(19) 대한민국특허청(KR)
(12) 등록특허공보(B1)

(45) 공고일자 2019년12월24일
(11) 등록번호 10-2059263
(24) 등록일자 2019년12월18일

(51) 국제특허분류(Int. Cl.)
G06Q 30/02 (2012.01) *G06N 99/00* (2019.01)
(52) CPC특허분류
G06Q 30/0269 (2013.01)
G06N 20/00 (2019.01)
(21) 출원번호 10-2018-0024413
(22) 출원일자 2018년02월28일
심사청구일자 2018년02월28일
(65) 공개번호 10-2019-0109631
(43) 공개일자 2019년09월26일
(56) 선행기술조사문헌
조영성 외 2인, "RFM기법과 k-means 기법을 이용한 개인화 추천시스템의 개발", 한국컴퓨터정보학회, 2012년 6월, 제17권, 제6호, pp.163-172. 1부.*
진병운 외 2인, "RFM 기법과 연관성 규칙을 이용한 개인화된 전자상거래 추천시스템", 한국컴퓨터정보학회, 2010년 12월, 제15권, 제12호, pp.227-235.*
KR101438050 B1
US20170372338 A1
*는 심사관에 의하여 인용된 문헌

(73) 특허권자
세종대학교산학협력단
서울특별시 광진구 능동로 209 (군자동, 세종대학교)
(72) 발명자
신동일
서울특별시 강남구 압구정로 347, 26동 1207호(압구정동, 한양아파트)
신동규
서울특별시 강남구 언주로 201, 1903호(도곡동, 에스케이리더스뷰)
윤현수
서울특별시 동작구 상도로62길 81-1, 105호(상도동, 씨티아모리움)
(74) 대리인
양성보

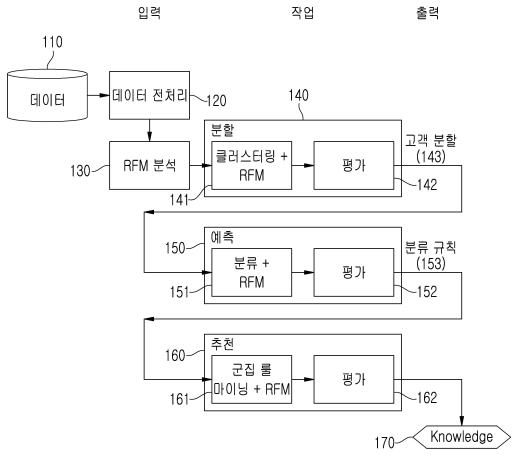
전체 청구항 수 : 총 7 항 심사관 : 이원재

(54) 발명의 명칭 **기계학습 알고리즘을 적용한 RFM(Recency, Frequency, Monetary) 분석 방법 및 시스템**

(57) 요 약

기계학습 알고리즘을 적용한 RFM 분석 방법 및 시스템이 개시된다. 일 실시예에 따른 데이터 분석 시스템에 의하여 수행되는 데이터 분석 방법은, 데이터에 대한 RFM 분석을 수행하는 단계; 상기 RFM 분석을 수행함에 따라 획득된 RFM 값에 기초하여 상기 데이터를 적어도 하나 이상의 클러스터로 분류하는 단계; 상기 클러스터의 RFM 값을 통하여 분류 규칙을 생성하는 단계; 및 상기 생성된 분류 규칙으로부터 추천 목록을 제공하는 단계를 포함할 수 있다.

대 표 도 - 도1



이 발명을 지원한 국가연구개발사업

과제고유번호 1711055356

부처명 과학기술정보통신부

연구관리전문기관 정보통신기술진흥센터

연구사업명 ICT유망기술개발지원

연구과제명 IOT 기반의 고객데이터 수집 및 자동인식을 통한 빅데이터분석 클라우드 고객센싱 서비스 시스템 구축

기 여 율 1/1

주관기관 (주)소프트자이온

연구기간 2017.05.01 ~ 2018.04.30

명 세 서

청구범위

청구항 1

데이터 분석 시스템에 의하여 수행되는 데이터 분석 방법에 있어서,

데이터에 대한 RFM 분석을 수행하는 단계;

상기 RFM 분석을 수행함에 따라 획득된 RFM 값에 기초하여 상기 데이터를 적어도 하나 이상의 클러스터로 분류하는 단계;

상기 클러스터의 RFM 값을 통하여 분류 규칙을 생성하는 단계; 및

상기 생성된 분류 규칙으로부터 추천 목록을 제공하는 단계

를 포함하고,

상기 데이터에 대한 RFM 분석을 수행하는 단계는,

상기 데이터에 제품의 거래 정보, 제품의 식별정보, 제품을 구매한 고객 정보, 제품의 주문 날짜, 제품의 수량, 제품의 주문 금액, 제품의 유형을 포함하는 데이터에 대한 식별 정보에 기초하여 데이터 전처리 과정을 수행하고, 상기 전처리 과정이 수행된 데이터에 대한 RFM 분석을 수행하는 단계

를 포함하고,

상기 RFM 분석을 수행함에 따라 획득된 RFM 값에 기초하여 상기 데이터를 적어도 하나 이상의 클러스터로 분류하는 단계는,

상기 전처리 과정이 수행된 데이터에 대한 RFM 분석을 수행함에 따라 획득된 RFM 값에 대하여 k-means++ 알고리즘을 사용하여 유사한 RFM 값을 가진 고객 세그먼트를 탐색하고, 상기 유사한 RFM 값을 가진 고객 세그먼트들이 할당된 적어도 하나 이상의 클러스터로 분류하는 클러스터링 과정을 수행하고, 상기 클러스터링 과정을 수행함에 따라 표준편차 또는 SSE(Sum of Squared Error)를 사용하여 클러스터의 품질을 평가하고 클러스터 중심에서 벗어나지 않는 클러스터를 보장하고 다른 클러스터와의 분리를 보장하는 클러스터링을 평가하는 과정을 수행하는 단계

를 포함하고,

상기 클러스터의 RFM 값을 통하여 분류 규칙을 생성하는 단계는,

상기 클러스터에 포함된 RFM 값과 상기 데이터를 분석하기 위한 별도의 정보를 사용하여 사용자의 행동을 예측하기 위한 C4.5 Decision Tree 알고리즘에 기초하여 분류 규칙을 생성하고, 상기 생성된 분류 규칙에 대하여 유효성 검사를 수행하는 단계

를 포함하고,

상기 C4.5 알고리즘은 우선 divide-and-conquer 전략을 사용하여 트리를 성장시킨 다음 트리를 잘라(Cut)내고, 트리를 잘라냄에 따라 모든 속성의 전체 엔트로피와 정보 이득을 계산하고, 계산된 정보 이득이 가장 높은 속성이 결정을 내리기 위해 선택되는 것을 포함하고,

상기 유효성 검사는, n-fold 교차 검증 기법을 사용하고, 데이터 집합을 n(n은 자연수) 개의 부분 집합으로 분할하고, 분할하는 과정을 n(n은 자연수)번 반복함에 따라 n 개의 서브 세트 중 하나가 테스트 세트로 사용되고 다른 n-1 개의 서브 세트가 함께 학습 세트를 형성하여 n 번의 시도에 대한 평균 오차가 계산되는 것을 포함하고,

상기 생성된 분류 규칙으로부터 추천 목록을 생성하는 단계는,

상기 분류 규칙과 ARM(Association Rule Mining)에 기반하여 상기 클러스터와 관련된 규칙들 중 기 설정된 기준에 의하여 소정의 규칙을 추출하고, 상기 클러스터의 정보, 상기 데이터와 연관된 사용자 식별 정보 및 상기 클

러스터에 포함된 데이터 간의 연관성을 식별하고, 상기 식별된 연관성에 기반하여 순위를 정렬하여 제품을 추천하는 단계

를 포함하고,

상기 추출된 소정의 규칙의 연관성은, Lift($\mathrm{Lift}(R)=\frac{P(XY)}{P(X)P(Y)}$)와 Loevinger($\mathrm{Loevinger}(R)=1-\frac{P(X)P(-Y)}{P(X-Y)}$)의 기준으로 평가되고, 상기 Lift와 Loevinger의 기준은 항목 집합 X, Y 및 규칙 R에 대해 정의되며 X->Y 형식의 설명 패턴이며 X는 연관 규칙의 조건부이고, Y는 결과적인 부분이고, Lift 기준은 X가 발생할 때 Y를 갖는 확률 계수를 나타내고, Loevinger 기준은 결과 부분 Y를 만족하지 않을 확률에 따라 규칙의 중심 신뢰를 표준화하고, Lift와 Loevinge 값이 클수록 더 강한 연관성을 나타내는,

데이터 분석 방법.

청구항 2

제1항에 있어서,

상기 RFM 분석을 수행함에 따라 획득된 RFM 값에 기초하여 상기 데이터를 적어도 하나 이상의 클러스터로 분류하는 단계는,

R 속성, F 속성 및 M 속성을 포함하는 RFM 속성 각각을 기 설정된 기준으로 분할하고, 상기 데이터를 상기 분할된 각각의 속성에 기반하여 RFM 분석을 수행함으로써 상기 데이터에 대한 RFM 값을 획득하는 단계

를 포함하는 데이터 분석 방법.

청구항 3

제2항에 있어서,

상기 RFM 분석을 수행함에 따라 획득된 RFM 값에 기초하여 상기 데이터를 적어도 하나 이상의 클러스터로 분류하는 단계는,

상기 데이터의 RFM 값에 따라 K-Means ++ 알고리즘을 사용하여 상기 데이터를 적어도 하나 이상의 클러스터로 분류함에 따라 상기 데이터의 클러스터링 결과를 평가하는 단계

를 포함하는 데이터 분석 방법.

청구항 4

삭제

청구항 5

삭제

청구항 6

삭제

청구항 7

데이터 분석 시스템에 의하여 수행되는 데이터 분석 방법을 실행시키기 위하여 컴퓨터 판독 가능한 기록매체에 저장된 컴퓨터 프로그램에 있어서,

데이터에 대한 RFM 분석을 수행하는 단계;

상기 RFM 분석을 수행함에 따라 획득된 RFM 값에 기초하여 상기 데이터를 적어도 하나 이상의 클러스터로 분류하는 단계;

상기 클러스터의 RFM 값을 통하여 분류 규칙을 생성하는 단계; 및

상기 생성된 분류 규칙으로부터 추천 목록을 제공하는 단계

를 포함하고,

상기 데이터에 대한 RFM 분석을 수행하는 단계는,

상기 데이터에 제품의 거래 정보, 제품의 식별정보, 제품을 구매한 고객 정보, 제품의 주문 날짜, 제품의 수량, 제품의 주문 금액, 제품의 유형을 포함하는 데이터에 대한 식별 정보에 기초하여 데이터 전처리 과정을 수행하고, 상기 전처리 과정이 수행된 데이터에 대한 RFM 분석을 수행하는 단계

를 포함하고,

상기 RFM 분석을 수행함에 따라 획득된 RFM 값에 기초하여 상기 데이터를 적어도 하나 이상의 클러스터로 분류하는 단계는,

상기 전처리 과정이 수행된 데이터에 대한 RFM 분석을 수행함에 따라 획득된 RFM 값에 대하여 k-means++ 알고리즘을 사용하여 유사한 RFM 값을 가진 고객 세그먼트를 탐색하고, 상기 유사한 RFM 값을 가진 고객 세그먼트들이 할당된 적어도 하나 이상의 클러스터로 분류하는 클러스터링 과정을 수행하고, 상기 클러스터링 과정을 수행함에 따라 표준편차 또는 SSE(Sum of Squared Error)를 사용하여 클러스터의 품질을 평가하고 클러스터 중심에서 벗어나지 않는 클러스터를 보장하고 다른 클러스터와의 분리를 보장하는 클러스터링을 평가하는 과정을 수행하는 단계

를 포함하고,

상기 클러스터의 RFM 값을 통하여 분류 규칙을 생성하는 단계는,

상기 클러스터에 포함된 RFM 값과 상기 데이터를 분석하기 위한 별도의 정보를 사용하여 사용자의 행동을 예측하기 위한 C4.5 Decision Tree 알고리즘에 기초하여 분류 규칙을 생성하고, 상기 생성된 분류 규칙에 대하여 유효성 검사를 수행하는 단계

를 포함하고,

상기 C4.5 알고리즘은 우선 divide-and-conquer 전략을 사용하여 트리를 성장시킨 다음 트리를 잘라(Cut)내고, 트리를 잘라냄에 따라 모든 속성의 전체 엔트로피와 정보 이득을 계산하고, 계산된 정보 이득이 가장 높은 속성이 결정을 내리기 위해 선택되는 것을 포함하고,

상기 유효성 검사는, n-fold 교차 검증 기법을 사용하고, 데이터 집합을 n(n은 자연수) 개의 부분 집합으로 분할하고, 분할하는 과정을 n(n은 자연수)번 반복함에 따라 n 개의 서브 세트 중 하나가 테스트 세트로 사용되고 다른 n-1 개의 서브 세트가 함께 학습 세트를 형성하여 n 번의 시도에 대한 평균 오차가 계산되는 것을 포함하고,

상기 생성된 분류 규칙으로부터 추천 목록을 생성하는 단계는,

상기 분류 규칙과 ARM(Association Rule Mining)에 기반하여 상기 클러스터와 관련된 규칙들 중 기 설정된 기준에 의하여 소정의 규칙을 추출하고, 상기 클러스터의 정보, 상기 데이터와 연관된 사용자 식별 정보 및 상기 클러스터에 포함된 데이터 간의 연관성을 식별하고, 상기 식별된 연관성에 기반하여 순위를 정렬하여 제품을 추천하는 단계

를 포함하고,

상기 추출된 소정의 규칙의 연관성은, Lift($\mathrm{Lift}(R)=\frac{P(XY)}{P(X)P(Y)}$)와 Loevinger($\mathrm{Loevinger}(R)=1-\frac{P(X)P(-Y)}{P(X-Y)}$)의 기준으로 평가되고, 상기 Lift와 Loevinger의 기준은 항목 집합 X, Y 및 규칙 R에 대해 정의되며 X->Y 형식의 설명 패턴이며 X는 연관 규칙의 조건부이고, Y는 결과적인 부분이고, Lift 기준은 X가 발생할 때 Y를 갖는 확률 계수를 나타내고, Loevinger 기준은 결과 부분 Y를 만족하지 않을 확률에 따라 규칙의 중심 신뢰를 표준화하고, Lift와 Loevinge 값이 클수록 더 강한 연관성을 나타내는,

컴퓨터 판독 가능한 기록매체에 저장된 컴퓨터 프로그램.

청구항 8

데이터 분석 시스템에 있어서,

데이터에 대한 RFM 분석을 수행하는 분석부;

상기 RFM 분석을 수행함에 따라 획득된 RFM 값에 기초하여 상기 데이터를 적어도 하나 이상의 클러스터로 분류하는 분류부;

상기 클러스터의 RFM 값을 통하여 분류 규칙을 생성하는 규칙 생성부; 및

상기 생성된 분류 규칙으로부터 추천 목록을 제공하는 추천부

를 포함하고,

상기 분석부는,

상기 데이터에 제품의 거래 정보, 제품의 식별정보, 제품을 구매한 고객 정보, 제품의 주문 날짜, 제품의 수량, 제품의 주문 금액, 제품의 유형을 포함하는 데이터에 대한 식별 정보에 기초하여 데이터 전처리 과정을 수행하고, 상기 전처리 과정이 수행된 데이터에 대한 RFM 분석을 수행하는 것을 포함하고

상기 분류부는,

상기 전처리 과정이 수행된 데이터에 대한 RFM 분석을 수행함에 따라 획득된 RFM 값에 대하여 k-means++ 알고리즘을 사용하여 유사한 RFM 값을 가진 고객 세그먼트를 탐색하고, 상기 유사한 RFM 값을 가진 고객 세그먼트들이 할당된 적어도 하나 이상의 클러스터로 분류하는 클러스터링 과정을 수행하고, 상기 클러스터링 과정을 수행함에 따라 표준편차 또는 SSE(Sum of Squared Error)를 사용하여 클러스터의 품질을 평가하고 클러스터 중심에서 벗어나지 않는 클러스터를 보장하고 다른 클러스터와의 분리를 보장하는 클러스터링을 평가하는 과정을 수행하는 것을 포함하고,

상기 규칙 생성부는,

상기 클러스터에 포함된 RFM 값과 상기 데이터를 분석하기 위한 별도의 정보를 사용하여 사용자의 행동을 예측하기 위한 C4.5 Decision Tree 알고리즘에 기초하여 분류 규칙을 생성하고, 상기 생성된 분류 규칙에 대하여 유효성 검사를 수행하는 것을 포함하고,

상기 C4.5 알고리즘은 우선 divide-and-conquer 전략을 사용하여 트리를 성장시킨 다음 트리를 잘라(Cut)내고, 트리를 잘라냄에 따라 모든 속성의 전체 엔트로피와 정보 이득을 계산하고, 계산된 정보 이득이 가장 높은 속성이 결정을 내리기 위해 선택되는 것을 포함하고,

상기 유효성 검사는, n-fold 교차 검증 기법을 사용하고, 데이터 집합을 n(n은 자연수) 개의 부분 집합으로 분할하고, 분할하는 과정을 n(n은 자연수)번 반복함에 따라 n 개의 서브 세트 중 하나가 테스트 세트로 사용되고 다른 n-1 개의 서브 세트가 함께 학습 세트를 형성하여 n 번의 시도에 대한 평균 오차가 계산되는 것을 포함하고,

상기 추천부는,

상기 분류 규칙과 ARM(Association Rule Mining)에 기반하여 상기 클러스터와 관련된 규칙들 중 기 설정된 기준에 의하여 소정의 규칙을 추출하고, 상기 클러스터의 정보, 상기 데이터와 연관된 사용자 식별 정보 및 상기 클러스터에 포함된 데이터 간의 연관성을 식별하고, 상기 식별된 연관성에 기반하여 순위를 정렬하여 제품을 추천하는 것을 포함하고,

상기 추출된 소정의 규칙의 연관성은, Lift($\mathrm{Lift}(R)=\frac{P(XY)}{P(X)P(Y)}$)와 Loevinger($\mathrm{Loevinger}(R)=1-\frac{P(X)P(-Y)}{P(X-Y)}$)의 기준으로 평가되고, 상기 Lift와 Loevinger의 기준은 항목 집합 X, Y 및 규칙 R에 대해 정의되며 X->Y 형식의 설명 패턴이며 X는 연관 규칙의 조건부이고, Y는 결과적인 부분이고, Lift 기준은 X가 발생할 때 Y를 갖는 확률 계수를 나타내고, Loevinger 기준은 결과 부분 Y를 만족하지 않을 확률에 따라 규칙의 중심 신뢰를 표준화하고, Lift와 Loevinge 값이 클수록 더 강한 연관성을 나타내는,

데이터 분석 시스템.

청구항 9

제8항에 있어서,

상기 분류부는,

R 속성, F 속성 및 M 속성을 포함하는 RFM 속성 각각을 기 설정된 기준으로 분할하고, 상기 데이터를 상기 분할된 각각의 속성에 기반하여 RFM 분석을 수행함으로써 상기 데이터에 대한 RFM 값을 획득하는

것을 특징으로 하는 데이터 분석 시스템.

청구항 10

제9항에 있어서,

상기 분류부는,

상기 데이터의 RFM 값에 따라 K-Means ++ 알고리즘을 사용하여 상기 데이터를 적어도 하나 이상의 클러스터로 분류함에 따라 상기 데이터의 클러스터링 결과를 평가하는

것을 특징으로 하는 데이터 분석 시스템.

청구항 11

삭제

청구항 12

삭제

청구항 13

삭제

발명의 설명

기 술 분 야

[0001] 아래의 설명은 RFM(Recency, Frequency, Monetary) 분석 기술에 관한 것이다.

배 경 기 술

[0002] RFM은 최근성, 빈도 및 금전적 가치를 나타낸다. RFM 분석은 최근에 고객이 구매한 시기(최근성), 고객 구매 빈도(빈도) 및 고객이 지출한 금액(화폐)과 같은 고객 행동을 분석하는 데 사용되는 마케팅 기법이다. 향후 개별화 서비스를 위해 고객을 여러 그룹으로 나누고 프로모션에 응답할 가능성이 더 높은 고객을 식별하여 고객 세분화를 향상시키는 것은 유용한 방법이다.

[0003] 최근에는 컴퓨터 보안, 자동차 산업 및 전자 산업과 같은 다양한 분야에서 RFM 개념을 기반으로 하는 데이터 마이닝 응용 프로그램이 제안되었다. RFM 변수를 사용한 데이터 마이닝의 연구 사례로는 신경 네트워크 및 의사 결정 트리, 거친 집합 이론, 자체 구성 지도, CHAID, 유전 알고리즘 및 순차 패턴 마이닝과 같은 다양한 데이터 마이닝 기술이 있다.

[0004] RFM 분석 및 데이터 마이닝 기술의 통합은 현재 및 신규 고객에게 유용한 정보를 제공한다. RFM 특성을 기반으로 하는 클러스터링은 다른 클러스터 분석보다 고객의 실제 마케팅 수준에 대한 더 많은 행동 관련 정보를 제공한다. 고객 인구 통계 변수 및 RFM 변수에서 발견된 분류 규칙은 관리자가 최근에 고객이 구매할 가능성, 고객 구매 빈도 및 구매 가치를 비롯하여 향후 고객 행동을 예측하는 데 유용한 지식을 제공한다. RFM 측정을 기반으로 한 연관 규칙 마이닝은 제품 속성과 고객의 기여/충성도의 관계를 분석하여 고객의 요구를 충족시키는 데 더 나은 권장 사항을 제공한다.

[0005] 이에 따라 고객의 세그먼트, 고객의 현재 RFM 값, 잠재적인 미래 고객 행동 및 함께 자주 구매되는 제품과 같은 여러 매개 변수를 함께 고려하여 단순 권장 사항보다 우수한 제품 권장 사항을 제공하는 기술이 제안될 필요가 있다.

발명의 내용

해결하려는 과제

[0006] 고객의 세그먼트, 고객의 현재 RFM 값, 잠재적인 미래 고객 행동 및 함께 자주 구매되는 제품 등 다양한 매개변수를 함께 고려하여 단순 권장 사항보다 우수한 제품 권장 사항을 제공하는 방법 및 시스템을 제공할 수 있다.

과제의 해결 수단

[0007] 데이터 분석 시스템에 의하여 수행되는 데이터 분석 방법은, 데이터에 대한 RFM 분석을 수행하는 단계; 상기 RFM 분석을 수행함에 따라 획득된 RFM 값에 기초하여 상기 데이터를 적어도 하나 이상의 클러스터로 분류하는 단계; 상기 클러스터의 RFM 값을 통하여 분류 규칙을 생성하는 단계; 및 상기 생성된 분류 규칙으로부터 추천 목록을 제공하는 단계를 포함할 수 있다.

[0008] 상기 RFM 분석을 수행함에 따라 획득된 RFM 값에 기초하여 상기 데이터를 적어도 하나 이상의 클러스터로 분류하는 단계는, R 속성, F 속성 및 M 속성을 포함하는 RFM 속성 각각을 기 설정된 기준으로 분할하고, 상기 데이터를 상기 분할된 각각의 속성에 기반하여 RFM 분석을 수행함으로써 상기 데이터에 대한 RFM 값을 획득하는 단계를 포함할 수 있다.

[0009] 상기 RFM 분석을 수행함에 따라 획득된 RFM 값에 기초하여 상기 데이터를 적어도 하나 이상의 클러스터로 분류하는 단계는, 상기 데이터의 RFM 값에 따라 K-Means ++ 알고리즘을 사용하여 상기 데이터를 적어도 하나 이상의 클러스터로 분류함에 따라 상기 데이터의 클러스터링 결과를 평가하는 단계를 포함할 수 있다.

[0010] 상기 클러스터의 RFM 값에 기반하여 예측된 데이터의 식별 정보를 통하여 분류 규칙을 생성하는 단계는, 상기 클러스터에 포함된 RFM 값과 상기 데이터를 분석하기 위한 별도의 정보를 사용하여 사용자의 행동을 예측하기 위한 분류 규칙을 생성하고, 상기 생성된 분류 규칙에 대하여 유효성 검사를 수행하는 단계를 포함할 수 있다.

[0011] 상기 생성된 분류 규칙으로부터 추천 목록을 생성하는 단계는, 상기 분류 규칙과 ARM(Association Rule Mining)에 기반하여 상기 클러스터와 관련된 규칙들 중 기 설정된 기준에 의하여 소정의 규칙을 추출하는 단계를 포함할 수 있다.

[0012] 상기 생성된 분류 규칙으로부터 추천 목록을 생성하는 단계는, 상기 클러스터의 정보, 상기 데이터와 연관된 사용자 식별 정보 및 상기 클러스터에 포함된 데이터 간의 연관성을 식별하고, 상기 식별된 연관성에 기반하여 순위를 정렬하여 제품을 추천하는 단계를 포함할 수 있다.

[0013] 데이터 분석 시스템에 의하여 수행되는 데이터 분석 방법을 실행시키기 위하여 컴퓨터 판독 가능한 기록매체에 저장된 컴퓨터 프로그램은, 데이터에 대한 RFM 분석을 수행하는 단계; 상기 RFM 분석을 수행함에 따라 획득된 RFM 값에 기초하여 상기 데이터를 적어도 하나 이상의 클러스터로 분류하는 단계; 상기 클러스터의 RFM 값을 통하여 분류 규칙을 생성하는 단계; 및 상기 생성된 분류 규칙으로부터 추천 목록을 제공하는 단계를 포함할 수 있다.

[0014] 데이터 분석 시스템은, 데이터에 대한 RFM 분석을 수행하는 분석부; 상기 RFM 분석을 수행함에 따라 획득된 RFM 값에 기초하여 상기 데이터를 적어도 하나 이상의 클러스터로 분류하는 분류부; 상기 클러스터의 RFM 값을 통하여 분류 규칙을 생성하는 규칙 생성부; 및 상기 생성된 분류 규칙으로부터 추천 목록을 제공하는 추천부를 포함할 수 있다.

[0015] 상기 분류부는, R 속성, F 속성 및 M 속성을 포함하는 RFM 속성 각각을 기 설정된 기준으로 분할하고, 상기 데이터를 상기 분할된 각각의 속성에 기반하여 RFM 분석을 수행함으로써 상기 데이터에 대한 RFM 값을 획득할 수 있다.

[0016] 상기 분류부는, 상기 데이터의 RFM 값에 따라 K-Means ++ 알고리즘을 사용하여 상기 데이터를 적어도 하나 이상의 클러스터로 분류함에 따라 상기 데이터의 클러스터링 결과를 평가할 수 있다.

[0017] 상기 규칙 생성부는, 상기 클러스터에 포함된 RFM 값과 상기 데이터를 분석하기 위한 별도의 정보를 사용하여 사용자의 행동을 예측하기 위한 분류 규칙을 생성하고, 상기 생성된 분류 규칙에 대하여 유효성 검사를 수행할

수 있다.

[0018] 상기 추천부는, 상기 분류 규칙과 ARM(Association Rule Mining)에 기반하여 상기 클러스터와 관련된 규칙들 중 기 설정된 기준에 의하여 소정의 규칙을 추출할 수 있다.

[0019] 상기 추천부는, 상기 클러스터의 정보, 상기 데이터와 연관된 사용자 식별 정보 및 상기 클러스터에 포함된 데이터 간의 연관성을 식별하고, 상기 식별된 연관성에 기반하여 순위를 정렬하여 제품을 추천할 수 있다.

발명의 효과

[0020] 일 실시예에 따른 데이터 분석 시스템은 데이터를 보다 빠르고 정확하게 분석을 수행할 수 있다.

[0021] 일 실시예에 따른 데이터 분석 시스템은 마케팅 분야에서 고객의 미래의 행동을 정확하게 예측하여 제품을 추천함으로써 제품의 구매율을 높일 수 있다.

도면의 간단한 설명

[0022] 도 1은 일 실시예에 따른 데이터 분석 시스템의 동작을 설명하기 위한 도면이다.

도 2는 일 실시예에 따른 데이터 분석 시스템의 구성을 설명하기 위한 블록도이다.

도 3은 일 실시예에 따른 데이터 분석 시스템의 데이터 분석 방법을 설명하기 위한 흐름도이다.

발명을 실시하기 위한 구체적인 내용

[0023] 이하, 실시예를 첨부한 도면을 참조하여 상세히 설명한다.

[0024] 도 1은 일 실시예에 따른 데이터 분석 시스템의 동작을 설명하기 위한 도면이다.

[0025] 데이터 분석 시스템은 기계학습 알고리즘을 적용한 RFM 분석을 수행할 수 있다. 데이터 분석 시스템은 클러스터링, 분류 및 연관 규칙 마이닝을 비롯하여 데이터 마이닝 작업에서 RFM 분석을 사용하여 시장 인텔리전스를 제공하고 마케팅 관리자가 보다 나은 마케팅 전략을 개발할 수 있도록 지원하는 방안을 제안할 수 있다. 데이터 분석 시스템은 클러스터링 작업이 유사한 RFM 값을 갖는 고객 세그먼트를 찾는데 사용되면, 고객 세그먼트와 고객 인구 통계 변수를 사용하여 미래의 고객 행동을 예측하는 분류 규칙을 발견하고, 제품 추천을 위해 연관 규칙 마이닝을 수행할 수 있다. 이와 같이 "미래 고객 행동의 가장 좋은 예측 인자는 과거의 고객 행동"이라는 모델을 제공할 수 있다.

[0026] RFM은 마케팅 분야에서 적용할 때 매우 효과적이다. RFM 분석은 고객의 향후 구매 가능성에 영향을 미치는 3가지 중요한 구매 관련 변수인 Recency(R), Frequency(F) 및 Monetary(M) 측정값에 따라 다르다. Recency는 최신 소비 행위가 발생하는 시간과 현재 시간 사이의 간격을 나타낸다. 많은 마케터들은 가장 최근의 구매자가 최근 구매자보다 다시 구매할 가능성이 높다고 생각한다. Frequency는 고객이 특정 기간 내에 내린 거래 수이다. 이 측정은 구매가 많은 고객이 구매가 적은 고객보다 제품을 구매할 확률이 높다는 가정 하에 사용된다. Monetary란 특정 고객이 소비한 누적 금액을 나타낸다. RFM 분석은 과거의 행동을 기준으로 각 고객에게 가치 점수를 할당할 수 있다. 예를 들면, RFM은 고객의 행동을 정량화할 수 있다.

[0027] 데이터 분석 시스템은 관리자가 데이터 마이닝 및 RFM 분석 결과로 획득된 지식을 최대한 활용하는 마케팅 전략을 개발하는 데 도움을 제공할 수 있다. 든 고객이 동일한 금액을 구매한 것이 아니고, 일부는 더 자주 주문하고 일부는 최근에 주문한 것이 아니기 때문에 인구 통계학적 변수에 따라 고객 행동을 예측하는 데 유용하다. 또한 고객의 세그먼트, 고객의 현재 RFM 값, 잠재적인 미래 고객 행동 및 함께 자주 구매되는 제품과 같은 여러 매개 변수를 함께 고려하여 단순 권장 사항보다 우수한 제품 권장 사항을 제공한다.

[0028] 도 1을 참고하면, IPO(Input, Process and Output) 다이어그램을 나타낸 것이다. 데이터 분석 시스템은 데이터 전처리(120), RFM 분석(130), 고객 세분화(140), 예측(150), 평가(160) 프로세스를 통하여 제품을 권장할 수 있다. 데이터 분석 시스템은 데이터(110)로부터 지식 검색을 보다 쉽고 정확하게 하기 위하여 데이터 전처리 과정을 수행할 수 있다(120). 이때, 데이터에 대한 식별 정보(예를 들면, 데이터에 대한 사용자 정보, 데이터의

식별 정보 등)가 포함될 수 있다. 예를 들면, 데이터에 제품의 거래 정보, 제품의 식별정보(ID), 제품을 구매한 고객 정보, 제품의 주문 날짜, 제품의 수량, 제품의 주문 금액, 제품의 유형 등의 속성 정보가 포함될 수 있다. 데이터 분석 시스템은 속성 수 감소, 이상치 검출, 정규화, 이산화, 개념 계층 생성과 같은 데이터 준비 작업을 개선할 수 있다. 데이터 전처리 과정을 통하여 사실 예측 정확성을 높이고 경과 시간을 절약할 수 있다.

[0029] 데이터 전처리(120) 과정은 채우기, 처리, 변환, 이산화(Discretization), 개념 계층 구조 생성 등의 과정을 수행할 수 있다. 데이터 분석 시스템은 불필요한 특성 (예를 들면, 값이 복수 개 (다른 값은 null 임) 또는 단일 값만 갖는 특성)을 삭제하는 차원성을 감소시킬 수 있다. 구체적으로, 채우기 과정은 누락된 값을 채워져야 하는 과정을 의미한다. 처리 과정은 이상치 및 부정확한 값을 데이터 집합에서 처리되고 제거하는 것을 의미한다. 변환 과정은 데이터가 기 설정된 형식으로 변환되는 것을 의미한다. 이산화 과정은 연관 규칙 마이닝 작업 전에 원래의 값을 적은 수의 값의 범위로 이산화함으로써 연속 속성을 부호화하는 과정을 의미한다. 모든 사건에 대해 거의 다른 가치를 가지고 있기 때문에 높은 카디널리티로 인하여 연관 규칙 마이닝 프로세스에는 거의 의미가 없다. 이 현상의 일례로, 연령 값을 저장하는 속성이다. 연령 속성은 하위 (0-12), 십대 (13-19), 성인 (20-59) 및 상위 (60+)와 같은 네 가지 범위로 그룹화할 수 있다. 개념 계층 구조 생성 방법은 저수준 개념(예를 들면, 이스탄불, 앙카라 또는 이즈미르 도시)을 상위 수준 개념(예를 들면, Marmara, Central Anatolia 또는 Aegean)으로 대체하는 데 사용할 수 있다.

[0030] 데이터 분석 시스템은 RFM 분석(130) 과정을 수행할 수 있다. RFM 분석은 R-F-M 속성의 스케일링을 정의하여 적용될 수 있다. 데이터 분석 시스템은 세 가지의 R 속성, F 속성 및 M 속성 각각의 데이터를 기 설정된 기준, 예를 들면, 내림차순 또는 오름차순으로 정렬할 수 있다. 데이터 분석 시스템은 3 개의 R-F-M 속성을 각각 기 설정된 개수(예를 들면 5 개)의 동일한 크기의 부분으로 분할할 수 있다. 이때, 분할된 각각의 부분은 전체의 20%와 동일하다. 분할된 다섯 개의 부분에는 고객 기부금을 나타내는 5, 4, 3, 2 및 1 점수가 지정된다. '5'는 가장 많은 고객 기여도를 나타내며 '1'은 매출 기여도가 가장 낮은 것을 의미한다. 각각의 R-F-M 속성에 대해 상기 설명한 과정을 개별적으로 반복할 수 있다. R-F-M 속성에서 각각의 속성에는 5 개의 점수(5, 4, 3, 2 및 1)가 있으므로 총 125(5 x 5 x 5)개의 조합이 생성될 수 있다.

[0031] 데이터 분석 시스템은 고객 세분화(140) 과정을 수행할 수 있다. 데이터 분석 시스템은 고객을 유사한 RFM 값을 가진 복수의 클러스터(그룹)으로 분류할 수 있다. 각 고객을 클러스터의 세그먼트에 할당할 수 있다. RFM 분석은 고객 충성도를 평가하고 클러스터링 분석을 통하여 기 설정된 값 이상의 높은 RFM 값을 가진 대상 고객을 식별할 수 있다. 이에 따라 서로 다른 고객 세그먼트에 대하여 서로 다른 마케팅 전략을 채택할 수 있게 된다. 이때, 고객 세그먼트는 클러스터에 포함된 각각의 고객을 의미할 수 있다. 또한, 고객을 여러 고객으로 그룹화하여 추천의 질을 높이고 의사 결정자가 시장 부문을 보다 명확하게 식별하여 효과적인 전략을 개발할 수 있도록 지원할 수 있다. 고객 세분화 과정(140)은 두 개의 하위 단계를 수행할 수 있다. 고객 세분화 과정(140)에서 클러스터링 과정(141) 및 클러스터링 평가 과정(142)이 수행될 수 있다.

[0032] 클러스터링 과정(141)에서 고객에 대한 R-F-M 속성에 따라 데이터는 K-Means ++ 알고리즘을 사용하여 k 개의 클러스터로 분할된다. 실시예에서는 K-Means, SOM과 같은 다른 클러스터링 알고리즘 대신에 K-Means ++ 알고리즘을 적용할 수 있다. K-Means++는 유사한 RFM 값을 가진 고객 세그먼트를 탐색할 수 있다. K-Means++는 클러스터를 일관되게 찾아내고 중지하기 전에 반복되는 반복 횟수를 궁극적으로 결정하는 초기화 절차로 인해 훨씬 빠른 속도를 제공한다. K-Means++는 속성에 따라 n개의 벡터를 k개의 파티션으로 그룹화하는 파티셔닝 클러스터 알고리즘이다(k<n). 이름은 k개의 클러스터가 결정되고 클러스터의 중심이 클러스터 내의 모든 벡터의 평균이라는 사실에서 유래한다. K-Means++는 k개의 적절한 초기 중심을 결정한 다음, 유클리드 거리를 사용하여 가장 가까운 중심에 벡터를 할당하고 할당된 데이터 벡터의 수단으로 새로운 중심을 다시 계산한다. 벡터가 더 이상 반복 사이에 클러스터를 변경하지 않을 때까지 이 과정을 반복한다.

[0033] D는 집합 $A = \{A_1, A_2, ..., A_p\}$의 p 속성으로 표현된 데이터 집합이고, $A_r \in A$는 마지막 트랜잭션 이후의 간격을 포함하고, $A_f \in A$는 특정 기간 내의 트랜잭션 수를 포함하고, $A_m \in A$는 특정 기간 내에 소요된 금액을 포함한다. 각각의 튜플 $t \in D$는 p 개의 튜플 $t = (CustomerID, r_i, f_i, m_i, ...)$을0 가지며, 여기서 $r_i \in Range(A_r)$는 속성

A_r의 범위 내의 값이고, $f_i \in Range(A_f)$는 속성 A_f의 범위 내의 값이고, $m_i \in Range(A_m)$는 속성 A_m의 범위 내의 값이다. $D \leq (1, r_1, f_1, m_1, \ldots), (2, r_2, f_2, m_2, \ldots), \ldots >$로 표현된 데이터 세트 D는 k(k는 자연수)개의 클러스터 $C = (C_1, C_2, \ldots, C_k)$로 분할된다.

[0034] 클러스터링 평가 과정(142)의 목적은 클러스터의 품질을 평가하고 클러스터 중심에서 벗어나지 않는 컴팩트 클러스터를 보장하고 다른 클러스터 간의 분리를 크게 보장하는 것이다. 수학식 1에 정의 된 표준 편차(σ, 수학식 2에 정의 된 SSE(Sum of Squared Error)와 같은 데이터 분할의 효율성을 평가하는 데 여러 가지 방법을 사용할 수 있다.

[0035] 수학식 1:

[0036] $$\sigma = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(x_i - c)^2}$$

[0037] 여기서, X_i (i = 1, 2, .., . N)는 N 개의 객체가 있는 클러스터의 요소이고 c는 클러스터의 중심을 의미한다.

[0038] 수학식 2:

[0039] $$SSE = \sum_{i=1}^{K}\sum_{x \in c_i} dist(c_i, x)^2$$

[0040] K는 클러스터의 수이고, C_i는 i번째 클러스터의 중심을 의미한다.

[0041] 데이터 분석 시스템은 예측 과정(150)을 수행할 수 있다. 예측 과정에서는 향후 고객의 행동을 예측하기 위하여 인구 통계 변수(예를 들면, 연령, 성별, 교육 수준 등) 및 고객 세그먼트의 RFM 값을 사용하여 분류 규칙을 생성할 수 있다. 예를 들면, age = teenager이고 gender = male이고 state = Aegean이면 R ↑ F ↑ M ↓ 이고, 기호 ↑는 값이 평균보다 크다는 것을 나타내고 기호 ↓는 값이 평균보다 작은 것을 나타낸다. 여기서, 고객의 인구 통계학적 가치가 유사하다면 유사한 RFM 값을 가질 가능성이 높다고 판단한다. 치열한 경쟁 환경에서 의사 결정자가 고객 프로필을 보다 명확하게 타겟팅할 수 있도록 고객 인구 통계값을 사용하여 분류 규칙을 탐색할 수 있다. 또한 분류 규칙이 권장 사항에 미치는 영향을 조사하여 보다 효과적인 마케팅 전략을 수립한다. 이러한 예측 과정(150)은 분류 과정(151) 및 분류 평가 과정(152)을 포함할 수 있다.

[0042] 분류 과정(151)에서 고객 인구 통계 변수와 R-F-M 속성을 사용하여 분류 규칙은 C4.5 Decision Tree(Quinlan, 1993) 알고리즘에 의해 적용될 수 있다. C4.5 Decision Tree 알고리즘은 데이터 분석에서 큰 데이터 집합을 분류할 수 있다. C4.5 알고리즘은 우선 divide-and-conquer 전략을 사용하여 초기 트리를 성장시킨 다음 과도한 문제를 피하기 위해 트리를 잘라(Cut)낸다. 트리를 잘라냄에 따라 모든 속성의 전체 엔트로피와 정보 이득을 계산한다. 이때, 정보 이득이 가장 높은 속성이 결정을 내리기 위해 선택될 수 있다. 따라서 트리의 각 노드에서 C4.5는 엔트로피 및 정보 획득에 따라 교육 데이터를 가장 효과적으로 잘라내는 부분 집합으로 가장 효과적으로 분할하는 하나의 특성을 선택한다. D를 세트 $A = \{A1, A2, \ldots, Ap\}$의 p 속성 및 세트 $C = (C1, C2, \ldots, Ck)$의 k 클래스로 표현 된 데이터 세트라고 할 때, 각 샘플 $d \in D$는 p + 1 튜플 $d = <V1, V2, \ldots, Vp; Cj>$ 이고, $Vi \in Range(Ai)$는 속성

$Ai \in A$와 $Cj \in C$의 범위에 있는 값이다. 결정 트리는 애트리뷰트 Ai와 애트리뷰트 값의 서브 세트 Vi를 선택하는 C4.5 알고리즘을 사용하여 구성된다.

[0043] 분류 평과 과정(152)에서 분류에 일반적으로 사용되는 유효성 검사 기술은 예를 들면, 간단한 유효성 검사, 교차 유효성 검사, n뻜old 교차 유효성 검사 및 부트 스트랩 방법 등이 있다. 데이터 분석 시스템은 데이터가 어떻게 분리되는지 덜 중요하기 때문에 n-fold 교차 검증 기법을 사용하는 것을 예를 들어 설명한다. n-fold 교차 검증 기법은 데이터 집합을 n(n은 자연수) 개의 부분 집합으로 분할하고, 이러한 과정을 n(n은 자연수)번 반복할 수 있다. 매번 n 개의 서브 세트 중 하나가 테스트 세트로 사용되고 다른 n-1 개의 서브 세트가 함께 학습 세트를 형성함에 따라 모든 n 번의 시도에 대한 평균 오차가 계산될 수 있다.

[0044] 데이터 분석 시스템은 추천 과정(160)을 수행할 수 있다. 데이터 분석 시스템은 고객에게 세그먼트, RFM 값 및 인구 통계 변수에 따라 더 나은 제품 권장 사항을 제공하기 위해 클러스터링 및 분류 작업 후에 ARM을 적용할 수 있다. 추천 과정(160)은 분류 규칙을 고려하고 FP 성장 알고리즘을 사용하여 각 고객 그룹에서 권장 규칙을 추출한다. 이를 위하여 함께 구입한 고객 세그먼트, 고객 프로필 및 제품 항목 간의 연관을 식별할 수 있다. 데이터 분석 시스템은 추천 과정을 수행함에 따라 관련 순위가 있는 제품을 추천할 수 있으므로 고객 만족도를 높이고 교차 판매를 유도할 수 있다. 추천 과정(160)은 마이닝 과정(161) 및 연결 규칙 평가 과정(162)을 포함할 수 있다.

[0045] 마이닝 과정(161)에서 연관 규칙은 X->Y 형식의 설명 패턴이며 X는 왼쪽편이라고 하며 연관 규칙의 조건부이다. 그 사이에, Y는 우측이라고 칭하고, 결과적인 부분이다. 연관 규칙 마이닝(ARM)은 minsup 임계값 이상을 지원하는 데이터베이스의 항목 간에 숨겨진 흥미로운 연관 규칙을 검색하는 작업이다. 연결 규칙의 지원은 해당 규칙이 데이터에서 얼마나 자주 발생 하는지를 나타낸다. 높은 지원은 데이터베이스의 항목 간의 더 강력한 상관관계에 해당한다. RFM 변수를 사용하여 고객 행동을 분석하는 ARM이 적용될 수 있다. ARM에서 FP 성장(자주 패턴 증가)은 ARM(Association Rule Mining)알고리즘 중 하나이다. FP 성장(자주 패턴 증가)은 ARM 알고리즘 중에서 접두어 트리를 구성하고 규칙을 생성하기 위해 이 트리를 탐색하여 데이터에서 매우 빠른 규칙을 추출할 수 있다. FP 성장(자주 패턴 증가)은 데이터베이스를 두 번만 검색한다. FP-성장은 데이터베이스를 자주 패턴 트리(FP-Tree)로 압축하는 것으로 시작된다. 이 과정에서 트리 탐색의 성능을 향상시키기 위해 모든 빈번한 항목 집합(예를 들면, 1-항목 집합)을 나열하는 헤더 테이블도 생성할 수 있다. 예를 들면, 머리글 표의 각 항목은 두 개의 필드로 구성될 수 있고, 항목 이름과 노드 링크 머리글은 트리에서 첫 번째 항목을 가리킨다. FP-Tree와 헤더 테이블을 생성한 후 헤더 테이블의 항목을 고려하고 조건부 FP-Tree를 재귀적으로 작성하여 FP-tree를 채굴하기 시작한다.

[0046] 데이터 분석 시스템은 마이닝 과정을 수행함에 따라 지원 및 신뢰 임계 값을 사용하고 일반적으로 흥미가 없을 수 있는 많은 수의 연결 규칙을 생성할 수 있다. 이때, 연결 규칙은 일부 평가 조치를 만족하는 경우 유효하다. 연결 규칙 평가 과정(162)은 흥미를 평가하기 위해 측정을 다루기 위한 평가를 수행할 수 있다. 데이터 분석 시스템은 채광된 규칙의 흥미도를 평가하고 Lift와 Loevinger의 두 가지 설명 기준으로 규칙의 관련성을 표현할 수 있다. 이 두 가지 기준은 항목 집합 X, Y 및 규칙 R에 대해 정의되며 X->Y 다음과 같다.

[0047] 수학식 3:

[0048] $$\mathbf{Lift(R)=\frac{P(XY)}{P(X)P(Y)}}$$

[0049] 수학식 4:

[0050] $$\mathbf{Loevinger(R)=1-\frac{P(X)P(-Y)}{P(X-Y)}}$$

[0051] 리프트 기준은 X가 발생할 때 Y를 갖는 확률 계수를 나타낸다. Loevinger 기준은 결과 부분 Y를 만족하지 않을 확률에 따라 규칙의 중심 신뢰를 표준화한다. 일반적으로 Lift와 Loevinge 값이 클수록 더 강한 연관성을 나타낸다.

[0052] 일례로, 데이터 세트가 터키의 스포츠 상점에서 제공되고, 2년 이내에 전자 상거래 웹 사이트를 통해 수집되었다고 하자. 예를 들면, 전체 데이터 세트에는 54개의 하위 그룹에 1584 개의 서로 다른 제품 요구 사항과 2666명의 개별 고객에 대한 6149개의 구매 주문이 포함될 수 있다. 구매 주문에는 거래 ID, 제품 ID, 고객 ID, 주문 날짜, 수량, 주문 금액 (가격), 판매 유형, 할인 및 판촉의 참여 여부 등의 정보가 포함될 수 있다. 고객과 관련된 정보에는 연령, 성별, 결혼 상태, 교육 수준 및 지역과 같은 인구 통계 변수가 포함될 수 있고, 제품과 관련된 정보에는 바코드, 브랜드, 색상, 카테고리, 하위 카테고리, 사용 유형 및 계절과 같은 속성이 포함될 수 있다.

[0053] 데이터 분석 시스템은 데이터에 대한 이상치를 처리하고 누락된 값을 채우고 치수 감소, 변환, 개념 계층 생성, 정규화 및 이산화를 수행할 수 있다. 스포츠 데이터 세트에서 데이터 마이닝에 사용하기에는 부적절한 특성은 제외할 수 있다. 예를 들면, 연령 속성은 하위 (0-12), 십대 (13-19), 성인 (20-59) 및 상위 (60+) 네 가지 범위로 그룹화하고, 아이들의 수는 0, 1, 2 및 3+의 네 그룹으로 대체할 수 있고, 성별 속성은 남성 및 여성 대신 m 및 f로 인코딩할 수 있다.

[0054] 최근성은 각 고객에 대한 최종 거래 날짜와 현재 사이의 시간 간격을 계산하여 구성될 수 있다. 빈도 속성은 각 고객이 특정 기간 내에 만든 거래 수에 기초하여 구성될 수 있다. 통화 속성은 각 고객이 소비한 누적 금액을 계산하여 구성될 수 있다.

[0055] 데이터 분석 시스템은 모든 고객에 대하여 최근성, 빈도 및 금전적 가치를 고려하여 순위를 정렬함에 따라 R-F-M 값으로 정량화할 수 있다. 예를 들면, 데이터의 ID가 5인 고객은 R-F-M 값이 각각 4-3-4로 표기될 수 있다. 또한, RFM 값과 관련하여 고객 수의 분포를 나타낼 수도 있다.

[0056] 데이터 분석 시스템은 동일한 또는 유사한 RFM 값을 가진 고객을 그룹화하기 위해 K-Means++ 클러스터링을 수행할 수 있다. 고객은 최종 거래 (성수기), 구매 빈도 및 총 구매 지출 (금전) 이후의 기간 측면에 기반하여 복수 개의 그룹을 생성할 수 있다. 예를 들면, 각각의 그룹에 따라 최고의 고객(가장 가치있는 고객), 귀중한 고객, 구매자, 최초 고객, 고객 이탈 고객, 빈번한 고객, 지출자 및 불확실한 고객 등으로 분류될 수 있다. 클러스터를 나열함에 따라 고객 유형에 대한 RFM 패턴을 획득할 수 있다.

[0057] 일례로, 특정 RFM 값을 가진 클러스터를 선택함에 따라 선택된 클러스터에 속한 모든 고객이 가장 적합한 마케팅 전략을 수립할 후보가 될 수 있다. 고객 세분화 후에 표준 편차 및 SSE 메트릭을 사용하여 클러스터링 결과를 평가할 수 있다. 데이터 분석 시스템은 인구 통계 변수 (연령, 성별, 교육 수준 등) 및 고객 세그먼트의 RFM 값을 사용하여 분류 규칙을 추출 또는 생성할 수 있다.

[0058] 데이터 분석 시스템은 분류 규칙을 생성한 후 추천 규칙, 즉, 고객 그룹별로 빈번한 구매 패턴을 추출하기 위해 연관 규칙 마이닝을 적용할 수 있다. 추출된 빈번한 구매 패턴은 유사한 RFM 값 및 유사한 인구 통계학적 변수를 가진 고객의 일반적인 구매 행동을 나타낸다. 예를 들면, 45-54 세 여성 모두가 제품을 구매하는 경향이 동일하지 않기 때문에 RFM 값, 고객 세그먼트 및 자주 구입한 제품을 해당 제품과 함께 고려해야 한다. 이에 따라 서로 다른 RFM 값에 따라 서로 다른 제품을 고객에게 추천할 수 있다.

[0059] 도 2는 일 실시예에 따른 데이터 분석 시스템의 구성을 설명하기 위한 블록도이고, 도 3은 일 실시예에 따른 데이터 분석 시스템의 데이터 분석 방법을 설명하기 위한 흐름도이다.

[0060] 데이터 분석 시스템(100)은 분석부(210), 분류부(220), 규칙 생성부(230) 및 추천부(240)를 포함할 수 있다. 이러한 구성요소들은 데이터 분석 시스템(100)에 저장된 프로그램 코드가 제공하는 제어 명령에 따라 프로세서에 의해 수행되는 서로 다른 기능들(different functions)의 표현들일 수 있다. 구성요소들은 도 3의 데이터 분석 방법이 포함하는 단계들(310 내지 340)을 수행하도록 데이터 분석 시스템(100)을 제어할 수 있다. 이때, 구성요소들은 메모리가 포함하는 운영체제의 코드와 적어도 하나의 프로그램의 코드에 따른 명령(instruction)을 실행하도록 구현될 수 있다.

[0061] 프로세서는 데이터 분석 방법을 위한 프로그램의 파일에 저장된 프로그램 코드를 메모리에 로딩할 수 있다. 예를 들면, 데이터 분석 시스템(100)에서 프로그램이 실행되면, 프로세서는 운영체제의 제어에 따라 프로그램의 파일로부터 프로그램 코드를 메모리에 로딩하도록 서버를 제어할 수 있다. 이때, 프로세서 및 프로세서가 포함하는 분석부(210), 분류부(220), 규칙 생성부(230) 및 추천부(340) 각각은 메모리에 로딩된 프로그램 코드 중 대응하는 부분의 명령을 실행하여 이후 단계들(310 내지 340)을 실행하기 위한 프로세서의 서로 다른 기능적 표현들일 수 있다.

[0062] 단계(310)에서 분석부(210)는 데이터에 대한 RFM 분석을 수행할 수 있다.

[0063] 단계(320)에서 분류부(220)는 RFM 분석을 수행함에 따라 획득된 RFM 값에 기초하여 데이터를 적어도 하나 이상의 클러스터로 분류할 수 있다. 분류부(220)는 R 속성, F 속성 및 M 속성을 포함하는 RFM 속성 각각을 기 설정된 기준으로 분할하고, 데이터를 분할된 각각의 속성에 기반하여 RFM 분석을 수행함으로써 데이터에 대한 RFM 값을 획득하는 단계를 포함할 수 있다. 분류부(220)는 데이터의 RFM 값에 따라 K-Means ++ 알고리즘을 사용하여 적어도 하나 이상의 클러스터로 분류함에 따라 데이터의 클러스터링 결과를 평가할 수 있다.

[0064] 단계(330)에서 규칙 생성부(230)는 클러스터의 RFM 값을 통하여 분류 규칙을 생성할 수 있다. 규칙 생성부(230)는 클러스터에 포함된 RFM 값과 데이터를 분석하기 위한 별도의 정보를 사용하여 사용자의 행동을 예측하기 위한 분류 규칙을 생성하고, 생성된 분류 규칙에 대하여 유효성 검사를 수행할 수 있다.

[0065] 단계(340)에서 추천부(340)는 생성된 분류 규칙으로부터 추천 목록을 제공할 수 있다. 추천부(340)는 분류 규칙과 ARM(Association Rule Mining)에 기반하여 클러스터와 관련된 규칙들 중 기 설정된 기준에 의하여 소정의 규칙을 추출할 수 있다. 예를 들면, 추천부(340)는 분류 규칙을 생성한 후, 추천 규칙, 예를 들면, 그룹 고객별로 빈번한 구매 패턴을 추출하기 위하여 연관 규칙 마이닝을 적용할 수 있다. 추천부(340)는 클러스터의 정보, 데이터와 연관된 사용자 식별 정보 및 클러스터에 포함된 데이터 간의 연관성을 식별하고, 식별된 연관성에 기반하여 순위를 정렬하여 제품을 추천할 수 있다.

[0066] 이상에서 설명된 장치는 하드웨어 구성요소, 소프트웨어 구성요소, 및/또는 하드웨어 구성요소 및 소프트웨어 구성요소의 조합으로 구현될 수 있다. 예를 들어, 실시예들에서 설명된 장치 및 구성요소는, 예를 들어, 프로세서, 콘트롤러, ALU(arithmetic logic unit), 디지털 신호 프로세서(digital signal processor), 마이크로컴퓨터, FPGA(field programmable gate array), PLU(programmable logic unit), 마이크로프로세서, 또는 명령(instruction)을 실행하고 응답할 수 있는 다른 어떠한 장치와 같이, 하나 이상의 범용 컴퓨터 또는 특수 목적 컴퓨터를 이용하여 구현될 수 있다. 처리 장치는 운영 체제(OS) 및 상기 운영 체제 상에서 수행되는 하나 이상의 소프트웨어 애플리케이션을 수행할 수 있다. 또한, 처리 장치는 소프트웨어의 실행에 응답하여, 데이터를 접근, 저장, 조작, 처리 및 생성할 수도 있다. 이해의 편의를 위하여, 처리 장치는 하나가 사용되는 것으로 설명된 경우도 있지만, 해당 기술분야에서 통상의 지식을 가진 자는, 처리 장치가 복수 개의 처리 요소(processing element) 및/또는 복수 유형의 처리 요소를 포함할 수 있음을 알 수 있다. 예를 들어, 처리 장치는 복수 개의 프로세서 또는 하나의 프로세서 및 하나의 콘트롤러를 포함할 수 있다. 또한, 병렬 프로세서(parallel processor)와 같은, 다른 처리 구성(processing configuration)도 가능하다.

[0067] 소프트웨어는 컴퓨터 프로그램(computer program), 코드(code), 명령(instruction), 또는 이들 중 하나 이상의 조합을 포함할 수 있으며, 원하는 대로 동작하도록 처리 장치를 구성하거나 독립적으로 또는 결합적으로(collectively) 처리 장치를 명령할 수 있다. 소프트웨어 및/또는 데이터는, 처리 장치에 의하여 해석되거나 처리 장치에 명령 또는 데이터를 제공하기 위하여, 어떤 유형의 기계, 구성요소(component), 물리적 장치, 가상 장치(virtual equipment), 컴퓨터 저장 매체 또는 장치에 구체화(embody)될 수 있다. 소프트웨어는 네트워크로 연결된 컴퓨터 시스템 상에 분산되어서, 분산된 방법으로 저장되거나 실행될 수도 있다. 소프트웨어 및 데이터는 하나 이상의 컴퓨터 판독 가능 기록 매체에 저장될 수 있다.

[0068] 실시예에 따른 방법은 다양한 컴퓨터 수단을 통하여 수행될 수 있는 프로그램 명령 형태로 구현되어 컴퓨터 판독 가능 매체에 기록될 수 있다. 상기 컴퓨터 판독 가능 매체는 프로그램 명령, 데이터 파일, 데이터 구조 등을 단독으로 또는 조합하여 포함할 수 있다. 상기 매체에 기록되는 프로그램 명령은 실시예를 위하여 특별히 설계되고 구성된 것들이거나 컴퓨터 소프트웨어 당업자에게 공지되어 사용 가능한 것일 수도 있다. 컴퓨터 판독 가능 기록 매체의 예에는 하드 디스크, 플로피 디스크 및 자기 테이프와 같은 자기 매체(magnetic media), CD-ROM, DVD와 같은 광기록 매체(optical media), 플롭티컬 디스크(floptical disk)와 같은 자기-광 매체(magneto-optical media), 및 롬(ROM), 램(RAM), 플래시 메모리 등과 같은 프로그램 명령을 저장하고 수행하도록 특별히 구성된 하드웨어 장치가 포함된다. 프로그램 명령의 예에는 컴파일러에 의해 만들어지는 것과 같은 기계어 코드뿐만 아니라 인터프리터 등을 사용해서 컴퓨터에 의해서 실행될 수 있는 고급 언어 코드를 포함한다.

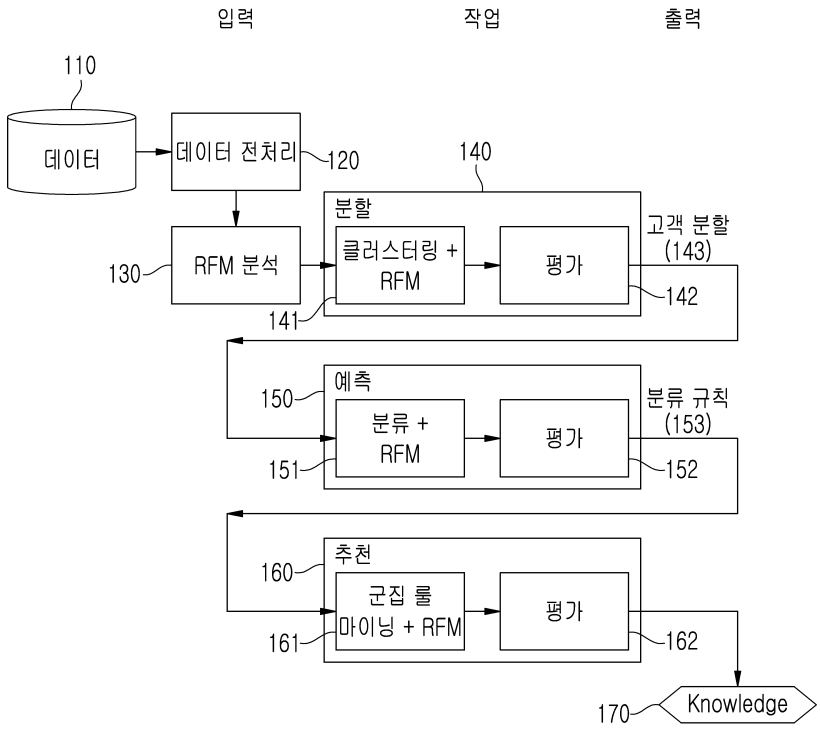
[0069] 이상과 같이 실시예들이 비록 한정된 실시예와 도면에 의해 설명되었으나, 해당 기술분야에서 통상의 지식을 가진 자라면 상기의 기재로부터 다양한 수정 및 변형이 가능하다. 예를 들어, 설명된 기술들이 설명된 방법과 다른 순서로 수행되거나, 및/또는 설명된 시스템, 구조, 장치, 회로 등의 구성요소들이 설명된 방법과 다른 형태로 결합 또는 조합되거나, 다른 구성요소 또는 균등물에 의하여 대치되거나 치환되더라도 적절한 결과가 달성될

수 있다.

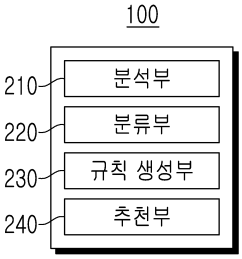
[0070] 그러므로, 다른 구현들, 다른 실시예들 및 특허청구범위와 균등한 것들도 후술하는 특허청구범위의 범위에 속한다.

도면

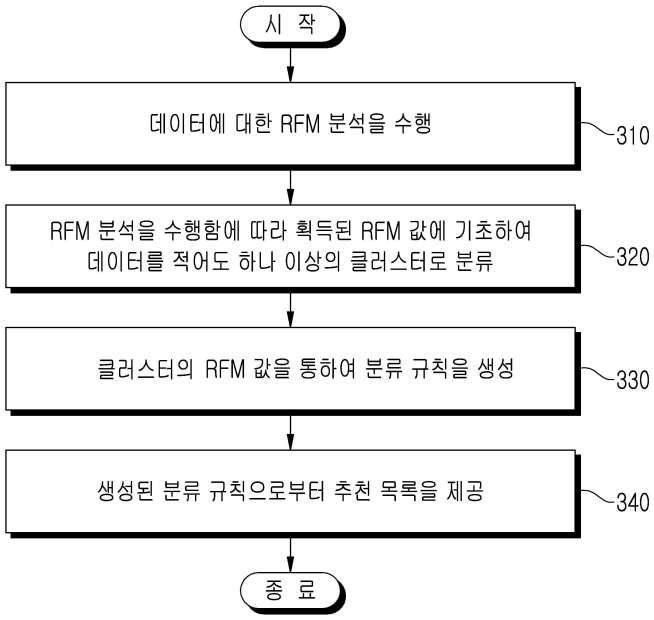
도면1



도면2



도면3


시 작
데이터에 대한 RFM 분석을 수행
310
RFM 분석을 수행함에 따라 획득된 RFM 값에 기초하여 데이터를 적어도 하나 이상의 클러스터로 분류
320
클러스터의 RFM 값을 통하여 분류 규칙을 생성
330
생성된 분류 규칙으로부터 추천 목록을 제공
340
종 료