개념 계층 이용 흥미로운 부분 데이터의 탐색[1]

홍정희*, 김성민*, 남도원*, 이동하**, 이전영*

*포항공과대학교 전자컴퓨터공학부

**포항공과대학교 정보통신연구소

(manifold, smkim, irene, dongha, jeon)@white.postech.ac.kr

Discovery of Interesting Knowlege using Concept Hierarchy

Jeong-Hee Hong*, Sung-Min Kim*, Do-Won Nam*, Dong-Ha Lee**, Jeon-Young Lee*

*Division of Electrical and Computer Engineering, POSTECH

** POSTECH Information Research Laboratories, POSTECH

요약(Abstract)

개념 계층(Concept Hierarchy)은 데이터베이스 분야에서 사용되는 대표적인 배경 지식(Background Knowledge)으로써, 데이터베이스에 내재되어 있는 구조적인 정보, 데이터의 분포, 영역전문가(Domain Expert)에 의해 주어지는 외부 지식 등이 반영되어 있다. 개념계층의 특성상 부모(parent)-자식(child) 관계가 있는 두 노드가 있을 때, 한 노드의 값으로부터 다른 노드의 값을 추정할 수 있다. 이 추정된 값을 기대치라고 하고, 한 노드의 값으로부터 추정된 기대치와 실제치가 상당히 상이한 값을 보이는 노드가 있을 때, 이를 흥미롭다(interesting)고 말할 수 있다. 그러나 아직까지 개념계층상에서의 흥미로운 부분 탐색에 대한 연구가 없었으며, 흥미로움(interestingness)의 척도(measurement)에 대한 연구로서는 신뢰도(confidence),리프트(lift),컨빅션(conviction)등이 있었다. 그러나 이런 흥미도의 척도에 관한 연구도 연관규칙에 한정되어 이루어졌으므로 개념계층상의 데이터에 적용하기 위해서는 약간의 수정 및 새로운 정의가 필요하다.

본 논문에서는 데이터의 특성에 따른 개념계층이 존재할 때, 이를 이용하여 기대치와 실제치가 상이한 흥미로운 부분을 발견하고자 하며, 이를 위하여 개념계층상에서의 흥미도의 척도를 제안하고 흥미로운 부분을 탐색하는 방법을 기술하고자 한다. 또한 데이터마이닝의 결과인 연관규칙을 개념계층에 적용하여 연관규칙을 통해 얻어질 수 있는 기대치를, 지지도(support),신뢰도(confidence), 리프트(lift), 컨빅션(conviction)등의 관계를 통해 다양한 방법으로 모색해본다.

이 연구에서 제안하는 이러한 개념계층상의 흥미로운 부분의 탐색은, 전자 상거래에서의 CRM(Customer Relationship Management)나 틈새시장(niche market) 마케팅 등에 적용가능하리라 여겨진다.

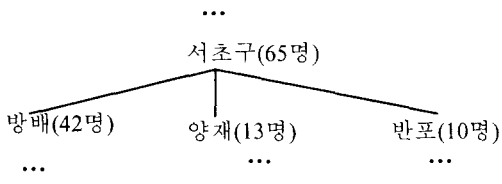
1. 서론

개념계층(Concept Hierarchy)은 데이터 집합이나 데이터의 도메인(domain)에 대하여 데이터의 분포나 데이터의 성격, 특성 등을 잘 표현해주므로 데이터베이스 분야에서 유용한 배경지식으로 쓰인다.[1][2] 개념계층(Concept Hierarchy)은 개념 공간상(Concept Space)에 존재하는 개념(Concept)들을 계층에 따라 일반적인 개념에서 구체적 개념(General-to-Specific Ordering)으로 나열한 것으로서, 개념계층에서의 부모노드와 자식노드와의 관계는, 자식노드 개념을 한 단계 일반화시킨 것이 부모노드이다. 그러므로, 부모노드가 뜻하는 성질의 대부분을 자식노드들이 그대로 지니고 있을 것이라는 가정이 성립하고 부모노드(일반적

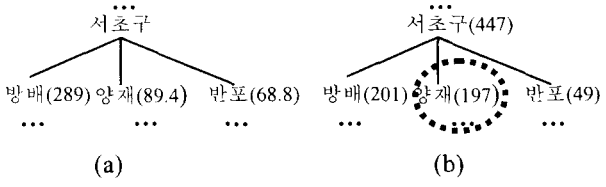
1. 본 연구는 BK21 사업을 통하여 포항공과대학교 전자컴퓨터공학부 에 주어진 교육부의 재정지원을 통해 이루어진 것입니다.

인 개념)를 통해서 그에 대한 자식노드(구체적인 개념)의 추측이 가능하다. 그러나 이러한 개념계층의 성질에 위배될 수 있는 개념, 즉 일반적 개념을 통해 추측한 개념과는 다른 예외적인 구체적 개념이 비록 다수는 아닐지라도 존재함이 가능하다.

이 논문의 주요 주제는, 데이터베이스에 저장된 데이터들에 대해 주어진, 일반적 성질을 통해 구체적 성질을 예측하는데 유용한 개념계층을 이용하여 이를 토대로, 상이한 성질을 보이는 예외적인 흥미로운 부분 데이터를 찾아내는 것이다. 예를 들어, 한 인터넷 서점의 데이타베이스에 저장된 고객 데이터에서 생성된 지역에 따른 개념계층을 고려해보자. [그림1]은 이러한 개념계층의 일부분이며, 괄호 안의 숫자는 그 노드에 속하는 고객들의 수이다. 또한 이러한 개념계층에 따른 도서의 구입수 대한 기대치와 실제치를 나타낸 것이 [그림2]이다.



[그림 1] 고객들에 대한 지역에 따른 개념계층



[그림 2] 지역에 대한 개념계층에 따른 도서 구입수의 기대치와(a) 실제치(b)

[그림2]에서 보듯이, 방배,반포동에 사는 고객들의 경우의 도서 구입수는 개념계층에 따라 계산된 기대치와 실제치가 거의 차이가 없지만, 양재동에 사는 고객들의 경우엔 도서 구입수의 기대치가 89.4권인데 비해 실체치는 197권으로 2배가 조금 넘게 차이가 나므로, 이 양재동에 사는 고객들의 도서 구입수를 흥미롭다(interesting)고 할 수 있다.

즉, 이 논문에서는, 이러한 데이터들의 특성이나 구조를 잘 나타낸 개념계층을 이용하여, 일반적인 특성을 통해 구체적인 특성을 추측할 수 있다는 개념계층의 성질에 따라, 형제노드들을 일반화시켜 구한 부모노드에 대해, 다른 형제노드들에 비해 상이하게 부모노드와 다른 특성을 지닌 노드가 존재하는지를 찾아내는 척도 및 방법을 제시하고자 한다. 특히, 응용이 가능한 인터넷의 쇼핑몰에서의 고객데이타에 대해, 조사하고자 하는 고객의 특정 성질에 대한 개념계층을 이용하여 기대와는 다른 흥미로운 부분을 찾는 것을 예로 들어 기술한다.

개념계층을 통해 이러한 예외적인 계층을 발견하고, 또한 발견하기 위해 필요로 하는 척도나 방법에 대한 연구는 그 자체로서도 흥미로울 뿐만 아니라, 전자 상거래에서의 고객 세분화(segmentation), 고객 차별 마케팅이나 신용카드 등의 사기 행각 발견 등 여러가지의 유용한 분야로의 응용도 가능한 가치있는 연구가 될 것이라 생각된다.

이 논문에서는 이러한 흥미로운 부분 데이터를 탐색하는데 있어서, 단일 특성 및 이를 일반화시킨 복수개의 특성에 대해서도 탐색 방법 및 척도를 제시한다. 또한 특성의 빈도수 대신 특정한 연관규칙을 만족하는 지지도(support)를 적용하여, 예외적으로 연관규칙이 맞지않는 부분 데이터를 발견하는 방법을 제시함으로서, 개념계층에 대한 연관규칙의 적용도 고려하였다. 참고로 주의할 것은, 보통 연관규칙에 대한 지지도는 전체 트랜잭션의 수에 대한 그 연관규칙을 만족하는 트랜잭션의 수로 나타내지만, 여기서는 편의상, 지지도를 단순히 그 연관규칙을 만족하는 트랜잭션의 수로서 나타내기로 한다. 그리고 마지막으로 항목들의 개념계층과 데이터마이닝의 결과인 연관규칙을 통해 얻을 수 있는 기대치를 여러가지 척도들의 관계를 통해 모색해보았다.

본 논문의 구성은 다음과 같다.

2장에서는 이 논문에서 풀고자 하는 문제에 대해 관련된 앞선 연구들에 관해 살펴보고, 3장에서는, 쇼핑몰의 데이터를 예로 들어 고객들의 지역에 따른 개념계층을 이용하여 지역에 따른 도서 구입수에 대해 기대치와 다른 고객층을 찾아냄으로써, 특정 개념계층을 이용하여 단일 항목에 대해 추측되는 기대치와 다른 패턴을 보이는 흥미로운 부분 데이터를 찾아내는 방법 및 측정하는 척도를 제시한다.(3.1) 또한 이를 좀더 일반화하여 복수개의 항목에 대

해 추측되는 기대치와 다른 패턴들을 보이는 부분 데이터를 찾아내는 방법 및 척도도 제시한다.(3.2) 그리고, 흥미로움(interestingness)을 측정할 때 항목에 대한 빈도수 대신 특정 연관규칙의 지지도를 적용하여 예외적으로 연관규칙이 적용되지 않는 데이터 부분을 발견하는 방법을 제시하고,(3.3) 마지막으로 항목들의 개념계층과 데이터마이닝의 결과인 연관규칙을 통해 얻을 수 있는 기대치를 여러가지 척도들의 관계를 통해 모색해본다.(3.4)
끝으로, 4장에서는 결론 및 향후 연구과제를 제시하며 끝을 맺겠다.

2. 관련연구

앞선 연구들 중에 개념계층을 이용하여 흥미로운 부분을 탐색하고자 하는 뚜렷한 시도는 아직까지 없었다. 따라서 대신 개념계층과 흥미로움(interestingness)에 관련된 앞선 연구들에 대해서 각각 살펴보기로 한다.

2.1 개념계층(Concept Hierarchy)

개념계층(concept hierarchy)이란 개념 공간상에 존재하는 개념들을 가장 일반적인 개념에서 출발하여, 그에 따른 구체적인 개념 순서로 그래프, 트리, DAG등으로 표현하여 레벨에 따라 나타낸 것이다. 즉, 개념계층에서의 자식노드와 부모노드와의 관계는, 어떤 구체적인 개념과 그에 대응하는 일반화된 개념과 같다.[1][3]
개념 계층에 대한 정의를 내리자면 다음과 같다.

[정의] 개념 계층(concept hierarchy) H 는 반순서관계(partial order) 집합 (H,<)이다. 이때 H는 개념들의 유한 집합이고, < 는 H 상에 정의된 반순서관계이다.[1][3]

개념계층은 지식탐사에서 유용한 배경지식으로 이용될 수 있는데, 그 이유는 일반화 과정에 유용하고 표현 방법이 간단하기 때문이다.
또한 개념계층을 데이터의 분포를 고려하여 자동적으로 생성 또는 수정하는 연구[1][2][3][4][5][6]도 몇몇 있었으나, 지식공학자(Knowledge engineer)나 영역전문가(domain-expert)의 역할이 크며, 본 논문에서는 개념계층이 고정되어 있다고 가정한다.

2.2 흥미로움(interestingness)

거의 틀림이 없을 것이라 기대했던 패턴에 대해 의외로 다른 패턴을 얻게 되면, 이러한 패턴은 '흥미롭다'고 말할 수 있다. 데이터마이닝 분야에서 이러한 흥미로운 데이터들에 대한 연구는, 대부분 연관규칙에 대해 이러한 흥미로운 정도를 측정할 수 있는 척도(measurement)에 대해 연구함으로써 이루어져 왔다. 이러한 흥미로움에 관해 연구된 척도에는, 신뢰도(confidence), 리프트(lift), 컨빅션(conviction) 등이 있다. [3][4]

신뢰도는 연관규칙의 중요한 척도로서 등장했지만, A→C라는 연관규칙에 대해 C의 지지도(support)를 고려하지 않기 때문에, A와 무관하게 C의 지지도가 클 경우 높은 신뢰도를 나타내어 잘못된 결과를 가져올 수 있다. 이러한 문제점을 해결하기 위해 리프트(lift)가 제안되어 A,C의 correlation의 경우를 잘 측정할 수는 있으나, 서로 대칭적(symmetric)이므로 연관규칙의 의미(implication)를 정확히 표현하고 있다고 하기는 어렵다. 리프트의 이러한 문제점을 해결하기 위해 제안된 것이 컨빅션(conviction)이다. 이것은 C가 일어날 확률대신 C가 일어나지 않을 확률을 고려함으로써 연관규칙의 의미(implication)를 잘 표현할 수 있다. 이러한 척도들의 자세한 의미는 [표1]에 정리되어 있다.

	연관규칙 A→C 에 대해		
흥미로움 척도들	신뢰도 (confidence)	리프트 (lift)	컨빅션 (conviction)
표현식	$\frac{sup(A \rightarrow C)}{sup(C)}$	$\frac{P(AandC)}{P(A)P(C)}$	$\frac{P(A)P(\neg C)}{P(Aand\neg C)}$

[표1]흥미로움(interestingness)에 대한 척도들

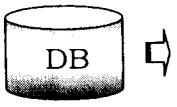
지금까지 연구된 이러한 흥미로움에 대한 척도들은 연관규칙을 대상으로 연구되고 제안되어 왔었다. 이 논문에서는 연관규칙만이 아닌 개념계층의 데이터들에 적용시키기 위하여 3장에서 이러한 척도들을 의미를 약간 수정하거나 혼합, 또는 새로 정의하기로 한다.

3. 개념계층의 예외적 패턴 탐사 방법

3.1 데이터의 개념계층 표현

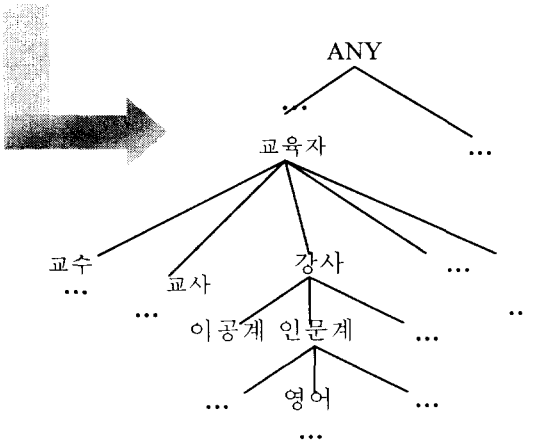
우선 데이터베이스에 저장된 데이터들이 개념계층으로 어떻게 표현되는가를, 한 쇼핑몰의 고객들의 직업에 따른 개념계층을 예를 들어 살펴보기로 한다.

[표2]과 같이 고객에 관한 데이터들이 데이터베이스에 저장되어 있다. 이 데이터들에 대해 직업에 따른 개념계층으로 표현하고 싶다면 가장 상위레벨인 루트(root)는 모든 것을 포함하는 'ANY'가 되고, 그 자식노드들로서 크게 '교육자'와 'Not교육자'로 나눌 수 있으며, '교육자'노드에 대한 자식노드들로서 '교사','교수','강사'등을 두어 [그림3] 처럼 나타낼 수 있다. 즉, 부모노드에 관한 좀더 구체적인 개념이 자식노드가 되는 것이다. 그리고 가장 하위레벨인 말단노드(leaf node)에는 모든 고객들이 나타난다. 개념계층에 관한 더 자세한 사항은 [1],[2]를 참조하기 바란다.



고객 ID	성명	나이	구체적 직업	…
30201	김은주	29	영어강사	…
30202	이전영	47	이공계 교수	…
31203	오은성	29	전자제품 판매업	…
…	…	…	…	…

[표2] 데이터가 데이터베이스에 저장된 상태



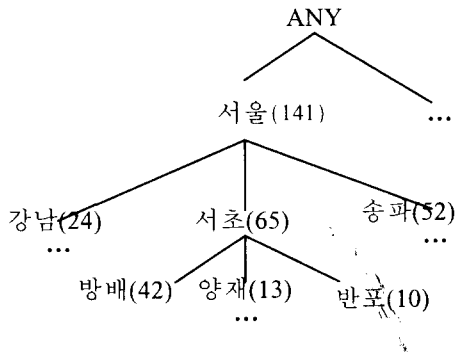
[그림3] 직업에 따른 개념 계층

그러면 이 3장의 나머지 부분에서는 우선 탐사하고자 하는 대상 항목이 하나일 때와, 좀더 일반적인 경우인 탐사 대상 항목이 복수개 일때에 대한 흥미로운 부분 탐사 방법을 제안한다., 이어서 개념을 측정할 때 항목리스트의 빈도수 대신 특정 연관규칙의 지지도를 사용하여 예외적으로 연관규칙이 잘 적용되지 않는 계층을 발견하는 방법을 제시한다.

3.2 개념계층의 단일 항목에 대한 흥미로움 탐사

앞에서 예로 들었듯이, 지역에 따른 고객들 중 서초구에 사는 고객들이 인터넷 쇼핑몰에서 구입하는 특정 항목에 대해 조사했을 때, 방배동과 반포동에 사는 고객들은 평균적으로 그 품목에 대해 개인당 4~5개 정도의 구매경향을 보여 기대치와 비슷한 반면, 양재동에 사는 고객들은 개인당 14개 정도의 항목를 구입하여, 기대치보다 3배나 많은 구매경향을 보였다면, 양재동에 사는 고객들의 구매패턴은 서초구에 사는 고객들의 평균 구매패턴에 비추어 보았을 때, 비교적 예외적인 것임을 알 수 있고, 그러므로 이 집단은 관심을 가질만하다는 것임을 알 수 있다. 그러면 비슷한 부류에 속하는 고객의 평균 항목구매 개수와 어느정도 동떨어진 개수만큼 물품을 구입하는 각 고객이나 계층을 탐사하는 방법에 대해 살펴보기로 한다.

앞의 서론부분에서 예로 들었듯이, 한 인터넷 서점에서 고객들의 지역에 따른 도서 판매량을 분석하여 기대치와 어긋나는 구매형태를 보이는 집단을 분석하여 마케팅 전략을 새롭게 구상하려 한다고 가정하자. 이를 해결하기 위해서는 우선 배경지식으로 [그림4]와 같은 고객들의 지역에 대한 개념 계층이 필요하다.

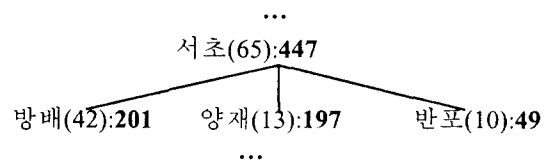


[그림4] 지역에 따른 개념계층과 고객수

각 노드의 괄호안의 숫자는 그 노드에 해당하는 고객의

수를 나타낸다. 위 개념계층의 가장 하위레벨인 말단노드(leaf node)에는 각 고객들이 나타난다.

위의 개념계층을 토대로 가장 하위레벨부터 도서구입수에 대해 살펴본 중간부분의 일부분이 [그림 5]이다.



[그림5]개념계층에 따른 도서 구입수

각 노드에서 괄호안의 숫자는 그 계층에 속하는 고객수이고, 그 옆의 굵은 글씨체의 숫자는 그 계층에서 구매한 총 도서수를 나타낸다. 즉, 양재에 사는 회원 고객들은 총 14명이고 이들이 구입한 도서수는 198권이다.

여기에 기대와는 달리 비슷한 부류에 속하는 고객들과 상이한 구매패턴을 보이는 고객 집단을 찾기 위한 하나의 척도(measurement)로서, 앞의 관련연구에서 살펴본 lift를 제안한다. 앞에서는 연관규칙을 위한 척도였으나, 이 논문에서는 개념계층을 이용하여 데이터들의 흥미로움을 측정하기 위해 다음과 같이 $lift_{li}$로 표현한다.

$lift_{li}$는 l 레벨의 i노드에 대해 그 노드에서 나타나는 항목의 개수의 기대값 $E(n_{li})$에 대한 실제값 n_{li}의 비율이다.

$$lift_{li} = \frac{n_{li}}{E(n_{li})}$$

l 레벨의 i노드에 대해 그 노드에서 나타나는 항목의 개수의 기대값 $E(n_{li})$는 다음과 같은 식으로 나타낼 수 있다.

$$E(n_{li}) = n_{l-1,j} \times \frac{c_{l,i}}{c_{l-1,j}}$$

$n_{l-1,j}$: l-1 레벨의 j번째 노드(부모노드)의 총 구입항목 개수

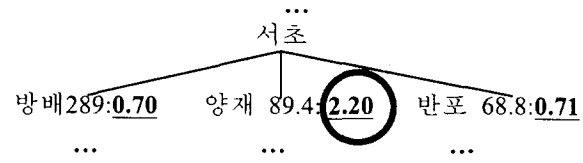
c_{li} : l 레벨의 i노드에 대해 그 노드에 해당되는 고객의 수

$c_{l-1,j}$: l-1레벨의 노드에 j노드(부모노드)에 대해 그 노드에 해당되는 고객의 수

$lift_{li}$의 값이 1에서 멀어질수록 기대한 수치를 만족하지 못하므로, $lift_{li}$의 값과 1과 차이가 임의의 한계치(threshold) t 값을 넘으면, l레벨의 i번째 노드는 우리가 얻고자 하는 흥미로운 구매패턴을 보이는 고객 부분 데이터이다. 또한 얻어진 $lift_{li} > 1$ 이면, 그 노드는 기대치보다 그 특성이 많이 나타나는 긍정적인(positive) 연관성을 지닌 것이고, $lift_{li} < 1$ 이면, 그 노드는 기대치보다 그 특성이 적게 나타나는 부정적인(negative) 연관성을 지녔다고 할 수 있다.

3.2.1 척도 적용 결과

앞에서 제시한 $lift_{li}$를 이용하여 [그림4]에 대해, 방배동,양재동, 반포동에 사는 고객들이 구입한 도서수에 대한 기대값과 $lift_{li}$(굵은숫자로 표시)를 구하여 보면 다음의 [그림6]과 같다.



[그림6] 지역에 따른 도서 구입수에 대한 기대값과 $lift_{li}$

위 [그림 5]에서 보면 방배동과 반포동에 사는 고객들에 대한 도서 구입수의 $lift_{li}$는 1에 가까운 반면, 양재동에 사는 고객들에 대한 도서구입수의 $lift_{li}$는 2.20으로서 기대치보다 2배가 넘는 양만큼 도서를 평균적으로 구입했음을 알 수 있다. 따라서, 역동적인(dynamic) 피드백(feedback) 과정을 통하여, 양재동에 사는 고객들에 대한 비중(weight)를 높여 이 고객들에게는 도서 구입시 특히 할인을 더 많이 해준다든지, 도서에 관련된 카탈로그나 정보들을 이메일이나 우편을 통해 보내는 등의 전략으로 구매를 더욱 유도할 수 있을 것이다. 또한 다시 양재동에 사는 고객들의 도서 구입수가 평균치로 낮아졌다면 이 또한 흥미로운 상황이므로 피드백 과정을 통해 비중을 다시 조절하여 적절한 전략을 마련할 수 있을 것이다.

3.3 개념계층의 복수 항목에 대한 흥미로운 부분 탐사

만약 전산업계에 종사하는 고객들이 인터넷 쇼핑몰에서 구입하는 도서들이 비슷하다는 가정하에 도서들를 종류별로 나눈 항목들에 대해 평균을 내었을 때, 네트웍 분야, 데이터베이스 분야, 알고리즘 분야에 종사하는 고객들이

평균적으로 어떤 항목에 대해 50% 이상의 구매경향을 보인 반면, 인공지능 분야에 속하는 직업을 가진 고객들은 평균적으로 10%이하의 구매경향을 보였다면 이 고객들의 도서 구매패턴은 전산업계에 종사하는 고객들의 평균 도서 구매패턴에 비추어 보았을 때, 비교적 예외적인 것임을 알 수 있고, 그러므로 이 고객층의 도서 구입패턴은 관심을 가질만하다는 것임을 알 수 있다.

이 장에서는, 이러한 관심을 가질만한 예외적인 항목을 구매하는 부분 데이터(고객층)를 다른 예에도 적용할 수 있도록 일반화시켜 탐사하는 방법에 대해 기술한다.

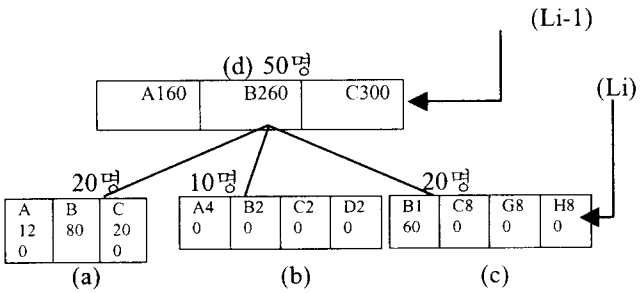
인터넷 쇼핑몰을 이용하는 고객들에 대한 A,B,C,D,..라는 항목에 대한 항목구매 리스트(product list)가 다음 [표3]과 같이 존재할 것이다.

CustID	A	B	C	D	E	....

[표3]고객의 항목(item)구매 리스트

아래의 [그림7]처럼 특정 개념계층에 따른 항목구매 리스트 고려했을 때, 가장 하위 계층 중 하나이며 말단노드(leaf node)들인 (L_i) 레벨을 먼저 살펴보기로 하자. (L_i)레벨에 속하는 각 고객들의 [표3]과 같은 항목구매 리스트 중에서 각 고객당 평균 구입수가 한계치 s값 이상인 것들만 고려한다. 직업에 따른 개념계층 중, 일부분에 해당하는 고객 데이터에 대한 항목구매리스트를 살펴보면 다음 [그림7]과 같다. 항목 옆의 숫자는 그 노드에서 그 항목을 구매한 총 개수를 나타낸 것이고, 항목구매리스트 위의 숫자는 그 노드에 속하는 총 고객수를 나타낸 것이다.

설명의 편의상, (a),(b),(c),(d)는 각 항목구매리스트 노드를 가리키고, (L_{i-1}),(L_i)는 각 레벨을 가르킨다고 하자.



[그림 7]특정 개념계층에 따른 일부 항목구매 리스트

이 항목구매 리스트의 개념계층의 생성은 가장 하위 레벨부터 이루어진다. 이 개념계층의 생성과정을 살펴보면 다음과 같다. 우선 (L_i)레벨의 항목구매 리스트 (a),(b),(c)는 각 노드에 속하는 고객들의 항목구매 리스트 중 각 고객의 평균 구입수가 s개 이상인 항목들만을 추려내어 만든다. 그 후, (a),(b),(c) 노드들의 항목구매리스트들을 합하여(union) 한 단계 위 레벨인 (L_{i-1})레벨의 후보(candidate) 항목구매 리스트 (d')를 만든다. 이 후보 항목구매 리스트 (d')에서 이 리스트에 있는 각 항목에 대해 각 고객이 그 항목을 구입한 평균 개수를 구하여 사용자에 의해 주어진 지지도 s를 만족하는 것만을 추려내 (L_{i-1})레벨의 항목구매 리스트를 만들면 (d)와 같이 된다.

여기서 (L_i)에 해당하는 노드들 중 예외적인 노드를 발견하기 위한 척도(measurement)로 즉, i 레벨에서의 j번째 노드와 그 한 단계 일반화된 상위 계층인 i-1 레벨의 노드에 대한 차이(difference)를 나타내는 척도(measurement) 로 다음과 같은 함수 $diff^{1st}$,$diff^{2nd}$ 를 제안한다.

[정의]

$$diff^{1st} = \sum \left| \left(\frac{n_{l-1jk}}{\sum n_{l-1j}} - \frac{n_{lik}}{\sum n_{li}} \right) \times 100 \right|$$

$n_{l-1,j,k}$: i-1 레벨 j 번째 노드의 k라는 항목의 property(ex: 항목 개수)

$n_{l,i,k}$: l 레벨의 i번째 노드의 k라는 항목의 property

$n_{l,i}$: l 레벨의 i번째 노드의 항목의 property

위 함수 $diff^{1st}$의 의미는 i 레벨의 한 노드에 대한 그 항목들의 퍼센트와 한 단계 일반화시킨 i-1레벨의 노드의 항목들의 퍼센트와의 차이의 합을 나타낸다. 즉, 한 노드와 그 부모 노드에서 각 항목이 차지하는 백분율의 차이의 합이다. 따라서 $diff^{1st}$ 의 값이 클수록 그 노드에 속하는 고객들은 평균에 비해 예외적인 항목을 구입하거나 예외적인 개수만큼 구입하는 경향이 있음을 알 수 있으므로 흥미롭다고 할 수 있다.

또한 두번째 척도로 제안하는 $diff^{2nd}$는 다음과 같다.

이 척도는 한 노드를 이루는 항목리스트의 각 항목들에 대해 그 노드에 속하는 총 고객수로 나눈 값을 부모노드에서의 값과 비교함으로써, 고객 1인당 각 항목의 평균 구입수와 부모노드에서의 고객 1인당 각 항목의 평균 구입수와의 차이의 합을 나타낸다. 따라서 $diff^{2nd}$값이 0에 가까울수록 평균에 가까우며, 값이 클수록 흥미로움이 커진

다고 할 수 있다.

[정의] $\mathrm{diff}^{2nd} = \sum \left| \left(\frac{n_{l-1,j,k}}{C_{l-1,j}} - \frac{n_{l,k}}{C_{l,j}} \right) \right|$

$n_{l-1,j,k}$: l-1 레벨 j 번째 노드의 k 라는 항목의 property(ex : 항목개수)

$C_{l,i}$: l 레벨의 i번째 노드의 customer수

$C_{l-1,j}$: l-1 레벨의 j번째 노드의customer수

위와 같은 척도 diff(diff^{1st} ,diff^{2nd})를 척도로 하여 기대치와 다른 형태의 항목구매리스트를 지닌 노드를 찾는 알고리즘은 다음과 같다.

```
Algorithm()
{
SIG ← ∅
L ← the last level index
Adjust(L) ;
While(L>1){
// To union its child nodes which are item
// lists of level L for each level L-1 node
  Union(L) ;
// To prun for each level L-1 node
  Adjust(L-1);
  For(j=1;j ≤# of nodes level L-1;j++)
  {
    while(exist child node k of node j){
       if  diff(L,j,k)  ≥ threshold T
    then  SIG ←node k  }
    }
   L ← L-1
  }
  return SIG ;
}
// To prun for each level L-1 node
function Adjust(L){
 while(exist node i of level L)
     prun item of node i whose number per one
  person < s
}
// To union its child nodes which are item
// lists of level L for each level L-1 node
function Union(L){
   while(exist node i of level L){
      items of node i ← union of child nodes of each node
           i  }
}
```

즉, 가장 하위레벨에서 시작하여 각 고객들의 항목리스트에서 각 고객당 평균 구입수가 s개 미만인 것은 제거하여 항목리스트를 조정한다. 이렇게 해서 형제 노드들에 대한 항목리스트들이 만들어지면, 이 항목리스트들로 한단계 일반화된 부모노드의 항목리스트를 만든다. 그 과정은, 우선 항목리스트를 이루는 모든 항목들을 합쳐서 부모노드의 후보(candidate)항목리스트를 만든후, 이를 구성하는 항목들 중 각 고객당 평균 구입수가 s개 미만인 것은 제거하여 부모노드의 항목리스트를 만든다. 그 후 각 노드에 대해 부모노드와 diff^{1st} 이나 diff^{2nd}의 척도를 적용하여 한계치 T가 넘는 값을 갖는 노드를 SIG에 넣는다.

이를 상위레벨로 올라가며 되풀이하여 루트(root)까지 도달하면 종료하고, SIG를 리턴한다.

리턴값 SIG는 한계치 T가 넘는 diff값을 가진 노드의 집합이다. 즉, 주위에 비해 예외적인 성질을 지닌 노드가 리턴된다.

이 알고리즘은 하위레벨부터 시작하여 상위레벨로 올라가는 bottom-up 형태를 취하고 있다. 그러므로 가장 하위레벨에서 데이터베이스를 한번만 스캔(scan)하면 상위레벨로 올라가며 데이터베이스를 다시 스캔할 필요가 없으므로 시간 및 공간적으로 많은 이익이 있다.

이러한 알고리즘을 diff^{1st} 에 대한 한계치(threshold) T를 80으로 주고, diff^{2nd} 에 대한 한계치 T를 12로 주고 각각을, 앞의 예([그림7])에 적용시켰을 경우에 [그림7]에 해당되는 결과는 다음과 같다.

node	(a)	(b)	(c)
diff^{1st}	32.2	75.6	87.8
diff^{2nd}	4.0	10.0	16.0

SIG ={ c }

이 결과로, 앞의 예에서는 (c) 노드에 속하는 고객들이 같은 레벨에 속하는 고객들의 평균에 비해 예외적인 흥미로운 항목들을 구입하거나 흥미로운 개수만큼 특정 항목을 구입하는 경향이 있음을 알 수 있다.

그리고 이러한 방법은 한계치에 따라 결과가 많이 나오거나 적게 나올 수 있으므로, 최적화를 위해 결과에 따라 적절한 피드백(feedback)을 통해 다시 보다 알맞은 결과를 얻을 수 있을 것이다. 또한 알고리즘을 좀더 확장하여 흥미로운 구매경향을 보이는 고객뿐만 아니라, 그에 따른 흥미로운 항목들을 찾아내고, 그 항목들을 좀더 세분화하여 좀더 다양한 결과들을 얻을 수 있을 것이다.

3.4 개념계층에서의 예외적인 연관규칙 계층 탐색

3.4.1 예외적인 연관규칙 계층 탐색

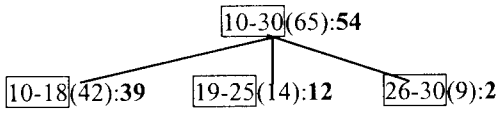
이번 장에서는 데이터의 특성에 대해 항목에 대한 구입 개수 즉, 항목의 빈도수 대신 특정 연관규칙 'A→C'의 지지도(support)를 이용하여, 이 연관규칙에 대해 비슷한 성향을 보일 거라고 예상되는 고객들의 특성에 따른 개념계층을 이용하여, 기대와는 다르게 예외적으로 이 연관규칙이 제대로 적용되지 않는 고객들의 집단 등 데이터의 흥미로운 부분을 발견하는 방법을 제시한다.

적용하려는 연관규칙이 '바지→벨트'라고 하자. 또한 이 연관규칙에 대해 비슷한 성향을 보일 거라고 예상되는 나이에 대한 개념계층을 고려해보자.

이 개념계층의 가장 하위 레벨인 말단노드(leaf node)에는 각 고객들에 대한 데이터가 있다. 각 고객들의 항목구매 리스트가 '바지→벨트'를 만족하면, 즉, 항목 바지와 벨트를 동시에 구매했으면 연관규칙 지지도(support)에 1을, 그렇지 않으면 0을 부여한다.

그 다음, 자식노드들에서 나타난 연관규칙의 지지도를 모두 합한 것으로 각 부모 노드의 연관규칙의 지지도를 나타낸다. 또한 자식노드들의 고객수의 총합도 유지한다.

다음 [그림 8]은 이러한 개념계층의 일부분이다. 괄호안의 숫자는 그 노드에 해당하는 고객수를 나타내고 그 옆의 굵은 글씨체의 숫자는 연관규칙의 지지도를 나타낸다.. 예를 들어, 10-18세에 속하는 고객은 42명이며, 이 노드의 연관규칙 지지도는 39이다.



[그림8]나이에 따른 개념계층 및 연관규칙 지지도

계층에 따른 연관규칙 지지도가 예상과는 달리 비슷한 계층 사이에서 기대치를 벗어나는 예외적인 흥미로운 계층을 탐사하기 위해, 여기서도 3.1장에서 쓰였던 $lift_{li}$를 이용한다. 여기서 $lift_{li}$은 l 레벨의 i노드에 대해 그 노드에서 나타나는 연관규칙 지지도의 기대값 $E(n_{li})$에 대한 실제값의 비율이다. 주의할 것은, 이 $lift_{li}$식에 나타나는 n값은 연관규칙에 적용하기 위해서 n=sup(A→C)가 될 것이다.

$$lift_{li} = \frac{n_{li}}{E(n_{li})}$$

앞의 3.1에서와 유사하게 l 레벨의 i노드에 대해 그 노드에서 나타나는 연관규칙 지지도의 기대값 $E(n_{li})$ 은 다음과 같이 나타낼 수 있다.

$$E\left(\sup\left(A \rightarrow C\right)_{li}\right) = \sup\left(A \rightarrow C\right)_{l-1,j} \times \frac{c_{li}}{c_{l-1,j}}$$

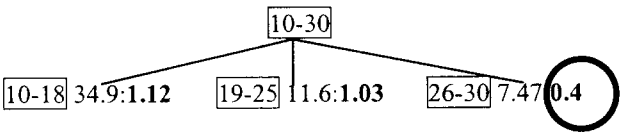
$\sup(A \rightarrow C)_{l-1,j}$: l-1 레벨의 j번째 노드(부모노드)의 연관규칙 지지도

c_{li} : l 레벨의 i노드에 대해 그 노드에 해당되는 고객의 수

$c_{l-1,j}$: l-1레벨의 노드에 j노드(부모노드)에 대해 그 노드에 해당되는 고객의 수

3.4.2 척도 적용결과

제시한 척도 $lift_{li}$에 의한 결과값을 살펴보면 다음의 [그림 9]와 같다. 이 그림은 앞에서 제시한 $lift_{li}$를 이용하여 [그림8]에 대해 10-18, 19-25, 26-30세에 해당하는 고객들의 '바지→벨트'라는 연관규칙 지지도에 대한 기대값과 $lift_{li}$ (굵은숫자로 표시) 를 얻은 값이다.



[그림 9] 연관규칙 만족 기대값과 lift 결과

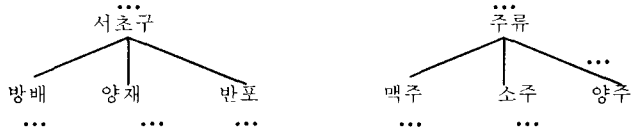
10-18세와 19-25세에 속하는 고객들의 $lift_{li}$값은 1과의 차이가 각각 0.12, 0.03으로 1에 가까운 값을 보여 기대한 것과 비슷한 연관규칙 지지도를 지녔으나, 26-30세에 속하는 고객들의 $lift_{li}$값은 0.4로 1과 부정적으로(negative) 0.6의 차이를 지녀, 10대에서 30대 사이에 속하는 고객들의 '바지→벨트'라는 연관규칙을 만족하는 구매패턴의 평균에 비해 낮은 구매패턴을 지녔음을 알 수 있다.

3.4.2 항목 개념계층에 따른 연관규칙 기대치

이 절에서는 고객에 대한 개념계층과 구입항목에 대한 개

념계층에 따른 연관규칙이 존재할 때, 일부 연관규칙과 개념계층을 이용하여 다른 연관규칙의 지지도(support), 신뢰도(confidence),리프트(lift)등에 대한 기대치를 얻을 수 있음을 보인다.

[그림2]와 같이 고객의 한 특성인 지역에 관한 개념계층과 항목인 주류에 관한 개념계층을 고려해보자.



(a) (b)

[그림 2] 지역에 대한 개념계층(a)와 항목에 대한 개념계층(b)

R1 : 방배동→주류
R2 : 방배동→맥주
R3 : 서초구→맥주

[그림2]에 따른 연관규칙

또한 이 두 개념계층 사이에 다음과 같이 R1,R2,R3의 연관규칙이 있다고 하자.

이러한 경우, 연관규칙 R2에 대한 지지도(support), 신뢰도(confidence),리프트(lift)등에 대한 기대치를 연관규칙 R1이나 R3를 통해 다음과 같이 얻을 수 있다.

$$E(sup(R2)) = E(sup(R1)) \times \frac{n(\text{맥주})}{n(\text{주류})}$$

$$E(sup(R2)) = E(sup(R3)) \times \frac{C(\text{방배})}{C(\text{서초})}$$

$$E(conf(R2)) = E(conf(R1)) \times \frac{n(\text{맥주})}{n(\text{주류})}$$

$$E(conf(R2)) = E(conf(R3)) \times \frac{C(\text{방배})}{C(\text{서초})}$$

$$E(lift(R2)) = E(lift(R1)) \times \frac{sup(R2)}{sup(R1)} \times \frac{n(\text{주류})}{n(\text{맥주})}$$

즉, 이러한 식으로 항목의 개념계층을 이용하여 각 항목에 관한 연관규칙의 기대치를 다양한 방법으로 얻을 수 있다. 앞으로 이에 추가하여 이러한 척도들의 관계와 3장의 연구결과로 얻어진 흥미로움을 통해, 주어진 데이터들에 대한 연관규칙을 좀더 효율적으로 발견하는 방법을 연구할 예정이다.

4. 결론

지금까지 살펴본 바와 같이, 이 논문에서는 데이터베이스에 내재되어 있는 데이터의 분포나 구조적인 정보 등 데이터의 배경지식을 반영한 개념계층을 이용하여, 이 개념계층을 통해 기대되는 특정 성질에 대해 예상외로 비슷한 패턴을 보이지 않는 흥미로운 부분 데이터를 척도들을 이용하여 찾아내는 방법 및 척도를 몇 가지 경우로 나누어 제시하였다. 즉, 인터넷 상의 쇼핑몰에 대한 고객의 데이터에 대해, 각 고객의 구매리스트가 단일 항목으로 이루어졌을 때(3.1)와 이를 좀더 일반화하여 각 구매리스트가 복수개의 항목으로 이루어졌을 때로 나누어 흥미로운 부분 데이터를 찾는 방법 및 척도를 제안하고 역동적인(dynamic) 피드백을 통한 최적화 및 확장 가능성에 대해 논하였다.(3.2)

그리고 데이터의 흥미로움(interestingness)을 측정할 때 항목 구매의 빈도수 대신 특정 연관규칙의 지지도를 사용하여 기대와는 다르게 연관규칙이 적용되지 않는 고객 집단을 발견하는 방법을 제시하였다.(3.3)또한 항목들의 개념계층과 데이터마이닝의 결과인 연관규칙을 통해 얻을 수 있는 기대치를 여러가지 척도들의 관계를 통해 모색해보았다.(3.4)

이 논문의 연구를 인터넷 쇼핑몰에 응용하면, 비슷한 항목들을 구매하는 경향이 있다고 추측되는 고객 개념계층을 통해, 이에 예외적인 항목을 구매하는, 관심을 가질만한 고객층이나 각 개인을 찾아낼 수 있을 것이다. 이렇게 개념계층에 따라 일반적인 고려대상에서 제외되었던 소수 그룹 및, 특정 항목들의 특성들도 찾아냄으로써, 쇼핑몰에서는 좀더 차별화된 개별 마케팅 전략이나 틈새 시장 마케팅 전략을 마련할 수 있을 것이고, 좀더 나아가서는 보험회사 등에서는 사기행각을 발견하기 위한 근거로서의 활용방안도 모색하는 등 다양한 분야로의 응용이 가능할 것이다.

이 논문은, 지금까지 개념계층과 흥미로움(interestingness)에 대한 각각의 논문을 정리하여 개념계층을 이용하여 데이터에서의 흥미로운 부분을 탐색하는 시도를 처음으로 하여 앞으로 유사한 연구를 하고자 하는 연구자에게 좋은 참고가 되리라 생각한다. 여기서는 예외적인 부분 데이터를 탐색하기 위한 척도로 기존의 척도를 제안하거나 새로 정의하였는데, 이 척도들은 상황에 따라 지금까지 제시된 척도들 중 가장 알맞은 척도를 사용하거나 다시 새로 개발하는 것을 고려할 수 있을 것이다.

앞으로, 이 논문에 제시된 척도나 방법외에도 좀더 연구하여 추가적인 방안들을 찾을 것이며, 이 논문에서 얻어지는 흥미로운 부분들과 개념계층에 대한 연구를 계속하여 연관규칙을 좀더 효율적으로 찾기위한 방법을 모색할 예정이다.

또한 개념계층은 지식 공학자나 영역 전문가, 또는 사용자에 의해 다양하게 주어질 수 있으므로, 고객의 데이타에 대한 개념계층을 어떻게 구성하느냐에 따라 달라지는 결과를 살펴보는 것도 흥미로울 것으로 생각된다.

5.참고문헌

[1] 김성민,남도원,이동하,김성훈,이전영, "데이터베이스에서의 지식탐사를 위한 개념계층의 자동생성", 제2차 데이터 마이닝 워크샵 SIGDM '99, Vol.15, No.1, 1999, pp.152-167.

[2] J.Han and Y.Fu, "Dynamic Generation and Refinement of Concept Hierarchies for Knowledge Discovery in Databases", KDD'94, July 1994, pp.157-168

[3]Yijun Lu, "Concept Hierarchy in Data Mining: Specification, Generation and Implementation", Ms.Thesis, Simon Fraser University,1997

[3] Sergey Brin, Rajeev Motwani, "Dynamic Itemset counting and Implication Rules for Market Basket Data", ACM'97, 1997.

[4] Roberto J.Bayardo Jr. , "Mining the Most Interesting Rules", ACM SIGKDD'99, 1999, pp145-154.

[5] 김성민, 이동하, 남도원, 이전영, "관계형 데이터베이스에서의 지식탐사를 위한 개념 계층의 제어기법," '98 한국전문가시스템학회 추계학술대회 논문집, 1998년 12월,pp83-90,

[6] 이동하, 남도원, 김성민, 이전영, "관계형 데이터베이스에서의 연관 규칙 탐사를 위한 객체 일반화 트리의 구성," '98 한국전문가시스템학회 추계학술대회 논문집, 1998년 12월,pp91-96,

[7]이동하, 이전영, "지능 질의처리를 위한 지식 탐사 시스템의 설계," '97 동계 데이터베이스 학술대회 논문집, 제13권 2호, 1997년 2월, pp.217-222.

[8]남도원, 이동하, 이전영, "관계형 데이터베이스에서의 연관규칙 탐사," '97 봄 한국정보과학회 학술발표논문집(B), 1997년 4월, pp.217-222.

[9]Dong-Ha Lee, Dong-Yal Seo, Kang-Sik Moon, Ji-Sook Chang, Do-Won Nam, Jeon-Young Lee, "Discovery and Application of Inter-Classes Patterns in Database," *Proceedings of the 8th International Workshop on Database and Expert Systems Applications*, Sept. 1-2, 1997, Toulouse, France, IEEE Computer Society, pp.326-331.

[10] R.Srikant and R. Agrawal, "Mining Generalized Association Rules," Proc. 21th Int. Conf. VLDB, 1995, pp407-419

[11]D.-Y. Seo, D.-H. Lee, K.-M.Lee, J.-Y.Lee, "Discovery of Schema Information from a Forest of Selectively Labeled Ordered Trees,", Workshop on Management of Semistructured Data (In Conjunction with ACM PODS/SIGMOD '97),Tucson, Arizona, May 16,1997, pp.54-59

[12] Sridhar Ramaswamy, Rajeev Rastogi, Kyuseok Shim, "Efficient Algorithms for Mining Outliers from Large Data sets",SIGMOD2000,2000.