강 현 경**

Visual Exploration based Approach for Extracting the Interesting Association Rules

Jun-Woo Kim*, Hyun-Kyung Kang**

요 약

연관 규칙 탐사는 다양한 분야에서 널리 쓰이는 데이터 마이닝 기법으로 트랜잭션 데이터에 포함된 이산적인 항목들 간의 인과관계를 추출하는데 목적을 둔다. 하지만 분석자들은 때로 방대한 양의 데이터에서 추출된 많은 연관 규칙들을 해석하고 활용하는데 곤란을 겪기도 한다. 이러한 문제점을 해결하기 위하여 본 논문에서는 주어진 트랜잭션 데이터에서 유용한 연관 규칙을 탐색하기 위한 새로운 방법인 HTM 접근법을 제안하고자 한다. HTM 접근법은 크게 계층 군집, 테이블 뷰 및 모자이크 플롯의 세 가지 단계로 구성되며, 각 단계는 분석자들에게 적절한 시각적 표현을 제공한다. 예시를 위해 본 논문에서는 상기 접근법을 건강 검진 결과 데이터 분석에 적용하였으며, 실험 결과 HTM 접근법을 통해 분석자들은 유용한 규칙들을 보다 효과적으로 탐색할 수 있을 것으로 기대된다.

▸ Keywords : 연관 규칙 탐사, 트랜잭션 데이터, 사전 처리, 시각화, 계층 군집

Abstract

Association rule mining is a popular data mining technique with a wide range of application domains, and aims to extract the cause-and-effect relations between the discrete items included in transaction data. However, analysts sometimes have trouble in interpreting and using the plethora of association rules extracted from a large amount of data. To address this problem, this paper aims to propose a novel approach called HTM for extracting the interesting association rules from given transaction data. The HTM approach consists of three main steps, hierarchical clustering, table-view, and mosaic plot, and each step provides the analysts with appropriate visual representation. For illustration, we applied our approach for analyzing the mass health

•제1저자 : 김준우 •교신저자 : 강현경
•투고일 : 2013. 7. 5, 심사일 : 2013. 7. 22, 게재확정일 : 2013. 8. 2.
* 동아대학교 산업경영공학과(Dept. of Industrial and Management Systems Engineering, Dong-A University)
* 신라대학교 치위생학과(Division of Dental Hygiene, Silla University)

examination data, and the result of this experiment reveals that the HTM approach help the analysts to find the interesting association rules in more effective way.

▸ Keywords : Association rule mining, Transaction data, Pre-processing, Visualization, Hierarchical clustering

I. 서 론

연관 규칙(association rule) 탐사는 널리 사용되는 데이터 마이닝 기법 중의 하나로, 트랜잭션 데이터에 존재하는 이산적인 항목 집합(itemset) 간의 인과관계를 분석하여 '$X \rightarrow Y$' 형태의 규칙을 산출하게 된다. 이러한 연관 규칙에서 X, Y는 각각 한 개 이상의 항목으로 구성된 항목 집합이며, X는 전항(antecedent), Y는 후항(consequent)으로 지칭하고, 이 연관 규칙은 X를 포함하는 트랜잭션은 Y까지 포함하는 경향이 있음을 의미한다[1, 2]. 일반적으로 연관 규칙 탐사에는 Apriori 알고리즘 및 그 변종들로 대표되는 자동화된 알고리즘이 사용되며, 이들은 분석자가 설정한 지지도 하한, 신뢰도 또는 여러 가지 유용성 지표(interestingness measure)의 하한 등과 같은 파라미터를 만족하는 모든 연관 규칙을 산출한다[3, 4].

그러나 항목의 종류가 많은 대용량 데이터를 분석하는 경우에는 분석 결과로 산출되는 연관 규칙의 개수도 함께 늘어나기 때문에 종래와 같이 텍스트 형태로 추출된 연관 규칙을 나열해서는 분석자가 이를 해석하고 유용한 연관 규칙을 찾아 활용하는데 많은 어려움을 겪게 된다[4, 5]. 이에 따라 분석자가 유용한 연관 규칙을 효과적으로 찾을 수 있게 하는 방법에 대한 연구가 지속적으로 이루어져 왔으며, 대표적으로는 연관 규칙 탐사 과정에서 적절한 시각화(visualization) 기능을 제공하여 사람의 시각적 인지 능력 및 정보 처리 능력을 활용하고자 하는 것을 들 수 있다. 일반적으로 시각화의 목적은 연관 규칙 탐사 과정 수행 중, 분석자에게 여러 가지 표나 그래프 등을 제공하고 분석자는 이를 보고 탐사 대상을 좁히거나 좀 더 상세한 분석을 진행하도록 하는데 있다. 그러나, 기존의 연구들에서는 시각화 도구 자체가 직관적이지 않아 해석이 어려운 경우도 있고, 연관 규칙의 수가 많아지는 경우에는 이러한 시각화 역시 복잡한 형태의 정보를 제공하게 되는 경우가 많다[6].

이러한 맥락에서 본 논문에서는 분석자가 좀 더 직관적인 절차를 통해 유용한 연관 규칙을 찾는데 사용할 수 있는 방법인 HTM(Hierarchy-Table-Mosaic) 접근법을 제안하고자 한다. HTM은 크게 계층 군집, 테이블 뷰, 모자이크 플롯의 3가지 단계로 구성되며, 1단계인 계층 군집에서는 트랜잭션 데이터에 존재하는 단일 항목들에 대한 계층 구조를 추출하여 보여주며, 이 때 항목 간의 연관성으로는 지지도나 신뢰도 등의 지표를 사용할 수 있다. 이러한 단일 항목 간 계층 구조는 분석자가 서로 연관성이 높은 항목들의 그룹을 직관적으로 파악할 수 있게 해 준다. 이를 통해 분석자가 분석 대상 항목을 선택하면, 선택된 항목들로 범위를 좁혀 연관 규칙 탐사를 실시하며 이 과정에서는 Apriori 알고리즘 등을 사용할 수 있다. 선택된 항목들에 대한 연관 규칙이 추출되면, 2단계로 이들에 대해 전통적인 테이블 형태의 요약 정보를 제공하며, 이를 통해 분석자는 추출된 연관 규칙의 종류 및 세부 정보들을 확인할 수 있다. 사용자는 이 때 보여지는 연관 규칙 중 특별히 관심이 가는 규칙 1개를 선택할 수 있으며, 선택된 연관 규칙에 대해서는 마지막 3단계에서는 모자이크 플롯을 통한 관찰을 실시하여 보다 세부적인 정보를 확인할 수 있다. HTM은 데이터 분석에 익숙하지 않은 분석자라 할지라도 비교적 직관적인 요약 절차인 1단계와 선택된 항목에 대한 기본 분석 절차인 2단계를 수행할 수 있으며, 3단계에서는 보다 심도 깊은 분석 결과를 제공한다. 이러한 구성은 Schneiderman[7]이 제안한 시각적 분석의 대원칙인 'overview first, zoom and filter, then details-on-demand'와도 부합하며, 다양한 분야, 다양한 목적의 분석에 활용될 수 있을 것으로 기대된다.

본 논문의 2장에서는 연관 규칙 시각화 및 제안하는 HTM 세부 요소들과 관련된 기존 문헌을 소개하고, 3장에서는 HTM 접근법의 분석 과정 및 절차를 설명한다. 이어 4장에서는 건강 검진 결과 데이터에 HTM을 적용한 사례 및 결과를 소개하며, 끝으로 5장에서 결론 및 추후 과제를 제시한다.

II. 관련 연구

1. 연관 규칙 시각화

연관 규칙 탐사는 그 개념이 비교적 단순하나, 분석 결과 방대한 양의 연관 규칙이 추출되는 경우에는 이들을 적절히 해석하고 유용한 규칙들을 선택하는 것이 까다롭기 때문에 이를 보완하기 위한 다양한 연구가 수행되어져 왔다[8].

분석자의 이해를 돕기 위한 가장 기본적인 방법은 추출된 연관 규칙들을 테이블 형태로 정리하여 연관 규칙 하나를 한 행에 기록하고 각 열에는 지지도, 신뢰도, Lift 등의 다양한 평가 지표들을 나열하는 방법이 있고, 지금도 다양한 데이터 마이닝 소프트웨어에서 이를 채용하고 있다[6, 9, 10]. 하지만 규칙의 개수가 많아질 경우, 테이블의 내용을 분석자가 일일이 관찰하기 어려워지기 때문에 적절한 그림이나 그래프를 활용하는 경우가 많다. 비교적 단순한 방법으로는 두 개의 평가 지표를 축으로 하여 각 연관 규칙들을 산점도 형식으로 나타내는 방법[11, 12]이 있었으며, 이는 분석자가 평가 지표가 높은 규칙을 찾아보도록 하는데 목적을 두었으나 항목들 간의 다 대 다 관계를 묘사하는 연관 규칙의 특성 상, 각 규칙이 포함하는 항목들을 살펴보는데 한계가 존재한다.

좀 더 발전된 형태의 시각화 방법으로는 각 항목을 수평선으로 나타내고 각 규칙이 포함하는 항목들만을 연결하는 형태의 bar 차트[13], 항목들을 포함하는 세로축을 여러 개 배치하고 빈발 항목 집합이나 연관 규칙이 포함하고 있는 항목들을 연결하여 나타내는 평행좌표(parallel coordinates)[14, 15, 16, 17], 항목집합을 노드로 표현하고 전항과 후항 사이를 아크로 연결하는 형태의 그래프[6, 10, 15], 격자의 좌측과 상단에 각각 규칙의 전항, 후항을 배치하고 격자의 각 셀들에 해당하는 연관 규칙을 표현하는 방법[6, 9, 18] 등이 있다.

이러한 방법들은 모두 추출된 연관 규칙들을 도식적으로 보여준다는 공통점이 있으나, 연관 규칙의 개수가 많아질 경우에는 그림이 복잡해지거나 선끼리의 엉킴(tangle) 등이 발생하여 분석자가 해석하기 어려워질 수 있다[6]. 이에 따라 항목들을 적절히 그룹을 지어 시각화할 내용을 줄이는 방법들이 제안되기도 하였으며, 이러한 목적으로는 쌍대척도법(dual scaling)을 통한 연관성 높은 항목의 추출[5], k-means 군집 분석을 통한 연관 규칙의 군집화[20], 추출된 연관 규칙을 구성하는 전항 및 후항들을 계층적으로 관찰하는 방법[4, 14] 등이 있다.

나아가, 최근에는 효과적으로 탐색 대상을 줄여나가기 위해서는 연관 규칙 탐사와 시각화를 각기 한 번씩만 수행하기보다 시각화 결과를 보고 사용자가 분석 과정에 적극적으로 개입하는 것이 보다 바람직한 것으로 받아들여진다[8]. 하지만 앞에서 언급한 시각화 방법들은 데이터 분석과 관련된 지식이 없는 분석자는 활용하기 어려운 경우가 많고, 최근에는 다양한 분야에서 대용량의 데이터가 수집되는 추세를 감안할 때, 보다 직관적이고 사용하기 쉬운 분석 방법이 필요하다.

2. 계층 군집과 모자이크 플롯

군집 분석(clustering analysis)은 본디 데이터를 구성하는 레코드들을 여러 개의 군집으로 나누는 것을 의미하며, 유사한 레코드끼리는 같은 군집에, 상이한 레코드는 서로 다른 군집에 소속시키는 것을 목표로 한다[3]. 계층 군집(hierarchical clustering)은 널리 사용되는 군집 분석 방법 중의 하나로, 레코드로 구성된 중첩된 군집(nested cluster)들을 형성하여 이들에 대한 계통도(dendrogram)가 얻어진다는 특징이 있다. 계층 군집 분석 알고리즘은 크게 모든 레코드를 각각 하나의 군집으로 본 상태에서 한 번에 두 개 군집을 합쳐나가는 것을 반복하는 병합형(agglomerative) 계층 군집[21]과 모든 레코드를 포함하는 한 개의 군집을 나누는 것을 반복하는 분할형(divisive) 계층 군집[22]으로 분류되며, 이들을 통해 얻어진 계통도는 매우 직관적으로 해석이 가능하다는 특징이 있다.

한편, 트랜잭션 데이터의 경우에는 계층 군집을 통해 종래와 같이 레코드들의 군집을 생성하는 대신, 항목들의 군집을 생성하기 위한 분석이 가능하며, 이러한 예로는 고객들의 상품 별 선호도 조사 데이터를 이용하여 같은 고객이 선호할 가능성이 큰 상품들이 가까이 배치되는 상품 계통도[23]나 IT 관련 기사들을 분석하여 같은 기사에 등장하는 키워드들을 가까이 배치하는 키워드 계통도[24]를 얻은 연구들이 존재한다. 본 논문에서는 이와 같이 트랜잭션 데이터를 구성하는 항목들에 대한 계통도를 생성하는 계층 군집을 통하여 분석자가 연관 규칙 탐사를 실시하기 전에 미리 대상 데이터에 대한 직관적인 이해를 하고, 분석 대상 항목들을 선별할 수 있도록 하고자 한다. 연관 규칙 탐사와 관련해서는 추출된 규칙들에 대하여 계층 군집을 적용, 연관 규칙들을 나열하는 사례[14]가 있었으나, 본 논문에서는 연관 규칙 탐사 이전에 사전 분석 목적으로 계층 군집을 이용한다는 점에서 차이가 있다.

실제 계층 군집을 실시할 때는 두 항목 간 유사도 및 두 군집간의 유사도를 측정하는 방법이 중요하며, 본 논문에서는 연관 규칙 탐사의 특성을 반영할 수 있도록 두 항목의 합집합

에 대한 지지도나 항목 간 신뢰도 등을 유사도 척도로 사용하고자 하며, 두 군집 간 유사도의 경우에는 항목 간 유사도 척도를 기반으로 전통적인 단일 링크(single link), 완전 링크(complete link) 등의 방법을 사용할 수 있다.

모자이크 플롯(mosaic plot) 및 그 변종[25]은 본디 앞 절에서 소개한 방법들과 함께 연관 규칙 탐사와 관련된 시각화 기술로 알려져 있다. 그러나, 추출된 많은 연관 규칙들의 시각적 표현 및 요약을 수행하는 다른 시각화 기술들과 달리 모자이크 플롯은 한 개의 연관 규칙을 대상으로 한다는 차이점이 있으며, 구체적으로는 연관 규칙에 대한 분할표(contingency table)을 도식적으로 나타내는데 사용된다.

분석자는 모자이크 플롯을 통해 선택된 연관 규칙의 유용성을 평가하거나, 지지도 임계치를 만족하지 못하여 연관 규칙 탐사에서 추출되지는 못하고 분석 결과에서는 누락되었으나 실제로는 의미 있는 잠재적인 연관 규칙을 탐색해볼 수 있다. 즉, 모자이크 플롯은 연관 규칙 탐사 이후 보다 세부적인 관찰을 수행하기 위한 시각화 방법이며, HTM 접근법에서는 가장 마지막 단계에 수행하는 절차이기도 하다.

III. HTM 분석 절차

그림 1은 유용한 연관 규칙의 효과적인 탐색을 위하여 본 논문에서 제안하는 HTM 접근법의 전체 절차를 요약하여 보여준다. HTM 접근법은 주어진 트랜잭션 데이터에 계층 군집, 테이블 뷰, 모자이크 플롯의 3가지 단계를 순서대로 적용하여 유용한 연관 규칙을 체계적으로 탐색하도록 구성되어 있다.

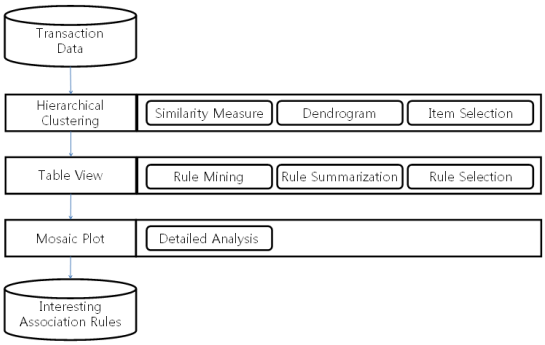

그림 1. HTM 전체 절차
Fig. 1. Overall HTM Procedure

첫 번째 단계에서 분석자는 주어진 트랜잭션 데이터 T의 특성을 전체적으로 관찰한 후, 실제 연관 규칙 탐사에 사용할 항목들을 선별하게 된다. 이를 위해서는 적절한 시각적 정보가 분석자에게 제공되어야 하며, 이 시각적 정보는 일반적인 분석자들이 가급적 직관적으로 이해할 수 있어야 한다. 본 논문에서는 계층 군집 분석을 통해 생성되는 항목 계통도를 이러한 목적으로 사용하고자 하며, 예를 들어 A, B, …, G의 총 7개 항목을 포함하는 트랜잭션 데이터의 경우, 그림 2와 같은 항목 계통도가 생성될 수 있다. 이러한 계통도에서 가까이 배치된 항목들은 서로 관련이 높은 항목들임을 의미하므로, 분석자는 계통도를 보고 특정 부분의 항목들만을 선별하여 연관 규칙 탐사를 진행할 수 있으며, 예를 들어, 그림 2에서는 서로 인접해 있는 A, B, C 세 개 항목, 또는 D, E두 개 항목을 선별하는 식으로 활용이 가능하다.

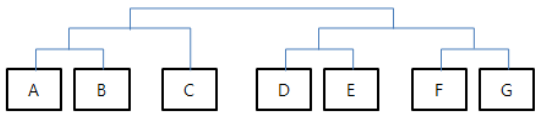

그림 2. 항목 계통도 예
Fig. 2. An Example of Item Dendrogram

본 논문에서는 계층 군집 분석 방법 중, 병합형 계층 군집 알고리즘을 사용하였으며, 분할형 계층 군집 알고리즘도 활용이 가능하다. 단, 계층 군집 분석을 실행하기 위해서는 두 개 항목 간 유사도 및 두 개 군집(항목 집합) 간 유사도 척도를 적절히 정의하는 것이 필요하다.

트랜잭션 데이터에 포함된 두 개 항목 A, B간 유사도 $sim(A, B)$를 측정하는 방법으로는 먼저, 트랜잭션 데이터에 포함된 항목집합 간의 거리(비유사도) 척도인 자카드 거리(Jaccard distance)[26]를 변형하여 (1)과 같이 측정할 수 있고, 단, $|A \cap B|$, $|A \cup B|$는 각각 전체 트랜잭션 중, 두 항목을 모두 포함한 트랜잭션의 개수 및 둘 중 적어도 한 개를 포함하는 트랜잭션의 개수를 의미한다.

$$sim(A, B) = \frac{|A \cap B|}{|A \cup B|} \quad (1)$$

(1)의 유사도는 두 항목 중 하나 이상을 가진 트랜잭션들 중, 둘 모두를 포함하고 있는 것들의 비율로, 같은 트랜잭션에 빈번하게 함께 등장하는 두 항목일수록 높은 값이 산출된다. 그러나 연관 규칙 탐사에서는 지지도, 신뢰도와 같은 유용성 지표를이 기준이 되므로, 좋은 연관 규칙이 많이 산출되는 항목 집합들의 집합을 계통도에서 선별하기 위해서는 이러한 지표들을 반영하는 것이 바람직할 수 있다. 예를 들어 (2)는 두 개 항목 간 유사도로 두 항목의 공기 정보

(co-occurrence)를 사용하기 위하여 두 항목의 합집합 지지도를 사용하는 척도이다. 단, $|T|$는 주어진 트랜잭션에 포함된 모든 레코드 개수를 의미한다.

$$sim(A, B) = support(A \cup B) = \frac{|A \cup B|}{|T|} \qquad (2)$$

한편, 실제 연관 규칙 탐사에 있어서는 지지도는 다소 낮더라도 신뢰도나 Lift 와 같은 상관관계 또는 조건부 확률 지표의 값이 높은 규칙들이 선호될 수 있는데, 이러한 지표들은 앞의 (1), (2)와는 달리 두 항목 간에 비대칭적인 (asymmetric) 값이 산출되므로, 두 항목을 각각 규칙 전항, 후항으로 한 번씩 사용하여 산출된 값 중 큰 것을 사용하여 유사도 지표를 정의할 수 있다. 예를 들어, (3)은 두 항목 A, B간 유사도로 규칙 '$A \rightarrow B$'와 규칙 '$B \rightarrow A$'의 신뢰도 중 큰 값을 사용하고 있다. 단, $con(A \rightarrow B)$는 규칙 '$A \rightarrow B$'의 신뢰도를 의미하며 (4)와 같이 산출된다. 참고로 신뢰도는 연관 규칙에 대한 전통적인 유용성 지표이나, 규칙 전항과 후항 간 상관관계를 반영하는데 한계가 있어 Lift 등의 발전된 지표를 사용하기도 하며, 이러한 지표들을 사용하는 경우에도 (3)과 같이 전항, 후항을 바꾸어가며 산출한 값의 최대값을 활용할 수 있다.

$$sim(A, B) = \max(con(A \rightarrow B), con(B \rightarrow A)) \qquad (3)$$

$$con(A \rightarrow B) = \frac{|A \cup B|}{|A|} \qquad (4)$$

병합형 계층 군집 분석이 진행되는 과정에서는 개별 항목들이 병합되어 항목집합을 형성하게 되므로, 이러한 항목집합 간 유사도를 산출하는 방법이 필요하다. 이는 전통적인 계층 군집에서 사용하는 그룹 간 비교 방법을 통해 수행할 수 있으며, 두 그룹에 속해 있는 항목들을 서로 하나씩 비교하여 (1)~(3)과 같은 방법으로 항목 간 유사도를 산출한 후, 이 중 가장 유사한 것끼리의 유사도를 항목집합 간 유사도로 사용하는 단일 링크, 가장 상이한 것끼리의 유사도를 사용하는 완전 링크, 유사도 값들의 평균을 사용하는 그룹 평균 등을 사용할 수 있다.

참고로 트랜잭션 데이터의 특성 상, 위에서 언급한 단일 항목 간 유사도를 기준으로 하는 그룹 간 비교 방법 외에도 다양한 지표를 생각해볼 수 있다. 예를 들어, (5)와 같이 자카드 거리에 기반한 항목 집합 X_i, X_j 유사도, 또는 (6)과 같은 두 개 항목 집합 전체 지지도 등이 가능하다. 그러나 이러한 지표들의 경우, 첫째, 사용 시 분석 과정이 오래 걸릴 수 있고, 둘째, 비교 대상 항목집합들에 포함된 항목의 개수가 늘어날수록 산출된 유사도 값의 수준이 극히 떨어지는 경향이 있어, 본 논문에서는 (1)~(3)의 지표들을 위주로 사용하였다.

$$sim(X_i, X_j) = \frac{|X_i \cap X_j|}{|X_i \cup X_j|} \qquad (5)$$

$$sim(X_i, X_j) = support(X_i \cup X_j) \qquad (6)$$

유사도 지표 및 그룹 간 비교 방법을 선택하여 항목 계통도를 조회한 뒤, 분석자는 계통도 상에서 인접한 곳에 배치된 항목들을 선택하는 것으로 첫 번째 단계인 계층 군집을 마치게 된다.

두 번째 단계인 테이블 뷰에서는 선택된 항목들을 대상으로 연관 규칙 탐사를 실시하게 되고, 이 때 apriori 알고리즘 등을 사용할 수 있다. 연관 규칙 탐사 알고리즘의 경우 본 논문의 범위가 아니므로 자세한 내용은 생략한다.

추출된 연관 규칙들은 표 1과 같이 테이블 형식의 시각화를 통한 관찰이 분석자의 이해를 도울 것으로 사용된다. 이와 같은 테이블 뷰는 전통적으로 다양한 데이터 마이닝 소프트웨어에서 지원되어왔고, 연관 규칙에 해당하는 행과 유용성 지표에 해당하는 열들로 구성된다는 특징이 있다. 이 방법은 단순하면서도 연관 규칙 탐사 결과를 요약시켜 보여준다는 특징이 있고, 특정 유용성 지표 기준으로 연관 규칙을 정렬시켜 보기 유리하다는 장점이 있으나, 연관 규칙 개수가 많아질 경우, 알아보기 어렵다는 한계도 있다. 하지만 본 논문에서 제안한 바와 같이 계층 군집 분석 단계를 거쳐 연관 규칙 분석 대상 항목을 선별한 경우에는 산출된 연관 규칙 개수가 상대적으로 적게 유지되면서도 관련 있는 항목들 간의 유용한 규칙들이 잘 산출될 것으로 기대된다.

표 1. 연관 규칙들에 대한 테이블 뷰 예
Table 1. An Example of the Table View for Associtaion Rules

연관 규칙	지지도	신뢰도	Lift	…
$\{B\} \rightarrow \{C\}$	0.500	0.800	1.280	…
$\{A, B\} \rightarrow \{C\}$	0.375	0.667	1.200	…
…	…	…	…	…

여기까지 설명한 HTM 접근법의 첫 번째, 두 번째 단계만을 가지고도 유용한 연관 규칙 탐사라는 목적을 어느 정도 달

성할 수 있으나, 보다 세부적인 관찰이 필요한 경우, 분석자는 최종적으로 테이블 뷰에 포함된 연관 규칙 하나를 선택하여 모자이크 플롯을 통한 관찰을 수행할 수 있다.

모자이크 플롯은 특정 연관 규칙 한 개에 대한 분할표를 도식화하여 관찰하는 도구로서, 사전에 정의된 지지도 및 신뢰도 하한을 넘지 못하여 연관 규칙 탐사 알고리즘에 의해 추출되지는 못하였으나 실제로는 높은 상관관계를 의미하는 연관 규칙들을 찾아내는 등의 목적으로 활용이 가능하며, 본 논문에서는 모자이크 플롯 중 가장 단순한 형태인 2차원 모자이크 플롯을 고려한다.

예를 들어, 표 2와 같이 두 개의 항목 A, B에 대한 분할표를 생각해보자. 지지도 하한 0.25, 신뢰도 하한 0.6을 가정할 때, 총 500개의 트랜잭션 중, '$A \rightarrow B$'는 지지도 0.5, 신뢰도 0.658로 연관 규칙으로 추출된다. 반면, $Not\ A \rightarrow Not\ B$의 경우에는 신뢰도 0.75로 상당히 가치 있는 연관 규칙임에도 불구하고, 지지도 0.18로 탐사 결과에서 제외된다. 이러한 연관 규칙들은 분석자가 추출된 개별 연관 규칙들에 대한 상세한 관찰을 통해 파악해야 한다.

표 2. 분할표 예
Table 2. An Example of the Contingency Table

	B	Not B
A	250	130
Not A	30	90

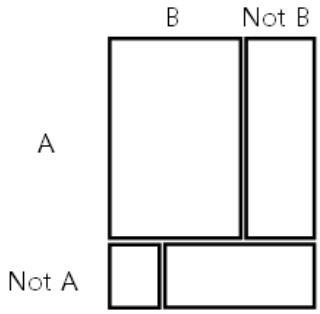

그림 3. 모자이크 플롯 예
Fig. 3. An Example of Mosaic Plot

모자이크 플롯은 표 2와 같은 분할표의 각 셀들을 그림 3에서와 같이 크기가 서로 다른 직사각형으로 나타내며, 좌측에는 규칙 전항 항목, 상단에 규칙 후항 항목을 배치한 다음, 각 행의 높이를 분할표 각 행의 도수 총합에 비례하여 설정하고, 각 행은 다시 분할표의 행을 구성하는 셀들의 도수에 비례하는 너비를 갖는 셀로 분할시켜둔다.

모자이크 플롯을 해석할 때는 한 번에 한 개 행을 관찰하면서, 너비가 큰 셀이 존재하는 경우 유용한 규칙이 발생할 가능성이 높은 것으로 볼 수 있다. 특히, 행의 높이는 짧으나 너비가 큰 셀이 존재하는 그림 3의 두 번째 행과 같은 경우, 유용성 지표가 높은 연관 규칙이 내포되어 있을 수 있다. 나아가, 표 2와 그림 3에서처럼 규칙 전항과 후항이 단일 항목으로 구성되지 않고, 2개 이상의 원소를 갖는 항목집합이 있는 경우에는 한 번에 전항 항목과 후항 항목을 하나씩 비교하는 작업을 여러 번 수행하여 상세한 분석을 실행할 수 있다.

IV. 건강 검진 데이터 분석 결과

본 논문에서는 HTM 접근법을 단체 건강 검진 결과 데이터에 적용하여 그 유용성을 검증해보고자 한다. 건강 검진 데이터에는 인구통계학적 문항과 키나 체중 등과 같은 연속 변수 문항도 존재하지만, 특정 질환 유무에 대한 검사 항목이나 생활 습관에 대한 설문 항목들의 대다수가 이진 문항으로 구성되어 있어, 이러한 항목들만 분리시킬 경우, 트랜잭션 데이터를 형성하게 된다. 나아가, 특정 수검자에 대하여 앞으로 발병 확률이 높은 질환 등을 조기에 파악하는데 연관 규칙 탐사를 활용할 수 있으나, 건강 검진을 담당하는 의료진들이 데이터 분석에 익숙하지 않은 경우가 많으므로, 본 논문에서 제안하는 것과 같은 체계적인 분석 방법의 활용이 중요할 것으로 생각된다. 참고로 본 논문에서는 2011년 부산 소재 D고등학교 1학년 278명에 대한 건강 검진 데이터에서 치위생 관련 문항들에 대한 분석을 실시하였다. 표 3은 이에 해당하는 19개 문항들의 목록을 보여주며, 이들은 모두 '예/아니오'로 응답하게 되어 있다.

표 3. 치위생 검진 문항
Table 3. The Dental Examination Variables

번호	문항
1	우식 치아 유무
2	유식 위험 치아 유무
3	부정교합 유무
4	구강 위생 상태 불량 여부
5	치주 질환 유무
6	구내염 및 연조직 질환 유무
7	깨지거나 부러진 치아 유무
8	차갑거나 뜨거운 식음료 섭취 시 통증 유무
9	치아가 욱신거리거나 이픔
10	잇몸이 아프거나 피가 남

11	불쾌한 입 냄새가 남
12	아침 식사 전 이를 잘 닦는지
13	아침 식사 후 이를 잘 닦는지
14	점심 식사 후 이를 잘 닦는지
15	저녁 식사 후 이를 잘 닦는지
16	잠자기 전 이를 잘 닦는지
17	간식 섭취 후 이를 잘 닦는지
18	단 음식이나 청량음료를 즐기는지
19	불소 함유된 치약 사용하는지

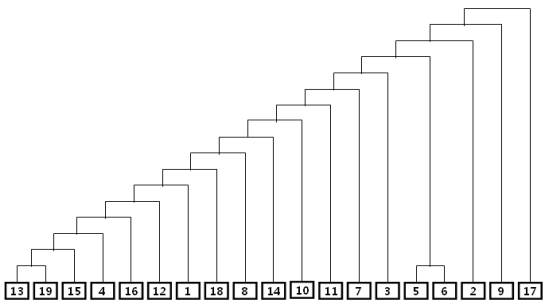

그림 4. 지지도 단일 링크에 의한 계층 군집
Fig. 4. Hierarchical Clustering based on Single Link of Support

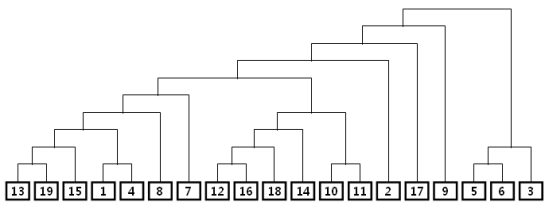

그림 5. 지지도 완전 링크에 의한 계층 군집
Fig. 5. Hierarchical Clustering based on Complete Link of Support

상기 데이터에 항목 간 유사도 지표로는 (2), (3)의 지지도 및 신뢰도를 사용하여 HTM 접근법의 첫 단계인 계층 군집을 실행한 결과가 각각 그림 4~그림 7에 요약되어 있고, 이들을 통해 다음 몇 가지 사항들을 관찰할 수 있다.

먼저, 트랜잭션 데이터에 대한 계층 군집 결과들을 보면 지정된 유사도 지표 및 그룹 간 비교 방법에 따라 각 항목들이 순차적으로 병합되는 계통도를 형성하고, 이러한 계통도는 분석자는 직관적으로 이해하기 쉬운 형태이다. 예를 들어, 그림 4에서는 가장 왼쪽의 항목 13, 19, 15, 4등이 비교적 인접하게 배치되었고, 그림 5에서는 {13, 19, 15}, {5, 6, 3} 등이 인접함을 볼 수 있다.

두 번째로는 유사도 지표 및 그룹 간 비교 방법에 따라 계층 군집을 통해 얻어지는 계통도의 모습은 상당히 상이함을 알 수 있다. 특히, 그림 4, 그림 6의 단일 링크와 그림 5, 그림7의 완전 링크의 주된 차이점은 계통도의 모양으로, 단일 링크를 사용하면 계통도의 깊이(depth)가 깊은 반면, 완전 링크를 사용하면 깊이가 상대적으로 얕다. 이는 단일 링크의 경우 양쪽 항목집합에서 가장 유사도 지표가 높은 한 쌍의 유사도만을 병합에 이용하는 반면, 완전 링크에서는 가장 유사도 지표가 낮은 한 쌍의 유사도만을 이용하기 때문이며, 나아가, 그룹 평균을 사용하는 경우에는 단일 링크와 완전 링크의 중간 정도 깊이를 갖는 계통도가 산출된다.

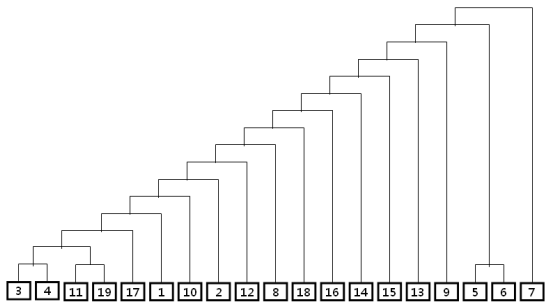

그림 6. 신뢰도 단일 링크에 의한 계층 군집
Fig. 6. Hierarchical Clustering based on Single Link of Confidence

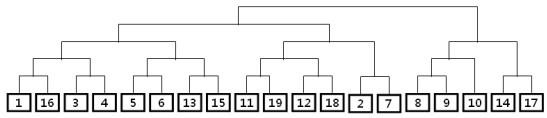

그림 7. 신뢰도 완전 링크에 의한 계층 군집
Fig. 7. Hierarchical Clustering based on Complete Link of Confidence

실제 유사도 지표와 그룹 간 비교 방법에 대한 선택은 연관 규칙 탐사 목적에 따라 달라질 수 있으며, 특정 항목 한 개를 후항으로 두는 클래스 연관 규칙(class association rule)을 탐사하는 경우에는 일반적으로 그림 4, 그림 6과 같이 단일 링크에 의해 생성된 깊은 계통도가 유리할 수 있다. 그 이유는 핵심적인 항목 몇 개가 다른 여러 개 항목들의 선택을 유발하는 경향이 있는 트랜잭션 데이터의 경우, 깊은 계통도에서 핵심 항목 및 인접한 항목들을 선택하여 이들과 관련된 연관 규칙을 집중적으로 탐사할 수 있기 때문이다. 예를 들어, 그림 4에서 항목 3, 4, 11, 19는 서로 인접하면서 다른 대부분의 항목들도 이들 중 한 개와 가장 큰 신뢰도 관계를 형성한다.

나아가, 이러한 깊은 트리는 일반적으로 핵심 항목들에 다른 항목들이 한 개씩 추가로 병합되는 형태가 만들어지는 경향이 있으나, 그림 4, 그림 6에서의 항목 5, 6처럼 핵심 항목들이 아니면서 서로 관련성이 높은 항목들이 나타날 수 있어, 이러한 정보들은 연관 규칙 탐사에서 놓칠 수 있는 규칙을 찾

아내는데 도움이 될 수 있다. 그럼에도 불구하고, 깊은 트리의 경우, 핵심적이지 않은 항목들 간의 그룹을 파악하는 데 한계가 있으며, 이러한 문제점은 완전 링크를 사용함으로서 어느 정도 해소된다. 예를 들어, 그림 5에서는 {13, 19, 15, 1, 4}, {12, 16, 18, 14}, {5, 6, 3} 등과 같이 서로 중복되지 않는 몇 개의 그룹들을 파악하여, 이러한 그룹들 중 하나를 선택하여 연관 규칙 탐사를 실시할 수 있다.

이러한 항목 계통도는 그 자체만으로도 어느 정도 데이터 분석에 도움이 된다. 예를 들어, 위 그림 4~그림 7에서는 항목 5, 6, 3이 서로 관련성이 있음이 파악되고, 이는 곧, 부정교합, 치주질환, 구내염이 서로 어느 정도 상관관계가 있음을 의미한다. 이번에는 이들에 대하여 좀 더 자세한 분석을 하는 상황을 가정해보자. HTM 접근법의 두 번째 단계는 선택된 항목들에 대한 연관 규칙 탐사를 실행하는 것으로 시작된다. 항목 5, 6, 3을 선택한 후, 지지도 하한으로 0.02, Lift 하한 1을 설정하여 연관 규칙 탐사를 실행한 후, 산출된 규칙들을 정리하면 표 4와 같다. 일반적으로 Lift는 1 미만일 경우 음의 상관관계, 1을 초과할수록 강한 양의 상관관계를 의미함을 생각해볼 때, 산출된 규칙들의 유용성 지표가 높음을 알 수 있다. 또한, 항목 5, 6, 3에 해당하는 부정교합, 치주질환, 구내염 3개 질환에만 초점을 맞춘 연관 규칙을 탐사하여 이러한 질환들 사이의 인과관계만을 집중적으로 관찰하는 데도 유리하다.

표 4. 연관 규칙들에 대한 테이블 뷰 예
Table 4. An Example of the Table View for Associtaion Rules

연관 규칙	지지도	신뢰도	Lift
{5}→{6}	0.09	0.79	8.42
{6}→{5}	0.09	1.00	8.42
{5}→{3, 6}	0.03	0.21	8.42
{3, 6}→{5}	0.03	1.00	8.42
{3, 5}→{6}	0.03	0.78	8.32
{6}→{3, 5}	0.03	0.27	8.32
{3}→{5}	0.03	0.24	2.00
{5}→{3}	0.03	0.27	2.00
{3}→{6}	0.03	0.18	1.97
{6}→{3}	0.03	0.27	1.97
{3}→{5, 6}	0.03	0.18	1.97
{5, 6}→{3}	0.03	0.27	1.97

반면, 동일한 지지도 하한과 Lift 하한을 사용하면서 19개 항목 전체에 연관 규칙 탐사를 적용한 경우에는 추출되는 연관 규칙의 수가 매우 늘어나기 때문에 이들을 활용하는 것이 어려우며, 특히 상기 3개 질환과 관련된 연관 규칙만을 필요로 하는 경우에도 이들을 식별하는데 불편이 따른다. 물론, 분석 대상 항목의 개수를 줄여 과다한 연관 규칙이 추출되는 문제를 어느 정도 줄일 수는 있으나, 무작위로 항목을 선택했을 경우에는 유용성 지표가 높은 규칙이 산출되지 않을 가능성이 있다. HTM 접근법에서 제안하는 것과 같이 계층 군집과 연관 규칙 탐사를 순차적으로 수행하면 이러한 문제점들을 해결하여 효과적인 연관 규칙 탐사가 가능하다.

만약 표 4에서와 같이 추출된 연관 규칙 중 특정 규칙에 대하여 좀 더 자세한 분석을 원할 경우에는 규칙 하나를 선택하여 모자이크 플롯을 통한 관찰을 실시할 수 있다. 예를 들어, 표 4의 첫 번째 규칙인 $\{5\}\rightarrow\{6\}$에 대한 교차표와 모자이크 플롯을 생성한 결과는 각각 표 5, 그림 8과 같다.

표 5. 항목 5, 6에 대한 분할표
Table 5. Contingency Table for Item 5 and 6

	6	Not 6
5	26	7
Not 5	0	245

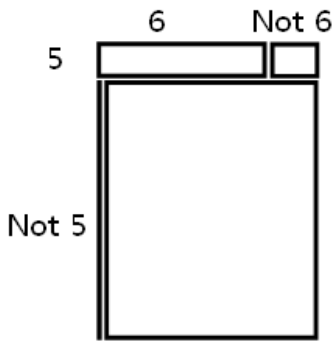

그림 8. 항목 5, 6에 대한 모자이크 플롯
Fig. 8. Mosaic Plot for Item 5 and 6

이들을 보면, 표 4에 나온 것처럼 5번 항목인 치주질환을 가지고 있는 경우 6번 항목인 구내염이 함께 있는 경우가 많다는 점이 드러난다. 아울러, 앞의 과정에서 명시적으로 드러나지는 않았지만 모자이크 플롯의 두 번째 행을 보았을 때, 치주질환이 없는 경우에는 구내염도 없는 것이 일반적이라는 부가적인 지식을 파악해낼 수 있다. 이러한 절차를 통해 HTM 접근법은 분석자가 데이터 전체에 대한 조망, 연관 규칙 탐사, 개별 규칙 세부 분석을 순서대로 수행할 수 있게 해준다.

V. 결론 및 추후 연구 과제

트랜잭션 데이터가 주어질 경우, 그 안에 숨겨진 연관 규

칙을 탐사하는 알고리즘은 많이 개발되어왔으나, 추출된 연관 규칙들을 해석하고, 그 중 유용한 것들을 선별하여 활용하는 것은 여전히 어려운 임무이다.

따라서 본 논문에서는 트랜잭션 데이터에 연관 규칙 탐사 알고리즘을 기계적으로 적용하기보다, 시각화에 기반한 순차적인 분석을 거쳐 유용한 연관 규칙을 선별하는 HTM 접근법을 제안하였다. 이 방법은 연관 규칙 탐사 알고리즘을 적용하기 전에 적절한 분석 항목을 선택하되, 임의로 선택하는 것이 아닌 항목 계통도라는 시각적 도구를 이용하고, 탐사 알고리즘 실행 결과 추출된 연관 규칙들에 대해서는 모자이크 플롯을 통해 세부적인 관찰을 시행하도록 구성되어 있다. 제안하는 분석 방법을 건강 검진 데이터에 적용해본 결과, 분석자가 시각적인 방법을 통해 선택한 항목 그룹에 대하여 유용성 지표가 높은 연관 규칙들이 추출되고, 개별 연관 규칙에 대한 모자이크 플롯을 통해 해당 규칙에 대한 보다 자세한 정보가 얻어지는 것을 확인할 수 있었다. 최근 대용량의 데이터가 다양한 분야에서 수집 및 활용되고 있는 상황에 비추어볼 때, 이렇게 직관적이고 체계적인 연관 규칙 탐사 방법은 향후 그 활용도가 매우 높을 것으로 기대된다. 반면, 제안하는 분석 방법에 대해서는 다음과 같은 추후 연구 주제들이 있어, 이들에 대한 보완이 필요할 것으로 생각된다.

첫 번째로는 항목 계통도를 생성할 때 사용하는 항목 간 유사도 및 항목집합 간 유사도를 측정하는 방법으로, 본 논문에서는 지지도나 신뢰도와 같은 비교적 단순한 지표만을 사용한 반면, 향후 다른 여러 가지 지표들을 개발하여 다양한 데이터에 적용해볼 필요가 있다. 아울러, 연관 규칙과 관련된 지표들은 항목 간 연관성이 서로 낮음을 의미할 수도 있는데, 예를 들어 0.5미만의 신뢰도 값이나 1.0미만의 Lift 값은 두 항목이 서로 음의 상관관계를 갖고 있음을 나타낸다. 이러한 경우에도 해당 지표들의 값을 이용하여 계통도를 생성할 것인지에 대해서는 추후 논의가 필요하다.

두 번째로는 분석자의 판단을 지원하는 방법에 관한 것으로, 항목 계통도가 다른 시각화 방법에 비해 직관적으로 이해하기 쉬운 형태이기는 하나, 항목의 개수가 많아지면 계통도 역시 복잡해질 수 있다. 따라서 계통도에서 인접한 항목들을 자동적으로 찾아주는 방법이나, 분석자가 직접 여러 항목들을 선택한 경우, 연관 규칙 탐사를 실시하기 전에 함께 분석하는 것이 바람직한 항목을 추천해주는 기능 등이 유용할 것이다.

끝으로, HTM 접근법은 서로 상이한 분석을 실시하는 세 가지 단계로 구성되어 있어, 이들을 통합적으로 지원하는 시스템을 개발하는 것이 필요하다. 본 논문의 저자들은 향후 이상의 추후 연구 과제를 수행하면서 다양한 데이터에 대한 적용을 통해 제안하는 분석 방법의 유용성을 검증하고, 한계점을 보완해나갈 계획이다.

참고문헌

[1] R. Agrawal, T. Imielinski, and R. Swami, "Mining Associations between Sets of Items in Massive Databases," Proceedings of the ACM-SIGMOD 1993 International Conference on Management of Data, pp. 207-216, 1993.

[2] R. Agrawal, and R. Srikant, "Fast Algorithms for Mining Association Rules," Proceedings of the International Conference on Very Large Databases, pp. 125-131, 1994.

[3] P.-N. Tan, M. Steinbach, and V. Kumar, "*Introduction to Data Mining*," Addison-Wesley, 2005.

[4] A. Jorge, "Hierarchical Clustering for Thematic Browsing and Summarization of Large Sets of Association Rules," Proceedings of the 2004 SIAM International Conference on Data Mining, 2004.

[5] L. A. Fernandes, and A. C. B. Garcia, "Association Rule Visualization and Pruning through Response-Style Data Organization and Clustering," *In Advances in Artificial Intelligence-IBERAMIA*, pp. 71-80, 2012.

[6] Y. A. Sekhavat, and O. Hoeber, "Visualizing Association Rules Using Linked Matrix, Graph, and Detail Views," International Journal of Intelligence Science, Vol. 3, pp. 34-49, 2013.

[7] B. Schneiderman, "The Eyes Have It: A Task by Data Type Taxonomy for Information Visualization," Proceedings of the IEEE Symposium on Visual Languages, pp. 336-343, 1996.

[8] M. Klemettinen, H. Mannila, P. Ronkainen, H. Toivonen, and A. I. Verkamo, "Finding Interesting Rules from Large Sets of Discovered Association Rules," Proceedings of the 3rd International Conference on Information and Knowledge Management, pp. 401-407, 1994.

[9] P. C. Wong, P. Whitney, and J. Thomas, "Visualizing Association Rules for Text Mining," Proceedings of the 1999 IEEE Symposium on Information Visualization, pp. 120-123, 1999.

[10] C. Romero, J. M. Luna, J. R. Romero, and S. Ventura, "RM-Tool: A Framework for Discovering and Evaluating Association Rules," Advances in Engineering Software, Vol. 42, No. 8, pp. 566-576, 2011.

[11] R. J. Bayardo, and R. Agrawal, "Mining the Most Interesting Rules," Proceedings of the 5th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 145-154, 1999.

[12] M. Hahsler, and S. Chellubonia, "Visualizing Association Rules: Introduction to the R-extension Package arulesViz.", *R project module*, 2011.

[13] K. Techapichetvanich, and A. Datta, "VisAR: A New Technique for Visualizing Mined Association Rules," I*n Advanced Data Mining and Applications*, Springer Berlin Heidelberg, pp. 88-95, 2005.

[14] Y. H. Fua, M. O. Ward, and E. A. Rundensteiner, "Hierarchical Parallel Coordinates for Exploration of Large Datasets," Proceedings of the Conference on Visualization '99, pp. 43-50, 1999.

[15] P. Buono, and M. F. Costabile, "Visualizing Association Rules in a Framework for Visual Data Mining," In *Integrated Publication and Information Systems to Information and Knowledge Environments*, Springer Berlin Heidelberg, pp. 221-231, 2005.

[16] L. Yang, "Pruning and Visualizing Generalized Association Rules in Parallel Coordinates," IEEE Transactions on Knowledge and Data Engineering, Vol. 17, No. 1, pp. 60-70, 2005.

[17] L. Yang, "Visual Exploration of Frequent Itemsets and Association Rules," In *Visual Data Mining: Theory, Techniques and Tools for Visual Analytics*, Springer Berlin Heidelberg, pp. 60-75, 2008.

[18] M. Hahsler, and S. Chellubonia, "Visualizing Association Rules in Hierarchical Groups," Proceedings of the 42nd Symposium on the Interface: Statistical, Machine Learning, and Visualization Algorithms, 2011.

[19] K. H. Ong, K. L. Ong, W. K. Ng, and E. P. Lim, "Crystalclear: Active Visualization of Association Rules," Proceedings of the ICDM-02 Workshop on Active Mining, 2002.

[20] O. Couturier, T. Hamrouni, S. B. Yahia, and E. M. Nguifo, "A Scalable Association Rule Visualization towards Displaying Large Amounts of Knowledge," Proceedings of 11th International Conference on Information Visualization IV, Vol. 7, pp. 657-663, 2007.

[21] W. H. E. Day, and H. Edelsbrunner, "Efficient Algorithms for Agglomerative Hierarchical Clustering Method," Journal of Classification, Vol. 1, No. 1, pp. 7-24, 1984

[22] A. Guenoche, P. Hansen, and B. Jaumard, "Efficient Algorithms for Divisive Hierarchical Clustering," Journal of Classification, Vol. 8, No. 1, pp. 5-30, 1991.

[23] B. Sarwar, G. Karypis, J. Konstan, and J. Riedl, "Item-based Collaborative Filtering Recommendation Algorithms," Proceedings of the 10th International Conference on World Wide Web, pp. 285-295, 2001.

[24] C.-J. Tsui, P. Wang, K. R. Fleischmann, A. B. Sayeed, and A. Weinberg, "Building an IT Taxonomy with Co-occurrence Analysis, Hierarchical Clustering and Multidimensional Scaling," Proceedings of iConference, pp. 247-256, 2010.

[25] H. Hofmann, A. P. Siebes, and A. F. Wilhelm, "Visualizing Association Rules with Interactive Mosaic Plots," Proceedings of the ACMKDD International Conference on Knowledge Discovery and Data Mining, pp. 227-235, 2000.

[26] A. Strehl, G. K. Gupta, and J. Ghosh, "Distance Based Clustering of Association Rules,"

Proceedings of ANNIE 1999, ASME Press, pp. 759-764, 1999.

감사의 글

이 논문은 2012년도 정부(교육과학기술부)의 재원으로 한국연구재단의 기초연구사업 지원을 받아 수행된 것임 (2012R1A1A1044834)

저 자 소 개

김 준 우

2001: 한국과학기술원
산업공학과 공학사.
2003: 한국과학기술원
산업공학과 공학석사.
2009: 한국과학기술원
산업 및 시스템공학과 공학박사
현 재: 동아대학교
산업경영공학과 조교수
관심분야: 데이터마이닝, 지능형시스템, 조합최적화, 데이터 시각화, 피지컬 컴퓨팅
Email : kjunwoo@dau.ac.kr

강 현 경

2000: 한국방송통신대학교
경영학과 경영학사.
2004: 고신대학교
보건관리학과 보건학석사.
2008: 고신대학교
의학과 의학박사
현 재: 신라대학교
치위생학과 조교수
관심분야: 구강보건, 치면세마, 치주학, 치과방사선학, 포괄치위생학
Email : icando@silla.ac.kr