Journal of Knowledge Information Technology and Systems(JKITS), Vol. 10, No. 3, pp. 305~317, June 2015

Journal of Knowledge Information Technology and Systems

ISSN 1975-7700

http://www.kkits.or.kr

Association Rule Visualization by Structured Association Map

Jun Woo Kim*

Department of Industrial & Management Systems Engineering, Dong-A University

ABSTRACT

Association rule mining is one of the most popular data mining techniques, and its aim is to extract the association rules, the cause-and-effect relations between the items, from the given transaction data. Several algorithms such as apriori and its variants have been developed in order to extract the association rules in efficient way, however, they often produce the plethora of the association rules that is difficult for the analyzers to interpret and exploit. To address this issue, this paper aims to propose a visualization method called structured association map. The structured association map is a variant of the well known cluster hear map, and it focuses on ordering the items in more meaningful way. The structured association map and the cluster heat map have in common that the dendrogram obtained by hierarchical clustering is appended to the matrix for data visualization and the items are ordered according to the dendrogram. On the contrary, the primary difference between the two visualization methods lies in the way the hierarchy of the column items is generated. In structured association map, the row items are ordered at first and their order is considered in ordering the column items, while the row items and column items of the cluster heat map are ordered in similar manner. Consequently, the structured association map can represent both antecedents and consequents of the association rules in more effective way, and it is expected to help the analyzers to understand the structure of the extracted association rules more conveniently.

KEYWORDS: Association rule mining, Data mining, Visualizations, Data matrix, Cluster heat maps, Hierarchical clusterings, Matrix reorderings

ARTICLE INFO: Received 13 April 2015, Revised 12 June 2015, Accepted 12 June 2015.

*Corresponding author is with the Department of Industrial & Management Systems Engineering, Dong-A University, 840, Hadan 2-dong, Saha-gu, Busan, 604-714, KOREA.

E-mail address: kjunwoo@dau.ac.kr

1. 서 론

1.1 연관규칙 탐사

연관분석은 대표적인 데이터마이닝 기법 중 하나로 오늘날 널리 활용되지만, 분석 결과 추출된 연관규칙의 개수가 많은 경우 이를 해석하기 어렵다는 문제가 있다. 이에 본 논문에서는 추출된 연관규칙 시각화를 위한 새로운 방법을 제안하고자 한다. 먼저, 이산적인 항목(item)들이 주어져 있을 때, 한 개 이상의 항목으로 구성된 항목집합(itemset) X, Y 간에 성립하는 $X \rightarrow Y$ 형태의 규칙을 연관규칙(association rule)이라 한다. 연관규칙 $X \rightarrow Y$는 한 개 트랜잭션(transaction) 내에 X가 존재할 경우, Y도 동일 트랜잭션에 존재함을 의미하며, 이 때 X는 규칙의 전항(antecedent), Y는 후항(consequent)이라 부른다. 나아가, 주어진 트랜잭션 데이터에 존재하는 모든 유용한 연관규칙을 추출하는 과정을 연관규칙 탐사라고 하며, 자동화된 연관규칙 탐사 방법으로는 apriori 알고리즘 및 그 변종들이 잘 알려져 있다[1][2]. 개별 연관규칙의 유용성을 평가하는 지표로는 지지도(support), 신뢰도(confidence) 및 리프트(lift) 등이 있으며, 연관규칙 탐사 시에는 지지도 및 신뢰도 하한을 정해두고, 이를 만족하는 연관규칙을 유용한 것으로 판단하는 것이 일반적이다. 따라서, apriori 등과 같은 연관규칙 탐사 알고리즘은 보통 이러한 연관규칙들을 효율적으로 추출하는데 초점을 두고 개발되었다[3][4].

1.2 기존 연관분석 주요 한계점

연관규칙은 그 형태가 매우 직관적이고 의미하는 바가 명확하기 때문에 최근까지도 마케팅[5], 추천[6] 및 예측이나 진단[7] 등과 같은 다양한 분야에서 폭넓게 활용되어 왔다. 반면, 경우에 따라 추출되는 연관규칙의 개수가 폭증할 수 있다는 점은 연관분석의 주요 한계점 중 하나이다. 특히, 많은 항목들이 서로 복잡한 인과관계를 형성하는 경우 이러한 점이 두드러져, 분석자가 이들을 적절히 해석하고 활용하는 것이 어려울 수 있다[8]. 따라서, 가능하다면 추출된 연관규칙들을 정리하여 보다 편리하게 활용할 수 있는 형태로 가공할 필요가 있다.

추출된 연관규칙들을 정리하는 방법으로는 시각화(visualization)나 가지치기(pruning), 요약 및 군집 등이 있으며, 이들은 서로 혼합되어 사용되기도 한다[9][10][11]. 그 중에서도 시각화는 연관규칙들의 내용을 다양한 차트나 표 형태로 도식화하는 것을 의미한다. 이러한 시각화는 인간의 우월한 시각적 인지 및 정보처리 능력을 활용한다는 장점이 있으며, 최근 자동화된 알고리즘만으로 처리하기 곤란한 대규모 데이터가 늘어나면서 그 중요성이 점점 더 강조되고 있다[12].

다만, 시각화의 효과를 극대화하기 위해서는 가급적이면 내용을 인간이 이해하기 쉬운 간결한 형태로 표현하는 것이 중요하다. 아울러, 최근 대규모 데이터의 등장으로 인해 연관규칙 탐사 이외의 분야에서도 데이터 시각화의 개념이 많이 활용되는데[13][14], 이러한 일반적인 시각화 도구를 연관규칙 탐사 분야에 적용할 때는 기존 도구를 바로 적용하기보다 연관규칙의 특성에 맞게 이들을 적절히 변형하여 사용하는 것이 필요할 것이다. 이러한 맥락에서 본 논문은 탐사 결과 추출된 연관규칙들을 시각화하는데 사용할 수 있는 구조화된 연관맵을 제안하고자 한다. 구조화된 연관맵은 기존의 클러스터 히트맵(cluster heatmap)을 일부 변형한 것으로, 연관규칙이 의미하는 전항과 후항 사이의 다대다 관계를 좀 더 효과적으로 표현할 수 있도록 항목들을 적절히 정렬하는데 초점을 두고 개발되었다.

본 논문의 구성은 다음과 같다. 제 2장에서는 연관규칙 시각화 및 클러스터 히트맵을 이용한 데이터 시각화에 대한 기존 문헌들을 간단히 소개하고, 제 3장에서는 구조화된 연관맵의 특성 및 작성 절차에 대해 설명한다. 이어 제 4장에서는 구조화된 연관맵을 건강검진 데이터에 적용한 사례를 제시하며, 끝으로 제 5장에서 결론 및 추후 연구 과제에 대해 기술하고자 한다.

2. 관련 연구

2.1 연관규칙 시각화

일반적으로 연관규칙 탐사에서는 지지도나 신뢰도 하한을 만족하는 모든 연관규칙을 찾아내기 때문에, 방대한 양의 연관규칙이 추출될 수 있다. 나아가, 연관규칙의 전항이나 후항에는 2개 이상의 항목이 포함될 수 있어 항목들 간의 다대다 관계가 표현될 수도 있고, 임의의 연관규칙 $X \rightarrow Y$가 추출되는 경우 이와 관련성이 높은 하위 규칙, 즉, 전항이 X의 부분집합이거나 후항이 Y의 부분집합인 규칙들도 함께 산출되기도 한다. 따라서, 개별 연관규칙 $X \rightarrow Y$가 비교적 단순한 구조와 의미를 갖는 것과 달리, 이들의 집합은 매우 복잡한 특성을 가져, 분석자가 관찰하기 어려울 수 있다.

추출된 연관규칙들을 나타내는 가장 단순한 방법은 이들을 테이블에 나열하는 것이다. 이 때는 각 연관규칙의 지지도나 신뢰도, 리프트 등과 같은 평가 지표의 값을 함께 보여주면서 이들의 값에 따라 연관규칙들을 정렬하는 경우가 많다. 그 구조가 매우 단순하기 때문에 이 방법은 다양한 데이터마이닝 소프트웨어에서 채용되고 있기도 하다[9][15][16]. 그러나 테이블에 연관규칙을 나열하는 것만으로는 앞에서 설명한 연관규칙 집합의 복잡한 구조를 효과적으로 나타내는 것이 어렵다.

이에, 추출된 연관규칙의 내용들을 보다 시각적으로 나타내고자 하는 시도들이 다양하게 이루어졌으며, 이러한 예로는 산점도를 비롯한 간단한 차트를 이용하는 방법[17][18][19], 항목들을 세로축에 배치한 평행좌표(parallel coordinate)를 이용하는 방법[20][21][22], 노드로 표현된 항목집합들로 이루어진 연관규칙을 아크로 나타내는 그래프를 이용하는 방법[9][16][20][23] 및 행렬을 이용하는 방법[9][15][24] 등이 있다.

이 중에서도 행렬 기반 시각화는 다른 방법들에 비해 보다 간결하고 이해하기 쉬운 형태를 가지고 있다. 행렬 기반 시각화란 행렬의 행 및 열에 트랜잭션 데이터를 구성하는 항목들을 배치한 후, 행렬의 각 원소에 연관규칙 행 항목→열 항목을 대응시키는 것을 의미한다. 일반적으로는 각 원소에 해당 연관규칙의 특성, 즉, 신뢰도나 리프트 등과 같은 평가 지표 값을 기록하거나, 그래프 또는 색깔을 이용하여 평가 지표 값을 표현하는 경우가 많으며, <그림 1>에서 간단한 행렬 기반 시각화의 예를 볼 수 있다. 단, A, B, C, D는 각각 항목을 의미한다. 그리고 행렬의 원소에는 행 항목→열 항목에 해당하는 연관규칙의 신뢰도를 대응시켜, 이 중 신뢰도 하한을 넘는 것만 색칠을 하였다. 즉, <그림 1>의 의미는 세 개의 연관규칙 $\{B\} \rightarrow \{C\}$, $\{B\} \rightarrow \{D\}$, $\{D\} \rightarrow \{B\}$가 신뢰도 하한을 만족하는 유용한 규칙들임을 나타내고 있다.

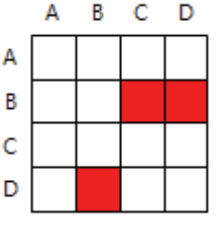

그림 1. 행렬 기반 시각화
Figure 1. Matrix based visualization

이러한 간단한 구조에도 불구하고, 행렬 기반 시각화는 다음과 같은 한계점을 갖는다. 첫째, 복수 항목으로 구성된 항목집합 간 연관규칙을 표현할 수가 없다. 행렬 기반 시각화에서는 행이나 열에 항목 1개가 배치되기 때문이다. 물론, 이 방법의 구조적인 특징으로 인해 복수 항목 간 연관규칙들을 직접적으로 나타내기는 어려우나, 사용자는 <그림 1>과 같은 행렬로부터 복수 항목 연관규칙에 대한 정보까지 얻고자 할 것이다. 두 번째 한계점은 항목들을 적절히 배열하지 않을 경우, 행렬 기반 시각화가 사용자에게 잘못된 정보를 제공할 수 있다는 점이다. 예를 들어, <그림 1>에서 직접적으로 표현되지는 않지만, $\{B\}\rightarrow\{C,D\}$도 유용한 규칙처럼 보일 수 있다. 하지만 이는 사실일 수도 있고, 그렇지 않을 수도 있다. 이러한 한계점들로 미루어볼 때, 행렬 기반 시각화를 연관규칙 탐사에 적용하기 위해서는 적절한 수정이 필요할 것이다.

2.2 행렬 기반 시각화와 클러스터 히트맵

최근 용량이 매우 크고 구조가 복잡한 데이터들의 활용이 늘어나면서 데이터의 시각적인 표현 방법에 많은 관심이 집중되었다[14][15]. 행렬 기반 시각화 역시 일반적인 데이터 시각화 방법 중의 하나이며, 기존에도 행 및 열 항목들을 적절히 배열하기 위한 연구가 이루어졌다. 다만, 일반적인 데이터 시각화에서는 행렬의 각 원소에 대응되는 값이 가급적이면 인접 원소와 비슷하도록 만들거나, 주 대각선(main diagonal) 인근에만 값이 큰 원소들을 배치하는 것이 주된 목적이었고, 이에 따라 발견적 기법(heuristic), 또는 특이값 분해(singular value decomposition)와 같은 선형대수적 분석이 활용되기도 하였다[25]. 반면, 클러스터 히트맵[26]은 행 및 열 항목들 간의 관련성에 기초하여 항목들을 배열하는 시각화 방법으로, 연관규칙 탐사에서도 유용하게 사용될 수 있다.

클러스터 맵은 <그림 2>와 같이 시각화를 위한 행렬에 항목 간의 계통도(dendrogram)가 추가된 형태이다. 이러한 계통도는 항목 간의 관련성을 계층적으로 보여주며, 관련성이 높은 항목들일수록 조기에 연결되고, 그렇지 못한 항목들은 계통도를 멀리 거슬러 올라가야 연결이 되는 특성을 갖는다. 즉, <그림 2>에서는 항목 B와 C의 관련성이 높고, 상대적으로 C와 D의 관련성은 낮다는 것을 의미한다. 따라서, 앞에서도 언급했던 연관규칙 $\{B\}\rightarrow\{C,D\}$가 그다지 유용하지 못할 것이라는 추론을 유도하는 효과를 얻을 수 있을 것이다.

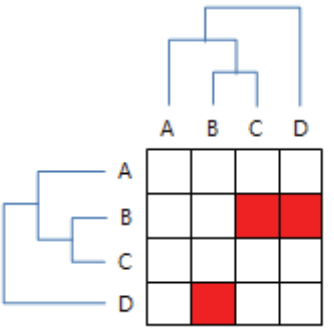

그림 2. 클러스터 히트맵
Figure 2. Cluster heat map

이러한 계통도를 작성하는 방법으로는 계층형 군집 분석이 있다[3]. 계층형 군집은 여러 항목들의 유사도(similarity) 또는 비유사도(distance)에 기반하여 항목들의 군집을 중첩적으로 생성해나가는 기법으로, 크게 병합형[27]과 분할형[28]으로 나누어지며, 클러스터 히트맵 작성 시에도 많이 활용되고 있다. 나아가, 클러스터 히트맵을 이용할 경우, 비교적 간결한 형태로 데이터의 내용을 표현할 수 있을 뿐만 아니라, 행 및 열 항목들을 어느 정도 의미있게 정렬하는 효과까지 거둘 수 있다. 이로 인해 최근에는 분자생물학 등의 분야에서 유전자

정보 등 대용량 데이터를 작성하는데 클러스터 히트맵이 활발히 사용된다[26]. 다만, 연관규칙 탐사에 적용하기 위해서는 클러스터 히트맵을 일부 개선할 부분이 있으며, 본 논문에서는 이러한 부분들에 대한 논의를 거쳐 연관규칙 시각화에 보다 적합한 구조화된 연관맵을 제안하고자 한다.

3. 구조화된 연관맵

클러스터 히트맵을 사용하기 위해서는 먼저, 행열의 각 원소에 대응되는 값이 무엇인지를 정의하는 것이 필요하다. 이는 기존 행렬 기반 시각화에서처럼 행 항목을 전항, 열 항목을 후항으로 하는 연관규칙의 유용성 지표, 즉, 신뢰도나 관심도를 사용할 수 있다.

두 번째로는 행 항목들에 대한 계통도를 생성하기 위해 이들 간의 유사도 또는 비유사도를 정의하는 것이 필요하다. 본 논문에서는 이를 내부 유사도(intra-similarity) 또는 내부 비유사도로 정의한다. 내부 유사도 또는 내부 비유사도는 연관규칙 탐사에서도 비교적 쉽게 정의할 수 있으며, 일반적으로 트랜잭션 데이터 내 항목 A, B 간의 내부 유사도로는 (1)과 같이 {A, B}의 지지도에 해당하는 공기 정보(co-occurrence)[29]를 사용할 수 있으며, A, B 간의 내부 비유사도로는 (2)와 같은 자카드(Jaccard) 거리[30]를 사용할 수 있다. 단, $|i|$는 항목 집합 i를 포함하는 트랜잭션의 개수를 의미하며, $|T|$는 데이터에 존재하는 모든 레코드의 개수를 의미한다.

$$SIM_{intra}(A,B) = SUPPORT(A \cup B) = \frac{|A \cup B|}{|T|} \quad (1)$$

$$DIST_{intra}(A,B) = 1 - \frac{|A \cap B|}{|A \cup B|} \quad (2)$$

클러스터 히트맵을 작성할 때, (1)이나 (2)와 같은 척도를 사용하여 행 항목들에 대한 계층형 군집 분석을 실시하면, <그림 2>에서 행렬 좌측에 보이는 것과 같은 계통도가 생성되어, 분석자에게 보다 의미 있는 정보를 제공함과 동시에 행 항목들을 적절히 배열할 수도 있게 된다. 나아가, 본 논문에서는 계통도의 단말(leaf)에 해당하는 행 항목들 중, 빈번하게 등장하는 것들이 상대적으로 위쪽에 배치될 수 있도록, <표 1>과 같은 계통도 수정 작업을 거치는 것을 제안한다. 단, m은 행 항목의 개수를 나타내고, 모든 행 항목의 집합을 I_R이라 할 때, *max_support* (X)는 I_R의 부분집합 X의 원소들의 지지도 중 최대값을 의미한다. 또한, $T_U(i)$, $T_L(i)$는 각각 병합형 계층 군집을 기준으로 i번째로 병합된 지점의 위쪽 및 아래쪽에 있는 항목들의 집합을 의미한다. 예를 들어, <그림 2>의 행렬 좌측에서 첫 번째 병합은 항목 B와 C의 연결을 의미하며, 두 번째 병합은 항목 A와 항목 집합 $\{B, C\}$의 연결이다. 따라서, 두 번째 병합을 기준으로 할 때, $T_U(2) = \{A\}$, $T_L(2) = \{B, C\}$이다. 한편, invert(i)는 i번째 병합 지점에서 위쪽과 아래쪽 부분의 위치를 서로 바꾸는 것을 의미하며, m개 항목을 대상으로 계층형 군집을 실시하는 경우, 총 m-1번의 병합이 일어나므로, <표 1>의 절차를 거쳐 계통도는 그 형태를 유지하면서도 지지도가 높은 항목을 최대한 위쪽에 배치하는 형태로 수정된다.

이제, 열 항목들에 대한 계통도를 생성하고, 이들을 적절히 행렬에 배치할 차례이다. 특히, 기존의 시각화 방법들은 보통 열 항목 역시 행 항목들과 동일한 방식으로 다루었지만, 본 논문에서는 연관규칙의 특성을 고려하기 위한 새로운 접근 방법을 제안하고자 한다.

표 1. 행 항목 계통도 수정 절차
Table 1. Refinement procedure for dendrogram of row items

```
For i=1 To m−1
  set S_U=max_support( T_U(i))
  set S_L=max_support( T_L(i))
  If S_L > S_U Then invert(i)
Next i
```

여기서 가장 중요한 것은, 앞에서도 설명한 바와 같이, 연관규칙 탐사에 적용할 경우, 행 항목들은 전항, 열 항목들은 후항에 해당한다는 점이다. 그리고 분석자들은 행렬 기반 시각화를 통해 복수 항목으로 구성된 연관규칙에 대한 정보까지 탐색하고자 하는 경향이 있다. 이러한 점을 고려하면, 열 항목들은 단순히 행 항목처럼 내부 유사도를 가지고 계통도를 생성하기보다, 연관규칙의 후항에서 함께 등장할 가능성이 높은 것들을 인접하게 배치하는 것이 보다 바람직하다. 이러한 맥락에서 본 논문은 열 항목들 간의 관련성을 측정하기 위한 상호 유사도(inter-similarity) 또는 상호 비유사도라는 개념을 사용하며, 이들의 목적은 열 항목 중 행 항목들에 대한 관련성이 유사한 것끼리 인접하게 배치하는 것이다.

n개의 열 항목이 존재하는 경우, 구조화된 연관맵에서는 먼저, 열 항목 j의 연관 벡터(association vector) a_j를 (3)과 같이 산출한다. 단, a_{ij}는 정렬된 행 항목 r_i ($r_j \in I_R$), 즉, i번째 행 항목과 열 항목 j 간의 연관성을 의미하며, (4)나 (5)와 같이 연관규칙 $\{i\} \rightarrow \{j\}$의 신뢰도 또는 리프트 등을 사용할 수 있다.

$$a_j = [a_{1j}\ a_{2j}\ a_{3j}\ \cdots\ a_{mj}] \tag{3}$$

$$confidence(i \rightarrow j) = \frac{support(r_i \cup j)}{support(i)} \tag{4}$$

$$lift(i \rightarrow j) = \frac{confidence(r_i \rightarrow j)}{support(j)} \tag{5}$$

열 항목들에 대한 계통도를 생성할 때는 (3)에서 주어진 연관 벡터를 이용하여 이들 간의 상호 유사도 또는 상호 비유사도를 산출한다. 즉, 예를 들어 두 개 열 항목 간의 상호 비유사도는 이들의 연관 벡터 사이 유클리드 거리 등으로 계산할 수 있다. 이렇게 상호 유사도 또는 비유사도가 산출된 이후에는 행 항목과 마찬가지로 계층형 군집을 통해 계통도를 생성할 수 있으며, 이를 <표 1>과 같은 방법으로 정렬하면 정렬된 열 항목들의 집합 $\{c_j | c_j \in I_C, j=1, 2, ..., n\}$을 얻는다. 단, c_j는 j번째 열 항목을 의미하며, n은 열 항목들의 개수, I_C는 모든 열 항목들의 집합을 의미한다. 결과적으로 이상과 같은 과정을 통해 행 항목과의 연관성이 서로 유사한 열 항목들을 인접한 곳에 배치하는 효과가 얻어진다.

정렬된 행 항목 r_i (i=1, 2, ..., m)와 정렬된 열 항목 c_j (j=1, 2, ..., n)가 얻어지면, 이제 시각화를 위한 $m \times n$ 행렬 V를 완성해야 한다. 이 행렬의 각 원소 v_{ij}에는 열 항목 r_i와 행 항목 c_j 간의 연관성을 대응시키며, 여기서 연관성이란 연관규칙 $\{r_i\} \rightarrow \{c_j\}$의 신뢰도나 리프트 등을 의미한다. 나아가, 각 v_{ij}에 이러한 연관성의 값을 직접 기록하기보다 <그림 1>이나 <그림 2>에서처럼 적절한 색깔을 통해 대응되는 연관성의 크기를 표현하면 보다 시각적으로 이해하는데 편리할 것이다.

지금까지 설명한 바에 따르면, 구조화된 연관맵에서는 행 항목들에 대해 그들 자체의 연관성, 즉, 내부 유사도 또는 비유사도를 이용하

여 계통도를 생성한다. 특히, 지지도나 자카드 거리와 같이 동일 트랜잭션 내 등장 빈도를 기준으로 할 경우, 빈발 항목집합을 형성하는 행 항목끼리 인접한 위치에 배치될 가능성이 높기 때문에 복수 항목으로 구성되는 전항에 대한 정보를 제공할 수 있다.

반면, 열 항목들은 행 항목과의 연관성이 유사한 것들끼리 인접한 위치에 배치되는 경향이 있다. 예를 들어, 구조화된 연관맵의 행 항목 r_i와 열 항목 c_j에 대해 연관규칙 $\{r_i\} \rightarrow \{c_j\}$의 신뢰도가 높다고 하자. 열 항목 c_j와 바로 인접한 다른 열 항목들인 c_{j-1}이나 c_{j+1}의 연관 벡터는 c_j와 유사하기 때문에, 연관규칙 $\{r_i\} \rightarrow \{c_j, c_{j-1}\}$ 또는 $\{r_i\} \rightarrow \{c_j, c_{j+1}\}$ 역시 유용할 가능성이 높을 것이다. 즉, 구조화된 연관맵은 복수 항목으로 구성되는 후항에 대한 정보를 제공할 수 있다. 나아가, 열 항목 계통도의 특성 상, 임의의 열 항목 c_j가 발생한 상황에서는 인접한 열 항목들인 c_{j-1}이나 c_{j+1}의 발생 가능성도 높다는 점을 추론해볼 수 있다. 이 열 항목들의 발생을 유도하는 행 항목들이 유사하기 때문이다. 다시 말해, 구조화된 연관맵에서는 인접한 열 항목들로 구성되는 $\{c_j\} \rightarrow \{c_{j-1}\}$이나 $\{c_j\} \rightarrow \{c_{j+1}\}$ 형태의 연관규칙들에 대한 관찰도 어느 정도 가능하다.

4. 적용 사례 및 실험 결과

다음으로는 구조화된 연관맵을 2011년 부산 소재 D고등학교 재학생 278명에 대한 치위생 건강검진 데이터에 적용해본 결과를 소개하고자 한다. 이 데이터에는 구강 내 건강 상태 및 생활 습관 등과 관련된 여러 가지 문항들이 포함되어 있으며, 이들 중에서도 연관규칙 탐사에 적합한 이진(binary) 문항 또는 적절히 이진화가 가능한 문항 19개를 선택하여 분석에 사용하였다. <표 2>는 각 문항들에 대한 설명과 응답 분포를 보여준다.

각 검진 문항들을 행 항목으로 배치하기 위해 (1)과 같은 지지도를 유사도로 사용하여 계통도를 생성한 후, <표 1>과 같은 절차에 의해 항목들을 정렬한 결과는 <그림 3>과 같다. 단, 계층형 군집분석 시, 병합 조건은 완전 링크(complete link)를 사용하였다. 이에 따라, 행 항목들은 19, 11, 12, 18, 7, 2, 13, 15, 5, 6, 4, 3, 16, 1, 8, 9, 10, 14, 17 순으로 배열된다. 단, 본 논문에서는 계통도에서 각 항목들간의 병합 관계만 나타내고, 병합되는 항목 집합 간의 유사도 또는 비유사도는 생략하였다.

표 2. 치위생 검진 문항
Table 2. The dental exmination variables

번호	문항	예	아니오
1	우식 치아 유무	82	196
2	우식 위험 치아 유무	26	252
3	부정교합 유무	38	240
4	구강 위생 상태 불량 여부	127	151
5	치주 질환 유무	33	245
6	구내염 및 연조직 질환 유무	26	252
7	깨지거나 부러진 치아 유무	53	225
8	냉・온 식음료 섭취 시 통증	70	208
9	치아가 욱신거리거나 아픔	27	251
10	잇몸이 아프거나 피가 남	54	224
11	불쾌한 입냄새가 남	48	230
12	아침 식사 전 이를 잘 닦음	104	174
13	아침 식사 후 이를 잘 닦음	165	113
14	점심 식사 후 이를 잘 닦음	66	212
15	저녁 식사 후 이를 잘 닦음	156	122
16	잠자기 전 이를 잘 닦음	124	154
17	간식 섭취 후 이를 잘 닦음	19	259
18	단 음식, 청량음료를 즐김	74	204
19	불소 함유된 치약 사용함	234	44

동일한 설문 항목들에 대해 <그림 3>에 나타낸 행 항목 계통도를 기반으로 (3)의 연관 벡터를 계산한 후, 이들 간의 유클리드 거리를 비유사도로 사용하여 열 항목 계통도를 생성한 결과는 <그림

4>에서 볼 수 있다. 열 항목 계통도 역시 항목들에 대한 정렬을 거쳤고, 병합 조건은 완전 링크를 사용하였다. 결과적으로 열 항목들은 19, 13, 15, 4, 1, 16, 12, 18, 14, 8, 10, 11, 9, 7, 2, 17, 3, 5, 6의 순서로 정렬되었다.

생성된 행 항목 계통도 및 열 항목 계통도를 이용하여 구조화된 연관맵을 작성한 결과는 <그림 5>에 나타나 있다. 단, 행렬의 각 원소에는 연관규칙 $\{r_i\} \rightarrow \{c_j\}$의 신뢰도를 대응시키고, 편의 상 0.7 이상의 신뢰도에 해당하는 원소들만 색칠하여 표시하였다. 이를 보면, 첫 번째 열은 모두 색칠이 되어 있는데, 이는 19번 항목의 경우 <표 2>에 보이듯이 대부분의 응답자가 '예'로 응답한 빈발 항목이기 때문이다. 그 외에는 6개의 원소가 색칠이 되어 있고, 이들은 행렬에 적절히 흩어져 있다.

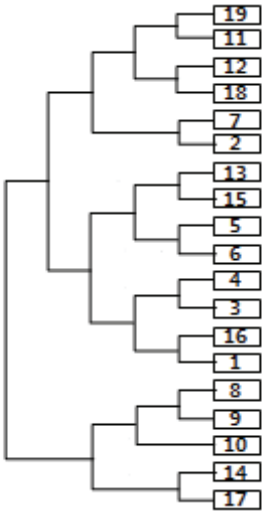

그림 3. 행 항목 계통도
Figure 3. Dendrogram of row items

이번에는 비교를 위해 기존의 클러스터 히트맵처럼 열 항목에 대해서도 행 항목의 계통도를 그대로 적용하여 <그림 6>과 같은 클러스터 히트맵을 작성하였다. 이를 보면 <그림 5>에서와 달리 색칠된 원소들이 좀 더 가까이에 모여 있음을 볼 수 있다. 따라서, 여기서는 우선, 본 논문에서 제안하는 구조화된 연관맵이 기존의 클러스터 히트맵과는 어느 정도 상이한 형태의 행렬을 생성한다는 점을 볼 수 있다.

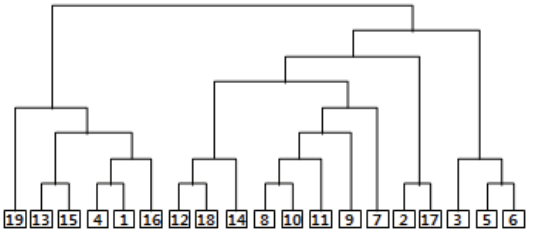

그림 4. 열 항목 계통도
Figure 4. Dendrogram of column items

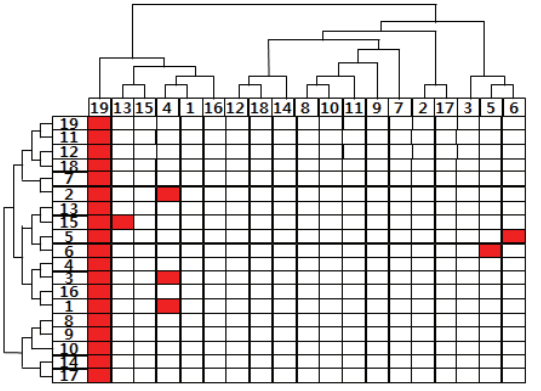

그림 5. 구조화된 연관맵
Figure 5. Structured association map

구조화된 연관맵과 클러스터 히트맵에서 행 항목들은 동일한 방식으로 정렬된다. 다만, 구조화된 연관맵은 새로운 방법으로 열 항목의 계통도를 생성하고 이들을 배치한다. 특히, 이러한 열 항목들은 연관규칙의 후항에 해당하기 때문에, 이번에는 데이터에서 추출된 유용한 연관규칙들의 후항이 구조화된 연관맵의 열 항목에서 어떻게 배치되는지를 평가해보고자 한다. 이를 위해, java 기반 오

픈소스 데이터마이닝 소프트웨어인 weka를 이용하여 치위생 건강 검진 데이터에 대표적인 연관규칙 탐사 방법인 apriori 알고리즘을 적용해보았다. 나아가, 알고리즘 적용 시에는 최소 지지도 0.1, 최소 리프트 1.1을 기준으로 하였으며, <표 3>은 추출된 결과에서 후항이 2-항목집합인 연관규칙들 중, 가장 리프트가 높은 상위 10개를 보여준다. 참고로 연관규칙은 리프트가 1보다 클수록 유용한 상관관계를 나타내기 때문에, 일반적으로 분석자는 <표 3>에 나열된 것과 같은 연관규칙을 파악하는데 매우 관심이 높을 것이다.

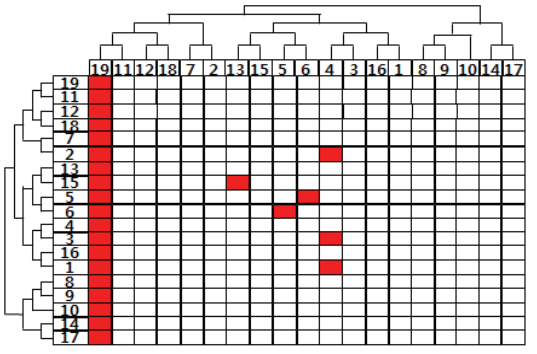

그림 6. 클러스터 히트맵
Figure 6. Cluster heat map

다만, 구조화된 연관맵이나 클러스터 히트맵과 같은 행렬 기반 시각화 방법에서는 전항과 후항이 모두 1-항목집합인 연관규칙만 직접적으로 표현되므로, <표 3>과 같은 연관규칙들은 분석자가 스스로 적절히 파악해야 한다. 이 때, 이 연관규칙들의 후항이 열 항목에서 가능한 인접하게 배치되어 있다면 분석자들은 행렬 기반 시각화로부터 좀 더 유용한 정보를 얻게 될 것이다. 이에 따라, <표 3>의 연관규칙들의 후항을 구성하는 두 개 항목들이 구조화된 연관맵 및 클러스터 히트맵의 가로축을 기준으로 얼마나 멀리 배치되었는지를 조사하여 <표 4>와 같은 결과를 얻었다. 단, 여기서 후항 항목들 간 거리란 이들이 각각 열 항목 c_a, c_b에 해당할 때, $|a-b|$를 의미한다.

표 3. 2-항목집합을 후항으로 갖는 유용한 연관규칙
Table 3. The interesting association rules with 2-itemset consequents

순위	연관규칙	리프트
1	$\{1\}\rightarrow\{4,15\}$	2.01
2	$\{1\}\rightarrow\{4,16\}$	1.88
3	$\{4\}\rightarrow\{1,15\}$	1.81
4	$\{4\}\rightarrow\{1,13\}$	1.76
5	$\{4\}\rightarrow\{1,16\}$	1.74
6	$\{1\}\rightarrow\{4,13\}$	1.72
7	$\{15\}\rightarrow\{13,14\}$	1.45
8	$\{13\}\rightarrow\{14,15\}$	1.37
9	$\{15\}\rightarrow\{8,13\}$	1.34
10	$\{15\}\rightarrow\{1,13\}$	1.30

표 4. 후항 항목들 간 거리
Table 4. The distance between the consequent items

순위	구조화된 연관맵	클러스터 히트맵
1	1	3
2	2	2
3	2	6
4	3	7
5	1	1
6	2	4
7	7	11
8	6	10
9	8	8
10	3	7

<표 4>에서 10개의 유용한 연관규칙들 중, 2, 5, 9번의 경우, 구조화된 연관맵 및 클러스터 히트맵에서의 후항 항목들 간 거리가 동일하고, 나머지 7개 연관규칙의 후항 항목들은 모두 구조화된 연관맵에서 더 인접함을 볼 수 있다. 따라서, 구조화된 연관맵이 복수 항목으로 구성된 연관규칙에 대한 정보를 더 많이 포함하고 있다는 결론을 내릴 수 있다.

5. 결론 및 추후 연구 과제

연관규칙 탐사는 데이터마이닝 기법 중에서도 개념이 비교적 단순하고, 다양한 분야에 널리 적용되어 왔다. 그럼에도 추출된 연관규칙의 개수가 많고 그 구조가 복잡할 경우, 내용을 파악하는 것이 어려워진다는 문제가 종종 지적되었으며, 최근에는 시각화를 통해 이러한 문제를 해소하기 위한 연구가 활발하였다. 그러나 복잡한 시각화 방법은 다양한 정보를 담을 수 있는 대신, 생성 과정이 까다롭고 사람이 이해하기 어려운 반면, 단순한 시각화 방법은 전달할 수 있는 정보가 제한적이라는 단점이 있었다. 한편, 본 논문에서 제안하는 구조화된 연관맵은 기존의 클러스터 히트맵처럼 이해하기 쉬우면서도 열 항목들의 적절한 배치를 통해 연관규칙 탐사에 필요한 정보를 더 많이 전달할 수 있도록 설계되었다.

구조화된 연관맵은 다음과 같이 활용될 수 있을 것으로 기대된다. 첫째, 추출된 연관규칙들의 시각화 자체가 목적인 경우, 구조화된 연관맵을 통해 보다 의미 있는 형태로 연관규칙들의 특성을 표현하는 것이 가능하다. 둘째, 항목이 매우 많아서 한꺼번에 연관규칙 탐사를 실시하는 것이 곤란한 경우, 탐사 알고리즘을 적용하기 전에 먼저 구조화된 연관맵을 작성하여, 유용한 연관규칙의 전항 및 후항을 구성할 가능성이 높은 항목들을 미리 선택한 다음, 선택된 항목들에 대해서만 연관규칙 탐사 알고리즘을 적용할 수도 있을 것이다. 즉, 구조화된 연관맵은 항목들의 적절한 필터링(filtering)을 통해 추출되는 연관규칙들을 사전에 가지치기하기 위한 용도로도 활용될 수 있다.

저자는 앞으로도 구조화된 연관맵의 개념을 보다 발전시켜 다양한 분야에 응용하고자 한다. 세부적인 추후 연구 과제들로는 다음과 같은 것들이 있다. 첫째, 구조화된 연관맵 생성을 위한 세부 사항들의 설정 방법에 대한 분석이 필요할 것으로 생각된다. 즉, 항목 간 유사도 또는 비유사도 산출이나 계층형 군집 분석 시 병합 기준 등에 대해 보다 다양한 방법을 적용하고, 각 상황에서 생성된 구조화된 연관맵의 특징을 분석하고자 한다. 둘째, 다양한 데이터에 구조화된 연관맵을 적용하면서 그 유용성을 검증함과 동시에, 추가적인 수정 및 보완 사항들을 도출하여 이를 반영하기 위한 연구도 지속될 것이다. 나아가, 구조화된 연관맵을 자동으로 생성할 수 있는 소프트웨어 모듈 또는 라이브러리를 개발하여 실제 현장의 분석자들이 편리하게 사용할 수 있도록 하는 것 역시 앞으로의 연구 과제 중 하나이다.

References

[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] B. Nath, D. K. Bhattacharyya, and A. Ghosh, *Incremental association rule mining: a survey*, Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, Vol. 3, No. 3, pp. 157-169, 2013.

[5] M. C. Chen, A. L. Chiu, and H. H. Chang, *Mining changes in customer behavior in retail marketing*, Expert Systems with Applications, Vol. 28, No. 4, pp.773-781, 2005.

[6] H. Kim, and H.-J. Kim, *A framework for tag-aware recommender systems*, Expert Systems with Applications, Vol. 41, No. 8, pp. 4000-4009, 2014.

[7] J. Jeon, and S. Y. Sohn, *Product failure pattern analysis from warranty data using association rule and Weibull regression analysis: a case study*, Reliability Engineering and System Safety, Vol. 133, pp. 176-183, 2015.

[8] B. Liu, W. Hsu, and Y. Ma, *Pruning and summarizing the discovered associations*, Proceedings of the fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 125-134, 1999.

[9] Y. A. Sekhavat, and O. Hoeber, *Visualizing association rules using liked matrix, graph and detail views*, International Journal of Intelligence Science, Vol. 3, pp. 34-49, 2013.

[10] Y. Djenouri, H. Drias, and A. Bendjoudi, *Pruning irrelevant association rules using knowledge mining*, International Journal of Business Intelligence and Data Mining, Vol. 9, No. 2, pp. 112-144, 2014.

[11] B. Lent, A. Swami, and J. Widon, *Clustering association rules*, Proceedings of 13th International Conference on Data Engineering, pp. 220-231, 1997.

[12] J.-W. Kim, and H.-K. Kang, *Visual exploration based approach for extracting the interesting association rules*, Journal of the Korea Society of Computer and Information, Vol. 18, No. 9, pp. 177-187, 2013.

[13] D. A. Keim, *Information visualization and visual data mining*, IEEE Transactions on Visualization and Computer Graphics, Vol. 7, No. 1, pp. 100-107, 2002.

[14] M. C. F. De Oliveira, and H. Levkowitz, *From visual data exploration to visual data mining: a survey*, IEEE Transactions on Visualization and Computer Graphics, Vol. 9, No. 3, pp. 378-394, 2003.

[15] P. C. Wong, P. Whiteney, and J. Thomas, *Visualizing association rules for text mining*, Proceedings of the 1999 IEEE Symposium on Information Visualization, pp. 120-123, 1999.

[16] 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.

[17] 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.

[18] M. Hahsler, and S. Chellubonia, *Visualizaing association rules: introduction to the R-extension package arulesViz*, R project module, 2011.

[19] K. Techapichetvanich, and A. Datta, *VisAR: a new technique for visualizing mined association rules*, in: X. Li, S. Wang, and Z. Y. Dong (Eds.), Advanced Data Mining and Applications, Springer Berlin, Heidelberg, pp. 88-95, 2005.

[20] P. Buono, and M. F. Constabile, *Visualizing association rules in a framework for visual data mining*, in: M. Hemmje, C. Niederee, and T. Risse (Eds.), From Integrated Publication and Information Systems to Information and Knowledge Environments, Springer Berlin, Heidelberg, pp. 221-231, 2005.

[21] 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.

[22] L. Yang, *Visual exploration of frequent itemsets and association rules*, Lecture Notes in Computer Science, Vol. 4404, pp. 60-75, 2008.

[23] S. Bornelöv, S. Marillet, and J. Komorowski, Ciruvis: a web-based tool for rule networks and interaction detection using rule-based classifiers, BMC Bioinformatics, Vol. 15, No. 1, pp. 139-150, 2014.

[24] M. Hahsler, and S. Chellubonia, *Visualizaing association rules in hierarchical groups*, Proceedings of the 42nd Symposium on the Interface Statistical, Machine Learning and Visualization Algorithms, 2011.

[25] I. Liiv, *Seriation and matrix reordering methods: an historical overview*, Statistical Analysis and Data Mining: The ASA Data Science Journal, Vol. 3, No. 2, pp. 70-91, 2010.

[26] L. Wilkinson, and M. Friendly, *The history of the cluster heat map*, The American Statistician, Vol. 63, No. 2, pp. 179-184, 2009.

[27] 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.

[28] A. Guenoche, P. Hansen, and B. Jaumard, *Efficient algorithms for divisive hierarchical clustering*, Journal of Classification, Vol. 8, No. 1, pp. 5-30, 1991.

[29] H.-K. Kang, and J.-W. Kim, *Hierarchical clustering based personalized feedback system for mass health examination*, Journal of Knowledge Information Technology and Systems, Vol. 6, No. 4, pp. 103-112, 2011.

[30] A. Strehl G. K. Gupta, and J. Ghosh, *Distance based clustering of association rules*, Proceedings of ANNIE 1999, ASME Press, pp. 759-764, 1999.

구조화된 연관맵에 의한 연관규칙 시각화

김준우

동아대학교 산업경영공학과

요 약

연관규칙 탐사는 대표적인 데이터 마이닝 기법 중의 하나로, 트랜잭션 데이터에 포함된 항목들 간의 인과관계를 나타내는 연관규칙의 추출을 목적으로 한다. 이러한 연관규칙을 효율적으로 추출하기 위해 apriori를 비롯한 다양한 알고리즘들이 개발되어 왔으나, 이들은 종종 매우 많은 양의 연관규칙을 생성하여, 분석자가 이를 해석하고 활용하는 것이 어려울 수 있다. 이러한 문제를 해결하기 위해 본 논문은 구조화된 연관맵이라는 시각화 방법을 제안한다. 구조화된 연관맵은 잘 알려진 클러스터 히트맵을 일부 변형한 것으로, 연관규칙의 전항과 후항 항목들을 보다 의미 있는 방식으로 정렬시키는데 초점을 맞춘다. 구조화된 연관맵과 클러스터 히트맵은 데이터 시각화를 위한 행렬에 계층 군집 분석을 통해 얻은 계통도를 추가하고, 이 계통도를 항목들의 정렬에 사용한다는 공통점이 있다. 반면, 두 시각화 방법의 주된 차이점은 열 항목의 계층 구조 생성 방식에 있다. 클러스터 히트맵에서 행과 열 항목들이 비슷한 방식으로 정렬되는 것과 달리, 구조화된 연관맵에서는 행 항목들을 먼저 정렬한 후, 관련 있는 행 항목들의 위치를 고려하여 열 항목을 정렬시킨다. 결과적으로 구조화된 연관맵은 유용한 연관규칙들의 전항과 후항 모두를 보다 효과적으로 표현할 수 있으며, 분석자들이 추출된 연관규칙들의 구조를 보다 편리하게 파악하는데 도움이 될 것으로 기대된다.

감사의 글

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

Jun Woo Kim received the bachelor's degree in the Department of Industrial Engineering from KAIST in 2001. He received the MS degree and the Ph.D. degree in the Department of Industrial and Systems Engineering from KAIST in 2003 and 2009, respectively. He has been a professor in the Department of Industrial and Management Systems Engineering at Dong-A University since 2011. His current research interests include data mining, intelligent system, combinatorial optimization, meta heuristic and data visualization, etc. He is a life member of the KKITS.

E-mail address: kjunwoo@dau.ac.kr