(19) 대한민국특허청(KR)
(12) 공개특허공보(A)

(11) 공개번호 10-2013-0070220
(43) 공개일자 2013년06월27일

(51) 국제특허분류(Int. Cl.)
G06F 17/16 (2006.01)
(21) 출원번호 10-2011-0137442
(22) 출원일자 2011년12월19일
심사청구일자 2011년12월19일

(71) 출원인
이화여자대학교 산학협력단
서울특별시 서대문구 이화여대길 52 (대현동, 이화여자대학교)
(72) 발명자
용환승
서울특별시 서대문구 대현동 이화여자대학교 아산공학관 331호
이인기
서울특별시 서대문구 대현동 이화여자대학교 아산공학관 331호
(74) 대리인
김인철

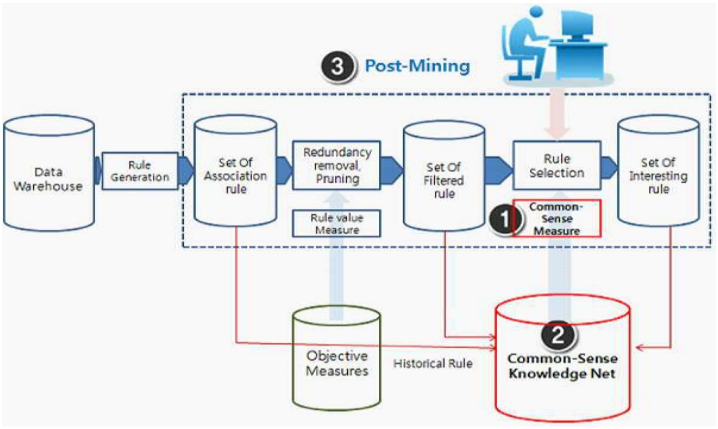
전체 청구항 수 : 총 5 항

(54) 발명의 명칭 **데이터 마이닝 결과 연관규칙에 대한 상식 기반 후 처리방법**

(57) 요 약

본 발명은 데이터 마이닝 결과 연관규칙에 대한 상식 기반 후 처리방법에 관한 것으로서, 데이터 마이닝되어 생성된 연관규칙과 상식의 유사 정도를 측정하는 방법으로서, (a) 데이터 마이닝 결과 생성된 연관규칙들이 각각 벡터 공간 모델에 기반하여 전건(antecedent)과 후건(consequent) 및 연관관계의 컴포넌트로 구성되는 연관관계 벡터로 정의되고, 도메인의 상식 지식망에 포함된 상식이 벡터 공간 모델에 기반하여 전개념(antecedent-concept)과 후개념(consequent-concept) 및 관계(reration)의 컴포넌트로 구성되는 상식벡터로 정의되는 단계; 및 (b) 상기 연관관계벡터와 상식벡터의 대응하는 컴포넌트 벡터 사이에 대한 코사인 거리 유사도를 계산하여 연관규칙과 상식간의 유사도를 산출하는 단계를 포함한다. 이에 의하여, 상식 기반의 의미론적 접근방법을 새롭고 유용한 지식을 찾기 위한 데이터 마이닝 기법에 적용함으로써, 사용자에게 새롭고 유용한 연관규칙들만을 제시할 수 있는 지능화된 데이터 마이닝 결과를 제공할 수 있는 데이터 마이닝 결과 연관규칙에 대한 상식 기반 후 처리방법을 제공할 수 있다.

대 표 도 - 도1



특허청구의 범위

청구항 1

데이터 마이닝되어 생성된 연관규칙과 상식의 유사 정도를 측정하는 방법으로서,

(a) 데이터 마이닝 결과 생성된 연관규칙들이 각각 벡터 공간 모델에 기반하여 전건(antecedent)과 후건(consequent) 및 연관관계의 컴포넌트로 구성되는 연관관계벡터로 정의되고, 도메인의 상식 지식망에 포함된 상식이 벡터 공간 모델에 기반하여 전개념(antecedent-concept)과 후개념(consequent-concept) 및 관계(reration)의 컴포넌트로 구성되는 상식벡터로 정의되는 단계; 및

(b) 상기 연관관계벡터와 상식벡터의 대응하는 컴포넌트 벡터 사이에 대한 코사인 거리 유사도를 계산하여 연관규칙과 상식간의 유사도를 산출하는 단계를 포함하는 데이터 마이닝 결과 연관규칙에 대한 상식 기반 후 처리방법.

청구항 2

제1항에 있어서, 상기 (a) 단계는

상기 연관규칙들과 상식에 대한 가중치를 부여하는 단계를 더 포함하는 것을 특징으로 하는 데이터 마이닝 결과 연관규칙에 대한 상식 기반 후 처리방법.

청구항 3

제2항에 있어서,

상기 연관규칙에 대한 가중치의 부여는

상기 도메인의 상식 지식망에 포함된 상식에 대한 도메인 데이터 집합인

$$AV = \{av_1, av_2, av_3, \ldots av_i\}$$

AV( ), 이때 a와 v는 각각 데이터 집합의 attribute와 value를 의미함)를 정의하고, 하기의 수학식 1 및 수학식 2를 이용하여 상기 연관규칙에 대한 차원 가중치를 부여하는 것을 특징으로 하는 데이터 마이닝 결과 연관규칙에 대한 상식 기반 후 처리방법.

[수학식 1]

$$w(r, av_i) = tf(av_i, r) * idf(av_i, R)$$

[수학식 2]

$$idf(av_i, R) = \log \frac{|R|}{|av_i \in r|}$$

이때, 수학식 1과 수학식 2에서 $w(r, av_i)$ 는 각 연관규칙의 차원 가중치를, $tf(av_i, r)$ 은 연관규칙에서 av_i 의 발생 빈도를 의미하고, $idf(av_i, R)$

는 av_i 항목이 연관규칙들에서 얼마나 자주 나타나는지를 나타내는 것으로, |R|는 가중치로 총 연관규칙의 개수이고, $|av_i \in r|$ 는 av_i 를 포함하는 연관규칙의 개수를 각각 의미한다.

청구항 4

제2항에 있어서,

상기 상식에 대한 가중치의 부여는

상기 도메인의 상식 지식망에 포함된 상식에 대한 데이터 집합인 AV($AV = \{av_1, av_2, av_3, \ldots av_i\}$, 이때 a와 v는 각각 데이터 집합의 attribute와 value를 의미함)에 상기 AV 항목에 그 동의어와 상위어 및 하위어를 포함하는 데이터 집합인 AVSH($AVSH = \{avsh_{11}, avsh_{12}, avsh_{13}, \ldots avsh_{ij}\}$)를 정의하고, 하기의 수학식 3과 수학식 4를 이용하여 상기 상식에 대한 차원 가중치를 부여하는 것을 특징으로 하는 데이터 마이닝 결과 연관규칙에 대한 상식 기반 후 처리방법.

[수학식 3]

$$w(cs, avi) = \sum_{j=1}^{n} tf(avshij, cs) * idf(avi, R) * sim(avi, avshij)$$

[수학식 4]

$$sim(av_i, avsh_{ij}) = \frac{2 * depth(lcs(av_i, avsh_{ij}))}{depth(av_i) + depth(avsh_{ij})}$$

이때, 상기 [수학식 3]와 [수학식 4]에 있어서, $w(r, av_i)$ 는 상식 cs의 차원 가중치이고, $sim(av_i, avsh_{ij})$ 는 av_i 와 av_i 의 시맨틱 키워드로 확장된 $avsh_{ij}$ 와의 유사도를 정의하는 것이며(이때, 유사도는 워드넷의 컨셉 hierarchy를 사용하여 측정할 수 있는 것임), lcs는 두 키워드들이 의미를 공유하는 최하포섭개념(lowest common subsume)이고, depth는 컨셉 hierarchy에서 키워드들의 루트 노드로부터의 깊이를 의미한다.

청구항 5

제4항에 있어서,

상기 (b) 단계에서의 상기 연관규칙과 상식간의 유사도는 하기의 수학식 5와 수학식 6을 이용하여 산출되는 것을 특징으로 하는 데이터 마이닝 결과 연관규칙에 대한 상식 기반 후 처리방법.

[수학식 5]

$$\mathbf{sim}(\vec{r}, \vec{cs}) = \alpha * \mathbf{sim}(\vec{r_{ant}}, \vec{cs_{ant}}) + \beta * \mathbf{sim}(\vec{r_{con}}, \vec{cs_{con}}) + (1 - \alpha - \beta) * \mathbf{sim}(\vec{r_{rel}}, \vec{cs_{rle}})$$

[수학식 6]

$$\mathbf{sim}(\vec{r_{ant}}, \vec{cs_{ant}}) = \frac{\sum_{k=1}^{i} w(r,av_k) * w(cs,av_k)}{\sqrt{w(r,av_k)^2} * \sqrt{w(cs,av_k)^2}}$$

이때, 상기 [수학식 5]와 [수학식 6]에서의 $\mathbf{sim}(\vec{r}, \vec{cs})$ 는 규칙 r과 상식 cs의 유사도인 common-sense 척도이고, α와 β는 상기 연관규칙과 상식간의 유사도를 계산하는 과정에서 영향을 주는 연관규칙의 구조에 따라 실험에 의해 정의된 각 컴포넌트(전건(antecedent), 후건(consequent), 연관관계(relation))의 가중치이다.

명 세 서

기 술 분 야

[0001] 본 발명은 데이터 마이닝 후처리 방법에 관한 것으로, 상식 기반의 의미론적 접근방법을 새롭고 유용한 지식을 찾기 위한 데이터 마이닝 기법에 적용함으로써, 사용자에게 새롭고 유용한 연관규칙들만을 제시할 수 있는 지능화된 데이터 마이닝 결과를 제공할 수 있는 데이터 마이닝 결과 연관규칙에 대한 상식 기반 후 처리방법에 관한 것이다.

배 경 기 술

[0002] 본 발명은 데이터 마이닝된 연관규칙에 대한 후처리 과정에 관한 것이다.

[0003] 데이터 마이닝(Data Mining)이란, 많은 데이터 가운데 숨겨져 있는 유용한 상관관계, 즉 연관규칙을 발견하여, 미래에 실행 가능한 정보를 추출해 내고 의사 결정에 이용하는 과정을 말한다.

[0004] 즉, 데이터 마이닝은 각 데이터의 상관관계를 인공 지능 기법을 통해 자동적으로 찾아 주는 과정을 말한다. 이때, 데이터 웨어하우스와 데이터 마트는 사용자가 원하는 테이블들을 미리 만들어 놓고 이를 꺼내 볼 수 있도록 한다. 예컨대, 비를 좋아하는 사람에 대한 데이터가 있고 색깔에 대한 선호도와 관계된 데이터가 있다면 이 둘의 관계를 밝혀 내는 기능을 수행한다.

[0005] 즉, 정확히 수치화하기 힘든 데이터 간의 연관을 찾아 내는 역할을 한다.

[0006] 데이터마이닝이 소개된 이후로 대용량 데이터로부터 패턴과 연관규칙들을 발견하기 위해 보다 효율적이고 확장성 있는 탐사 알고리즘들이 제안되어 왔고 한 차원 진보된 연구 성과를 보여주었다. 특히 연관규칙 마이닝은 상업, 통신, 보험, 생명공학등과 같은 다양한 분야에서 폭넓게 활용되어왔다.

[0007] 그러나 이와 같은 성능 향상과 더불어 실제 상업 데이터베이스들의 크기와 차원은 크게 증가하여, 수천 개 또는 수백만 개의 연관규칙이 생성되게 되었다. 이렇게 생성된 많은 연관규칙과 패턴들은 잠재적인 문제로 여전히 해결되지 못하고 있는 문제가 있다. 즉, 사용자들은 중복되거나 유용하지 않은 규칙들이 포함된 많은 연관규칙들 속에서 유용한 지식을 발견해 내기 위해 많은 시간과 노력을 필요로 하게 되는 문제가 있다.

[0008] 종래의 데이터 마이닝에 관한 발명들은 추출된 패턴들의 가치평가는 고려하지 않고 정확한 모델을 생성하는데 집중되고 있다.

[0009] 본 발명의 배경이 되는 기술은 대한민국 등록특허공보 제10-0919684호 내지 대한민국 공개특허공보 2003-0086038호 등에 개시되어 있으며, 상술한 문제를 여전히 내포하고 있다.

발명의 내용

해결하려는 과제

[0010] 상술한 문제점을 해결하기 위해 안출된 본 발명의 목적은 데이터 마이닝의 결과 생성된 연관규칙이 얼마나 상식(Common-Sense Knowledge)에 가까운지를 평가하는 척도를 제안하여 사용자에게 새롭고 유용한 연관규칙들만을 전달할 수 있는 데이터 마이닝 결과 연관규칙에 대한 상식 기반 후 처리방법을 제공하기 위한 것이다.

발명의 효과

[0011] 상기한 바와 같이 본 발명에 따르면, 상식 기반의 의미론적 접근방법을 새롭고 유용한 지식을 찾기 위한 데이터 마이닝 기법에 적용함으로써, 사용자에게 새롭고 유용한 연관규칙들만을 제시할 수 있는 지능화된 데이터 마이닝 결과를 제공할 수 있는 데이터 마이닝 결과 연관규칙에 대한 상식 기반 후 처리방법을 제공할 수 있다.

도면의 간단한 설명

[0012] 도 1은 본 발명에 의한 데이터 마이닝 결과 연관규칙에 대한 상식 기반 후 처리방법에 대한 블록도이다.

도 2는 Common-Sense 척도 평가결과의 예시도이다.

발명을 실시하기 위한 구체적인 내용

[0013] 이하, 본 발명에 따른 데이터 마이닝 결과 연관규칙에 대한 상식 기반 후 처리방법에 대하여 첨부된 도면을 참조하여 상세히 설명한다.

[0014] 데이터 마이닝의 과거 연구들이 추출된 패턴들의 가치평가는 고려하지 않고 정확한 모델을 생성하는데 집중하였다면, 최근 연구들은 이러한 문제에 초점을 맞추어 다양한 형태의 유용성 척도(Interestingness Measures)를 정의하는 것으로 확장해 가고 있다. 유용한 지식은 이미 알려져 있지 않고 이전에 볼 수 없었던 새로운 지식을 의미한다.

[0015] 본 발명은 이러한 문제점들을 해결하기 위한 방법으로 데이터 마이닝의 결과로 생성된 연관규칙이 얼마나 상식(Common-Sense Knowledge)에 가까운지를 평가하는 척도를 제시하기 위한 것이다.

[0016] 유용한 지식을 평가하기 위한 척도들의 최종 목표는 사람이 직접 그 지식이 유용한지를 판단했을 때와 가장 유사한 결과를 보여주는 것이다. 컴퓨터가 인간이 가지고 있는 지식들을 표현하고 습득하여 활용할 수 있는 방법을 찾기 위한 많은 노력들이 있었다. 특히 상식은 일반적으로 사회적 의사소통에 생략되어 있으므로 컴퓨터가 올바른 텍스트와 지식들을 이해하기 위해서는 많은 양의 상식 정보를 필요로 하게 된다. Open Mind Common Sense는 MIT Media lab에서 진행하고 있는 인공지능 프로젝트로 상식을 수집하기 위한 웹사이트를 제공하여 수천 명의 일반인들로부터 입력된 1,000,000의 상식들을 지식데이터베이스로 관리하고 있다[4]. 이 같은 방법은 컴퓨터가 가장 효과적인 방법으로 일반인으로부터 상식에 관한 지식을 습득하기 위한 방법이라고 볼 수 있다. ConceptNet은 상식을 기반으로 만들어진 시맨틱 네트워크로써 컴퓨터가 인간만이 인지할 수 있었던 상식에 접근할 수 있는 방법을 보여준다. 이러한 선진 기술들은 다양한 애플리케이션에 적용 가능하며 활발한 연구가 진행되고 있다. 특히 본 발명은 이러한 상식 기반의 의미론적 접근방법을 새롭고 유용한 지식을 찾기 위한 데이터 마이닝 기법에 적용함으로써 지능화된 데이터 마이닝 결과를 보여주고자 한다.

[0017] 본 발명에 따른 데이터 마이닝 결과 연관규칙에 대한 상식 기반 후 처리방법은 데이터 마이닝되어 생성된 연관규칙과 상식의 유사 정도를 측정하는 방법으로서, 데이터 마이닝 결과 생성된 연관규칙들이 각각 벡터 공간 모델에 기반하여 전건(antecedent)과 후건(consequent) 및 연관관계의 컴포넌트로 구성되는 연관규칙벡터로 정의되고, 도메인의 상식 지식망에 포함된 상식이 벡터 공간 모델에 기반하여 전개념(antecedent-concept)과 후개념(consequent-concept) 및 관계(reration)의 컴포넌트로 구성되는 상식벡터로 정의되는 (a) 단계(S10) 및 (b) 상기 연관규칙벡터와 상식벡터의 대응하는 컴포넌트 벡터 사이에 대한 코사인 거리 유사도를 계산하여 연관규칙과 상식간의 유사도를 산출하는 (b) 단계(S20)를 포함한다.

[0018] 이때, 상기 (a) 단계는, 상기 연관규칙들과 상식에 대한 가중치를 부여하는 단계를 더 포함할 수 있다.

[0019] 여기서, 상기 연관규칙에 대한 가중치 부여는 도메인 데이터집합의 attribute와 value로 구성되는 항목 집합인

$$\mathbf{AV} = \{\mathbf{av_1}, \mathbf{av_2}, \mathbf{av_3}, \ldots \mathbf{av_i}\}$$

AV{ }(이때, a와 v는 각각 데이터 집합의 attribute와 value를

의미함)를 정의하고, 하기의 수학식 1 및 수학식 2를 이용하여 상기 연관규칙에 대한 차원 가중치를 부여하는 것을 특징으로 한다. 즉, 집합 AV는 데이터마이닝을 위한 데이터웨어하우스의 attribute와 value로 구성되는 항목집합으로, 연관규칙들을 구성하는 attribute 와 value를 규칙의 키워드로 정의하고 각 키워드의 가중치를 부여하기 위한 것이다.

[0020] [수학식 1]

$$w(r, av_i) = tf(av_i, r) * idf(av_i, R)$$

[0021]

[0022] 상기 수학식 1에서 $w(r, av_i)$ 는 각 연관규칙의 차원 가중치를, $tf(av_i, r)$ 은 연관규칙에서 av_i 의 발생 빈도를 의미한다.

[0023] [수학식 2]

$$idf(av_i, R) = \log \frac{|R|}{|av_i \in r|}$$

[0024]

[0025] 상기 관계식 5에서 $idf(av_i, R)$ 는 av_i 항목이 연관규칙들에서 얼마나 자주 나타나는지를 나타내는 것으로, |R|는 가중치로 총 연관규칙의 개수이고, $|av_i \in r|$ 는 av_i 를 포함하는 연관규칙의 개수를 각각 의미한다.

[0026] 또한, 상기 상식에 대한 가중치의 부여는, 상기 AV 항목에 그 동의어와 상위어 및 하위어를 포함하는 데이터 집합인 AVSH($AVSH = \{avsh_{11}, avsh_{12}, avsh_{13}, \ldots avsh_{ij}\}$)를 정의하고, 하기의 수학식 3과 수학식 4를 이용하여 상기 상식에 대한 차원 가중치를 부여하는 것을 특징으로 하는 데이터 마이닝 결과 연관규칙에 대한 상식 기반 후 처리방법.

[0027] [수학식 3]

$$w(cs, av_i) = \sum_{j=1}^{n} tf(avsh_{ij}, cs) * idf(av_i, R) * sim(av_i, avsh_{ij})$$

[0028]

[0029] [수학식 4]

$$sim(av_i, avsh_{ij}) = \frac{2 * depth(lcs(av_i, avsh_{ij}))}{depth(av_i) + depth(avsh_{ij})}$$

[0030]

[0031] 이때, 상기 [수학식 3]과 [수학식 4]에 있어서, $w(r, av_i)$ 는 상식 cs의 차원 가중치이고, $sim(av_i, avsh_{ij})$

는 av_i와 av_i의 시맨틱 키워드로 확장된 $avsh_{ij}$

[0032] 와의 유사도를 정의하는 것이며(이때, 유사도는 워드넷의 컨셉 hierarchy를 사용하여 측정할 수 있는 것임), lcs는 두 키워드들이 의미를 공유하는 최하포섭개념(lowest common subsume)이고, depth는 개념 hierarchy에서 키워드들의 루트 노드로부터의 깊이를 의미한다.

[0033] 상기 (b) 단계에서의 상기 연관규칙과 상식간의 유사도는 하기의 수학식 5와 수학식 6을 이용하여 산출된다.

[0034] [수학식 5]

[0035] $$\mathbf{sim}(\vec{r},\vec{cs}) = \alpha * \mathbf{sim}(\vec{r_{ant}}, \vec{cs_{ant}}) + \beta * \mathbf{sim}(\vec{r_{con}}, \vec{cs_{con}}) + (1-\alpha-\beta) * \mathbf{sim}(\vec{r_{rel}}, \vec{cs_{rle}})$$

[0036] [수학식 6]

[0037] $$\mathbf{sim}(\vec{r_{ant}}, \vec{cs_{ant}}) = \frac{\sum_{k=1}^{i} w(r,av_k) * w(cs,av_k)}{\sqrt{w(r,av_k)^2} * \sqrt{w(cs,av_k)^2}}$$

[0038] 이때, 상기 [수학식 5] 에서의 $\mathbf{sim}(\vec{r},\vec{cs})$ 는 규칙 r과 상식 cs의 유사도인 common-sense 척도이고, α와 β는 상기 연관규칙과 상식간의 유사도를 계산하는 과정에서 영향을 주는 연관규칙의 구조에 따라 실험에 의해 정의된 각 컴포넌트(전건(antecedent), 후건(consequent), 연관관계(relation))의 가중치이다.

[0039] 우선 Common-Sense 척도에 관해 설명한다.

[0040] **Common-Sense 척도**

[0041] 도 1은 데이터 마이닝에서의 상식 기반 후처리 방법으로 데이터 마이닝 후 처리단계에서 선별된 연관규칙들을 상식 지식망의 누적된 지식을 기반으로 측정하고 평가하여 최종적으로 유용한 지식들을 사용자에게 전달하도록 하는 것을 나타내기 위한 블록도이다.

[0042] 연관규칙의 유용성 척도 중에서도 주관적 척도는 데이터와 관련된 사용자의 지식에 대한 접근이 중요하다. 패턴이나 연관규칙들은 사용자가 가지고 있는 지식으로 예측할 수 없을 때 더욱 흥미롭고 유용하기 때문에 사용자의 지식 표현이 핵심이라고 할 수 있다.

[0043] 본 연구에서는 사용자의 지식 표현을 위해 상식(Common-Sense Knowledge)을 활용한다. 상식은 일반적으로 사회적 의사소통을 위해 필요한 기본 지식으로, 예측가능하고 이미 사람들에게 알려져 있어 흥미롭지 못한 지식으로 표현할 수 있다. 우리가 발견하고자 하는 유용한 지식과 상식과는 상반된 특징을 보여주므로 Common-Sense 척도가 유용한 연관규칙을 선별하는데 효과적으로 사용될 수 있음을 알 수 있다.

[0044] 상식을 수집하기 위한 방법은 웹사이트를 제공하여 수천 명의 일반인들로부터 입력된 수많은 상식들을 지식데이터베이스로 관리하는 것이다. 상식기반 정보처리의 대표적인 연구자인 MIT Lab의 Push Sing는 상식을 기반으로 한 시맨틱 네트워크인 ConceptNet을 발표하였다[16]. 이와 같은 연구는 WordNet, Cyc Project 와 비교되는데 구조적으로는 WordNet과 유사하나 의미론적 관계가 훨씬 다양하고 유연하며 Cyc Project 보다는 쉽게

[0045] 사용할 수 있고 보다 많은 concept들을 다루고 있다. 또한, 노드 단계의 추론 기능들(contextual neighborhoods, analogy and projection)과 문서 단계의 추론 기능들(topic gisting, disambiguation/classification, novel-concept identification, affect sensing)을 제공한다.

[0046] Common-Sense 척도는 연관규칙이 얼마나 상식에 가까운지를 수치로 표현해 주는 의미론적 거리 척도이다. Common-Sense 척도를 정의하기 위해서는 연관규칙 마이닝을 통해 생성된 연관규칙과 사람들의 상식이 얼마나 유사한지를 계산할 수 있는 기법을 사용한다. 도 1은 데이터 마이닝에서의 상식 기반 후처리 기법으로 본 연구에서 제안하는 ① Common-Sense 척도는 ③과 같이 데이터 마이닝의 후 처리 단계에서 선별된 연관규칙들을 ② 상식 지식망의 누적된 지식을 기반으로 측정하고 평가하여 최종적으로 유용하지 않은 지식들을 제외한 연관규칙들

을 사용자에게 전달하도록 한다(도 1 참조).

[0047] 다음은 본 발명에 의한 데이터 마이닝 결과 연관규칙에 대한 상식 기반 후 처리방법을 위한 연관규칙과 상식에 대한 정의 단계에 대해 설명한다.

[0048] **연관규칙과 상식의 정의**

[0049] 먼저, 연관규칙(rule)에 관해 살펴본다.

[0050] 데이터 마이닝의 결과 생성된 연관규칙은 형식 $X \Rightarrow Y$의 함축적인 식이다. 이 연관규칙은 지지도 (Support)와 신뢰도(Confidence)로 측정될 수 있다.

[0051] 지지도는 한 연관규칙이 주어진 데이터 집합에 얼마나 자주 적용할 수 있는지를 결정하고, 신뢰도는 연관규칙에 의해 만들어지는 추론의 확실성을 측정한다. 연관규칙은 전건(antecedent)와 후건(consequent)으로 구성된다.

[0052] 연관규칙에 의해 만들어진 추론은 반드시 인과 관계를 함축하는 것은 아니고, 연관규칙의 전건과 후건에 있는 항목들 사이의 강한 동시 발생관계를 제시한다.

[0053] 따라서, 연관규칙은 아래의 관계식 1에 나타낸 바와 같이 벡터 공간 모델에 기반하여 전건(antecedent, ant), 후건(consequent, con), 연관관계(relation, rel)와 같은 세 개의 컴포넌트로 구성되고, 연관규칙은 아래의 관계식 2에 나타낸 바와 같이 상기 전건(ant), 후건(con), 연관관계(rel)의 세 컴포넌트로 구성되는 벡터로 정의할 수 있다.

[0054] [관계식 1]

[0055] **Rule** : {age=old} → {skin=wringkling}

[0056] **antecedent relation consequent**

[0057] [관계식 2]

[0058] $$\mathbf{Rule}\ \bar{r} = \{\ \overrightarrow{r_{ant}},\ \overline{r_{con}},\ \overrightarrow{r_{rel}}\ \}$$

[0059] 다음은, 상식(Common Sense, cs)에 관해 살펴본다.

[0060] MIT Media lab의 ConceptNet은 통합된 자연어 처리 엔진인 MontyLingua가 사람들로부터 입력 받은 상식 텍스트 문장들을 형태소, 품사 분석, 파싱, 추출단계와 표준화 단계를 거쳐 VSOO(verb-subject-object-object) 프레임들로 구성하게 된다. 이러한 프레임들은 concept과 concept의 relation으로 표현되는데 concept은 그래프의 노드로 보다 높은 차원의 프레임의 결합을 허용할 뿐만 아니라, relation은 20개 이상의 의미적 관계를 표현해준다[참고문헌 C. Havasi, R. Speer, J. Alonso, "ConceptNet 3: a Flexible, Multilingual Semantic Network for Common Sense Knowledge", Proceedings of Recent Advances in Natural Languges Processing, 2007. 참고]. 이렇게 저장된 상식 정보들은 ConceptNet API를 통해 사용 가능 하므로 본 연구에서는 이러한 상식들을 기반으로 변환, 저장하여 도메인의 상식 지식망을 필요한 형태로 표현할 수 있도록 정의한다. 도메인의 상식은 전-개념(antecedent concept), 후-개념(consequent-concept)과 관계 (relation), 상세관계(sub-relation)를 속성으로 갖는 데이터베이스에 저장될 수 있으며 Made-of, IsA, Partof 등과 같은 20개의 관계를 세부적으로 3000개의 상세관계로 정의할 수 있다.

[0061] 상식은 아래의 관계과 관계식 3에 나타낸 바와 같이, 연관규칙과의 유사도 측정을 위해 벡터 공간 모델을 기반으로 전-개념(antecedent-concept, ant), 관계(relation, rel), 후-개념(consquent-concept, con) 세 개의 컴포넌트로 구성되고, 상식은 아래의 관계식 4에 나타낸 바와 같이 전-개념(antecedent-concept), 관계(relation), 후-개념(consquent-concept) 세 개의 컴포넌트로 구성되는 벡터로 정의할 수 있다.

[0062] [관계식 3]

[0063] **Common-Sense Knowledge** :

[0064] Old aging has the effect of wrinkling the skin.

[0065] concept1 relation concept2

[0066] [관계식 4]

[0067] Common-Sense $\overrightarrow{cs}\overrightarrow{cs}$ = {$\overrightarrow{cs}_{ant}\overrightarrow{cs}_{ant}$, $\overrightarrow{cs}_{con}\overrightarrow{cs}_{con}$, $\overrightarrow{cs}_{rel}\overrightarrow{cs}_{rel}$}

[0068] 다음은 상기 정의된 연관규칙과 상식을 유사도 측정 기법을 사용하여 Common-sense 척도로 구현하는 방법에 대해 설명한다.

[0069] **Common-Sense 척도 구현(Common-Sense Measures, CS-M)**

[0070] 사용될 유사도 측정 기법은 텍스트 마이닝 기법에서 사용되는 벡터 공간 모델(vector space model)에서의 코사인 거리(cosine distance) 유사도 기법이다. 그러나 연관규칙의 항목들이 데이터베이스의 속성값으로 표현되는 특징, 상식의 개념(concept)이 단순 어휘 항목이 아닌 고차원 합성어인 개념(concept)으로 표현되는 특징들이 고려되어지기 위해서는 구조화된 지식을 위한 새로운 유사도 기법이 필요하다. 또한, 사람의 지식이 텍스트로 표현 되면서 발생하는 다양성의 문제를 고려했을 때 단순한 텍스트 유사도 기법보다는 시맨틱한 접근법이 필요하다. 즉, 동의어, 상위어, 하위어를 고려한 시맨틱 기반의 유사도 기법을 제안하다. 따라서 본 연구에서 제안한 Common-Sense 척도 측정 방법은 상기 연관규칙과 상식을 벡터를 구현하고, 그 다음 Common-Sense 척도를 계산하도록 한다.

[0071] 먼저, 벡터 공간 모델에 기반하여 데이터 마이닝의 결과로 추출된 연관규칙과 상식의 각 컴포넌트를 벡터로 정의하고, 도메인 데이터집합의 attribute와 value로 구성되는 항목 집합인 $\mathbf{AV} = \{\mathbf{av_1}, \mathbf{av_2}, \mathbf{av_3}, \ldots \mathbf{av_i}\}$ 를 정의한다.

[0072] 상기 집합 AV는 데이터마이닝을 위한 데이터웨어하우스의 attribute와 value로 구성되는 항목집합으로, 연관규칙들을 구성하는 attribute 와 value를 연관규칙의 키워드로 정의하고 각 키워드의 가중치를 부여하기 위한 것이다.

[0073] 연관규칙 r의 차원 가중치 $\mathbf{w(r, av_i)}$ 는 아래의 수학식 1와 수학식 2에 의해 산출된다.

수학식 1

[0074] $$\mathbf{w(r, av_i) = tf(av_i, r) * idf(av_i, R)}$$

수학식 2

[0075] $$\mathbf{idf(av_i, R) = \log \frac{|R|}{|av_i \in r|}}$$

[0076] 이때, 수학식 1과 수학식 2에서 $\mathbf{w(r, av_i)}$ 는 각 연관규칙의 차원 가중치를, $\mathbf{tf(av_i, r)}$ 은 연관규칙에서

av_i의 발생 빈도를 의미하고, $idf(av_i, R)$는 av_i 항목이 연관규칙들에서 얼마나 자주 나타나는지를 나타내는 것으로, |R|는 가중치로 총 연관규칙의 개수이고, $|av_i \in r|$는 av_i를 포함하는 연관규칙의 개수를 각각 의미한다.

[0077] 이때, 상식을 벡터 공간 모델에 정의하기 위한 차원은 연관규칙과 동일하게 AV 항목이다. 그러나 고차원 합성어로 이루어진 상식의 concept은 항목 집합인 연관규칙에서 키워드의 발생 빈도를 기반으로 가중치를 계산한 것과는 다른 기법이 필요하다. 상식의 concept을 벡터로 표현하기 위해서는 시맨틱 키워드 확장(Semantic keyword extension) 기법을 정의하고 사용한다. 이와 같은 기법은 벡터 공간 모델의 단점인 키워드 매칭 문제를 해결하고 의미론적 유사도를 측정하기 위함이다.

[0078] 이때, 상기 시맨틱 키워드 확장 기법은 상식의 concept에 매칭되는 키워드뿐만 아니라 워드넷의 의미관계 집합으로부터 키워드의 동의어, 상위어, 하위어를 적용하여 의미론적 유사도를 가중치에 반영하여 계산하는 기법으로, 아래의 관계식 6 내지 관계식 8을 이용하여 상식(cs)의 차원 가중치인 $w(cs, av_i)$를 산출한다.

[0079] 이때, AVSH는 상기 AV 항목에 그 동의어와 상위어 및 하위어를 포함하는 데이터 집합으로,

$$AVSH = \{avsh_{11}, avsh_{12}, avsh_{13}, \ldots avsh_{ij}\}$$

으로 정의된다.

수학식 3

[0080] $$w(cs, avi) = \sum_{j=1}^{n} tf(avshij, cs) * idf(avi, R) * sim(avi, avshij)$$

수학식 4

[0081] $$sim(av_i, avsh_{ij}) = \frac{2 * depth(lcs(av_i, avsh_{ij}))}{depth(av_i) + depth(avsh_{ij})}$$

[0082] 상기 수학식 3과 수학식 4에서 $sim(av_i, avsh_{ij})$는 av_i와 av_i의 시맨틱 키워드로 확장된 $avsh_{ij}$와의 유사도를 정의한다. 여기서 유사도는 워드넷의 컨셉 hierarchy를 사용하여 측정할 수 있으며 lcs는 두 키워드들이 의미를 공유하는 최하포섭개념(lowest common subsume)이고, depth는 개념 hierarchy에서 키워드들의 루트 노드로부터의 깊이를 의미한다.

[0083] 예를 들어 설명하면 다음과 같다.

[0084] 연관규칙 sex=female à cat=clothing 과 높은 유사도를 나타낸 아래의 표 1의 상식의 경우, 가중치를 구하기 위해서는 female과 clothing과 매칭되는 의미 관계집합 AVSH의 항목들에 대한 유사도를 계산하여 가중치에 반영시킨다. 즉 woman은 female의 동의어로 clothes는 clothing의 동의어로 가중치에

계산되어 의미론적 유사도가 높은 상식이 높은 유사도를 보여주게 된다.

표 1

[0085]

Common-Sense	concept	relation	concept
You are likely to find a woman in a clothing store.	A woman	You are likely to find {1} in {2}	a clothing store
A young woman wants new clothes.	A young woman	{1} wants {2}	new clothes

[0086] 상기의 방법에 의해 연관규칙과 상식의 관계(Relation) 유사도를 측정하고, 연관규칙의 전건과 후건의 동시 발생관계와 상식(Common-Sense)으로부터 3000개의 상세 관계(relation)사이의 유사도를 분석하여 0에서 1사이의 값을 갖는 의미 관계 유사도를 측정한다.

[0087] 다음은 Common-Sense 척도 계산에 대해 설명한다.

[0088] 연관규칙(r)과 상기(cs)의 유사도인 Common-Sense 척도는 하기의 수학식 5와 수학식 6에 의해 산출된다.

[0089] 연관규칙의 각 컴포넌트와 상식의 각 컴포넌트 벡터간의 각도를 비교하는 코사인 거리 유사도를 계산하여 하나의 연관규칙과 특정 상식간의 유사도를 계산할 수 있다. 코사인 값이 0인 경우는 연관규칙과 상식 벡터가 직각인 경우로, 둘 간의 유사성이 전혀 없음을 의미한다. 연관규칙과 관련된 다양한 상식들과의 유사도 중에서 최대

값을 그 연관규칙의 Common-Sense 척도로 정의한다. 즉, 하기 수학시 5의 $\mathrm{sim}(\vec{r},\vec{cs})$ $\mathrm{sim}(\vec{r},\vec{cs})$인 규칙 r과 상식 cs의 유사도인 common-sense 척도를 계산하기 위해서 전건(antecedent), 후건(consquent), 연관관계(relation)와 같은 세 개의 컴포넌트의 유사도를 계산하여야 한다. 이때, 하기의 α와 β는 상기 연관규칙과 상식간의 유사도를 계산하는 과정에서 영향을 주는 연관규칙의 구조에 따라 실험에 의해 정의된 각 컴포넌트(전건(antecedent), 후건(consequent), 연관관계(relation))의 가중치이다. 상기 각 컴포넌트의 유사도는 하기 수학식 6으로 계산된다.

수학식 5

[0090]
$$\mathrm{sim}(\vec{r},\vec{cs}) = \alpha * \mathrm{sim}(\overrightarrow{r_{ant}}, \overrightarrow{cs_{ant}}) + \beta * \mathrm{sim}(\overrightarrow{r_{con}}, \overrightarrow{cs_{con}}) + (1-\alpha-\beta) * \mathrm{sim}(\overrightarrow{r_{rel}}, \overrightarrow{cs_{rle}})$$

수학식 6

[0091]
$$\mathrm{sim}(\overrightarrow{r_{ant}}, \overrightarrow{cs_{ant}}) = \frac{\sum_{k=1}^{i} w(r,av_k) * w(cs,av_k)}{\sqrt{w(r,av_k)^2} * \sqrt{w(cs,av_k)^2}}$$

[0092] 다음은 본 발명에 의한 데이터 마이닝 결과 연관규칙에 대한 상식 기반 후 처리 방법을 이용한 실험 및 그 실험결과에 대해 설명한다.

[0093] 본 발명에 의한 데이터 마이닝 결과 연관규칙에 대한 상식 기반 후 처리방법인 Common-Sense 척도를 측정하기 위한 연관규칙과 상식의 유사도 기법을 평가하기 위해서 Microsoft paraphrase phrase tables와 5800 문장 페어로 구성된 corpus 데이터를 사용하였다.

[0094] 상기 Microsoft paraphrase phrase tables와 5800 문장 페어로 구성된 corpus 데이터를 평가에 이용하기 위해서 concept과 relation의 형태로 변환하여 상술한 CS-M(Common-Sense Measures) 유사도 기법을 적용하였고, 비교 평가를 위해 다른 텍스트 유사도 기법을 함께 평가하였다. Semantic(W&P)는 The Wu and Palmer의 대표적인

시맨틱 유사도 기법으로 워드넷 텍사노미를 활용하여 두 개의 컨셉 사이 깊이를 계산하여 유사도를 측정하는 기법이다. Lexical matching은 매칭되는 어휘의 수를 기반으로 계산하였고, Cosine/tf.idf는 전통적인 텍스트 유사도 기법으로 tf.idf 가중치를 코사인 유사도 기법으로 계산하였다. 표 2는 데이터에서 제공하는 문장 페어의 유사도 값인 0과 1을 기반으로 각 기법의 accuracy, precision, recall, f-measure를 측정한 결과이다. Concept과 Relation으로 구조화된 지식, 데이터항목으로 구성된 연관규칙과 같은 지식들의 유사도를 평가하는데 본 연구에서 제안한 기법이 가장 좋은 결과를 보여주었다.

[0095] 유용한 연관규칙을 찾기 위한 Common-Sense 척도의 실험을 위해 Knowledge Discovery and Data Mining Data set의 인터넷 쇼핑몰 데이터를 사용하였다. 9개의 속성을 갖는 10,268건의 데이터를 Apriori 알고리즘을 수행하고 객관적 척도인 confidence, lift, conviction, leverage 를 적용하여 200건의 베스트 규칙들을 선별하였다. 인터넷 쇼핑몰 도메인의 상식 지식망을 구성하여 연관규칙들의 Common-Sense 척도를 측정하고 4인의 도메인 사용자들을 대상으로 평가 기준표를 통해 연관규칙이 얼마나 유용한 지식인지를 평가하도록 하였다. 평가 항목은 Expected/Unexpected, Actionable/Unactionable, Noble, Accuracy로 구성되며 최종 연관규칙의 유용성은 0에서 1사이의 수로 평가되었다. 표 2에서 cs=1은 가장 상식에 가까운 지식을 의미하며, u=1은 가장 유용한 지식으로 평가된 것이다. 유용한 지식들은 상식 척도가 0에 가까웠으며, 전체 데이터의 71%가 일치하는 결과를 보여주었고, 상식척도가 0.5보다 큰 연관규칙들의 88%가 유용하지 않은 지식으로 평가되었다.

표 2

[0096]

CS척도 사용자평가	cs=0	0<cs<0.5	0.5<cs<1	cs=1
u=1	7%	0%	0%	0%
0.5<u<1	10%	3%	9%	0%
0<u<0.5	7%	0%	57%	0%
u=0	1%	0%	2%	4%

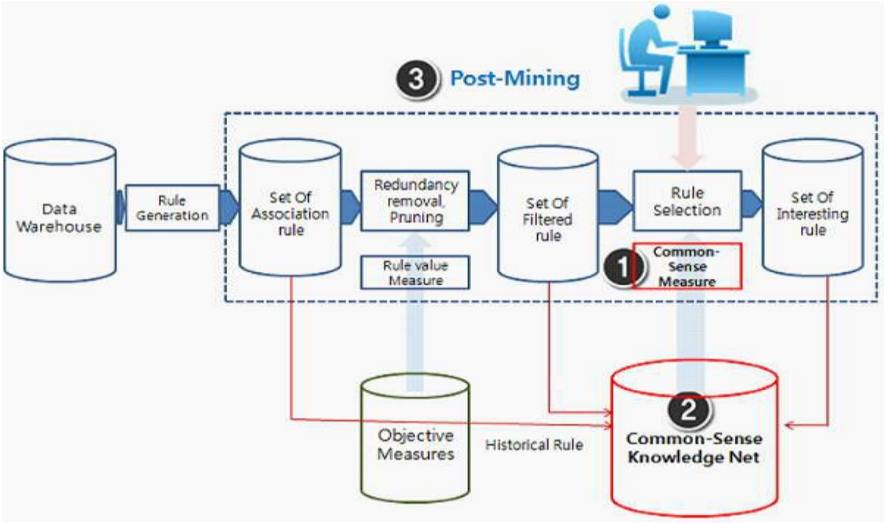
[0097] 그림 2는 Common-Sense 척도의 accuracy, precision, recall, F-measure를 보여주는 그래프이다.

[0098] 본 발명에서는 데이터 마이닝에서 생성되는 대량의 연관규칙과 패턴 문제를 해결하기 위한 후처리 기법으로 상식을 기반으로 하는 Common-Sense 척도를 정의하고 구현하여 유용한 지식들만을 선별할 수 있도록 하였다. Common-Sense 척도는 시맨틱 네트워크 형태의 상식 데이터베이스를 이용하여 연관규칙이 얼마나 상식에 가까운지를 수치로 표현해 주는 척도이다. 유사도 측정 기법은 벡터 공간 모델(vector space model)에서의 코사인 거리(cosine distance) 유사도 기법을 기반으로 시맨틱 키워드 확장 기법을 사용하여 구조화된 지식간의 시맨틱 유사도를 계산할 수 있는 방법을 제시하였다. 실험 결과 본 연구에서 제안한 유사도 기법이 구조화된 지식간의 유사도 계산에서 높은 정확성을 보여주었으며, 기존의 객관적 척도들에 의해 제거되지 못했던 불필요한 연관규칙들을 효과적으로 선별할 수 있음을 보여주었다.

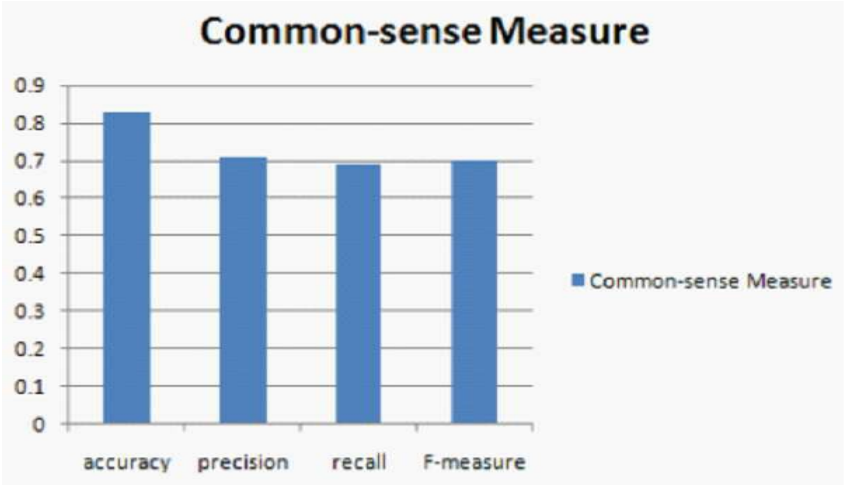
[0099] 이상에서 대표적인 실시예를 통하여 본 발명에 대하여 상세하게 설명하였으나, 본 발명이 속하는 기술 분야에서 통상의 지식을 가진 자는 상술한 실시예에 대하여 본 발명의 범주에서 벗어나지 않는 한도 내에서 다양한 변형이 가능함을 이해할 것이다. 그러므로 본 발명의 권리 범위는 설명된 실시예에 국한되어 정해져서는 안 되며 후술하는 특허청구범위뿐만 아니라 이 특허청구범위와 균등한 것들에 의하여 정해져야 한다.

도면

도면1


3 Post-Mining
Data Warehouse
Rule Generation
Set Of Association rule
Redundancy removal, Pruning
Rule value Measure
Set Of Filtered rule
Rule Selection
1 Common-Sense Measure
Set Of Interesting rule
Objective Measures
Historical Rule
2 Common-Sense Knowledge Net

도면2


Common-sense Measure
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
accuracy
precision
recall
F-measure
Common-sense Measure