

명세서

발명의 명칭: 데이터를 분석하는 방법, 장치 및 컴퓨터 프로그램

기술분야

- [1] 본 발명은 데이터를 분석하고 사용자 맞춤형 콘텐츠를 제공하는 방법에 대한 것이다. 보다 구체적으로 본 발명은 신규 사용자 분석에 최적화된 진단용 문제 세트를 추출하고, 머신 러닝 프레임워크가 적용된 데이터 세트를 레이블링하는 방법 및 장치에 대한 것이다.

[2]

배경기술

- [3] 지금까지 교육 콘텐츠는 일반적으로 패키지로 제공되어 왔다. 예를 들어 종이에 기록되는 문제집은 권당 최소 700문제가 수록되어 있으며, 온라인 또는 오프라인 강의 역시 1-2 시간 단위로 최소 한달간 공부할 양을 묶어서 한번에 판매된다.
- [4] 그러나 교육을 받는 학생들 입장에서는 개별적으로 취약한 단원과 취약한 문제 유형이 모두 상이하기 때문에 패키지 형태보다는 개인 맞춤형 콘텐츠에 대한 니즈가 존재한다. 자신이 취약한 단원의 취약한 문제 유형만을 골라서 학습하는 것이 문제집의 7백 문제 전체를 푸는 것보다 훨씬 효율적이기 때문이다.
- [5] 그러나 피교육자인 학생들 스스로 자신의 취약점을 파악하는 것은 매우 어렵다. 나아가 학원, 출판사 등 종래의 교육 업계에서도 주관적 경험과 직관에 의존하여 학생 및 문제들을 분석하기 때문에 개별 학생들에게 최적화된 문제를 제공하는 것을 쉽지 않다.
- [6] 이와 같이 종래의 교육 환경에서는 피교육자가 가장 효율적으로 학습 결과를 낼 수 있는 개인 맞춤형 콘텐츠를 제공하는 것이 쉽지 않으며, 학생들은 패키지 형태의 교육 콘텐츠에 대해 성취감과 흥미를 금방 잃게 되는 문제가 발생한다.

[7]

발명의 상세한 설명

기술적 과제

- [8] 본 발명은 상기와 같은 문제를 해결하는 것을 목적으로 한다. 보다 구체적으로 본 발명은, 사용자 분석을 위해 필요한 샘플 데이터를 효율적으로 추출하는 방법을 제공하는 것을 목적으로 한다. 나아가 본 발명은 비지도 학습 (unsupervised learning) 혹은 자율 학습 기반의 머신 러닝 프레임워크를 적용하여 분석한 데이터를 해석하기 위한 레이블링 방법을 제공하는 것을 목적으로 한다.

[9]

과제 해결 수단

- [10] 본 발명의 실시예를 따르는 데이터 분석 프레임워크의 신규 사용자에 대한 진단용 문제 세트를 구성하는 방법은, 복수의 문제를 포함하는 문제

데이터베이스를 구성하고, 상기 문제에 대한 사용자의 풀이 결과 데이터를 수집하고, 상기 풀이 결과 데이터를 상기 데이터 분석 프레임워크에 적용하여 상기 문제 및/또는 사용자의 모델링 벡터를 계산하는 a 단계; 상기 문제 데이터베이스에서 상기 진단용 문제 세트 구성을 위한 후보 문제를 적어도 하나 이상 추출하는 b 단계; 상기 후보 문제에 대한 풀이 결과 데이터가 존재하는 사용자 및 상기 사용자의 풀이 결과 데이터가 존재하는 다른 문제를 확인하는 c 단계; 상기 후보 문제에 대한 상기 사용자의 풀이 결과 데이터만 상기 데이터 분석 프레임워크에 적용하여 가상 사용자 모델링 벡터를 계산하는 d 단계; 상기 가상 사용자 모델링 벡터를 적용하여 상기 다른 문제의 가상 정답률을 계산하는 e 단계; 상기 가상 정답률과 상기 사용자의 상기 다른 문제의 실제 풀이 결과 데이터를 비교하고, 비교한 결과를 상기 사용자의 수에 따라 평균화하여 상기 후보 문제의 예측률을 계산하는 f 단계를 포함하는 것을 특징으로 한다.

- [11] 나아가 본 발명의 실시예를 따르는 데이터 분석 프레임워크를 통한 분석 결과를 해석하는 방법은, 복수의 문제를 포함하는 문제 데이터베이스를 구성하고, 상기 문제에 대한 사용자의 풀이 결과 데이터를 수집하고, 상기 풀이 결과 데이터를 상기 데이터 분석 프레임워크에 적용하여 상기 문제 및/또는 사용자에게 대한 클러스터를 적어도 하나 이상 형성하는 a 단계; 상기 클러스터에서 적어도 하나 이상의 제 1 데이터를 랜덤하게 추출하고, 상기 제 1 데이터를 해석하기 위한 제 1 레이블을 선정하는 b 단계; 상기 클러스터에 포함된 데이터 중, 상기 제 1 데이터와 임계값 이내의 유사도를 가지는 데이터에 상기 제 1 레이블을 부여하는 c 단계; 상기 제 1 데이터와 임계값 이외의 유사도를 가지는 데이터 중, 적어도 하나 이상의 제 2 데이터를 랜덤하게 추출하고, 상기 제 2 데이터를 해석하기 위한 제 2 레이블을 선정하는 d 단계; 상기 클러스터에 포함된 데이터 중, 상기 제 2 데이터와 임계값 이내의 유사도를 가지는 데이터에 상기 제 2 레이블을 부여하는 e 단계; 상기 제 1 레이블 및 상기 제 2 레이블을 이용하여 상기 클러스터를 해석하는 f 단계를 포함하는 것을 특징으로 한다.

- [12] 본 발명에 따르면, 새로 유입된 사용자 분석을 위해 필요한 최적화된 진단용 문제 세트를 구성할 수 있는 효과가 있다.

- [13] 나아가 본 발명의 실시예를 따르면 기계학습 프레임워크를 적용하여 분석된 결과를 효율적으로 해석할 수 있는 효과가 있다.

- [14]

도면의 간단한 설명

- [15] 도 1은 본 발명의 실시예를 따르는 데이터 분석 프레임워크에서 신규 사용자에게 대한 진단용 문제 세트를 구성하는 방법을 설명하기 위한 순서도

- [16] 도 2는 본 발명의 실시예를 따르는 자율학습 (unsupervised learning)기반의 데이터 분석 프레임워크에서 분석 결과를 해석하는 방법을 설명하기 위한

순서도

[17]

발명의 실시를 위한 형태

[18] 본 발명은 이하에 기재되는 실시예들의 설명 내용에 한정되는 것은 아니며, 본 발명의 기술적 요지를 벗어나지 않는 범위 내에서 다양한 변형이 가해질 수 있음은 자명하다. 그리고 실시예를 설명함에 있어서 본 발명이 속하는 기술 분야에 널리 알려져 있고 본 발명의 기술적 요지와 직접적으로 관련이 없는 기술 내용에 대해서는 설명을 생략한다.

[19] 한편, 첨부된 도면에서 동일한 구성요소는 동일한 부호로 표현된다. 그리고 첨부 도면에 있어서 일부 구성요소는 과장되거나 생략되거나 개략적으로 도시될 수도 있다. 이는 본 발명의 요지와 관련이 없는 불필요한 설명을 생략함으로써 본 발명의 요지를 명확히 설명하기 위함이다.

[20]

[21] 최근 IT 디바이스의 보급이 확대되면서, 사용자 분석을 위한 데이터 수집이 용이해지고 있다. 사용자 데이터를 충분히 수집할 수 있으면, 사용자의 분석이 보다 정밀해지고 해당 사용자에게 가장 적합한 형태의 콘텐츠를 제공할 수 있다.

[22] 이러한 흐름과 함께 특히 교육 업계에서 사용자 맞춤형 교육 콘텐츠 제공에 대한 니즈가 높다.

[23] 간단한 예를 들어, 어떤 사용자가 영어 과목에서 “동사의 시제”에 대한 이해도가 떨어지는 경우, “동사의 시제”에 대한 개념을 포함하고 있는 문제를 추천할 수 있으면 학습 효율은 보다 높아질 것이다. 그런데 이와 같이 사용자 맞춤형 교육 콘텐츠를 제공하기 위해서는 각각의 콘텐츠 및 사용자 개개인에 대한 정밀한 분석이 필요하다.

[24] 종래에는 콘텐츠와 사용자를 분석하기 위해 해당 과목의 개념들을 전문가에 의해 수작업으로 정의하고 해당 과목에 대한 각 문제가 어떤 개념을 포함하고 있는지 전문가가 개별적으로 판단하여 태깅하는 방식을 따랐다. 이후 각 사용자가 특정 개념에 대해 태깅된 문제들을 풀어본 결과 정보를 토대로 학습자의 실력을 분석하는 것이다.

[25] 그러나 이와 같은 방법은 태그 정보가 사람의 주관에 의존하는 문제점이 있었다. 사람의 주관이 개입되지 않고 수학적으로 생성된 태그 정보들이 수학적으로 문제에 부여되는 것이 아니기 때문에 결과 데이터에 대한 신뢰도가 높을 수 없는 문제가 있었다.

[26] 따라서 본 발명의 실시예를 따르는 데이터 분석 서버는 학습 데이터 분석에 머신 러닝 프레임워크를 적용하여 데이터 처리 과정의 사람의 개입을 배제할 수 있다.

[27] 이에 따르면, 사용자의 문제 풀이 결과 로그를 수집하고, 사용자와 문제로 구성된 다차원 공간을 구성하고, 사용자가 문제를 맞았는지 틀렸는지를

기준으로 상기 다차원 공간에 값을 부여하여, 각각의 사용자 및 문제에 대한 벡터를 계산하는 방식으로 사용자 및/또는 문제를 모델링할 수 있다.

- [28] 나아가 상기 사용자 벡터 및/또는 문제 벡터를 이용하여 전체 사용자에서 특정 사용자의 위치, 특정 사용자와 유사한 그룹으로 클러스터링할 수 있는 다른 사용자, 다른 사용자와 해당 사용자의 유사도, 전체 문제에서 특정 문제의 위치, 상기 문제와 유사한 그룹으로 클러스터링할 수 있는 다른 문제, 다른 문제와 해당 문제의 유사도 등을 수학적으로 계산할 수 있다. 나아가 적어도 하나 이상의 속성을 기준으로 사용자 및 문제를 클러스터링할 수 있다.
- [29] 이때 본 발명에서 상기 사용자 벡터, 상기 문제 벡터들이 어떤 속성, 또는 피처를 포함하고 있는지는 제한하여 해석될 수 없음을 유의해야 한다.
- [30] 예를 들어 본 발명의 실시예를 따르면, 상기 사용자 벡터는 상기 사용자가 임의의 개념에 대해 이해하고 있는 정도, 즉 개념의 이해도를 포함할 수 있다. 나아가 상기 문제 벡터는 상기 문제가 어떤 개념들로 구성되어 있는지, 즉 개념 구성도를 포함할 수 있다.
- [31] 그런데 머신러닝을 적용하여 학습 데이터를 분석하면, 몇가지 해결해야 할 문제가 존재한다.
- [32] 첫번째는 새로운 사용자나 문제가 추가되는 경우의 처리에 대한 것이다.
- [33] 신규 유입된 사용자나 문제의 경우, 해당 사용자나 문제에 대한 데이터가 축적되기 전에는 분석 결과를 제공할 수 없다. 따라서 초기 데이터, 즉 데이터 분석 프레임워크에서 초기 분석 결과를 임의의 신뢰도로 도출하기 위해 요구되는 학습 결과 데이터를 효율적으로 수집할 필요가 있다.
- [34] 보다 구체적으로, 새로 유입된 사용자를 분석하기 위해서는 해당 사용자의 문제 풀이 결과 데이터가 어느 정도 축적되어야 하는데, 신뢰성 있는 분석 결과를 제공하기 위한 진단 문제 세트를 구성하는 문제가 해결되어야 한다.
- [35] 문제 풀이 결과 데이터가 어느 정도 축적되지 않은 사용자에게는 신뢰도 있는 분석 결과를 제공할 수 없기 때문에 사용자는 진단용 문제를 풀어야 하고 진단용 문제는 많을수록 보다 정밀한 분석이 가능하다. 그러나 사용자 입장에서는 보다 빨리 학습 효율을 높일 수 있는 맞춤형 문제가 제공되기를 바랄 것이다.
- [36] 따라서 사용자 분석 결과의 신뢰도가 임의의 범위 이상 확보될 수 있는 최소한의 문제로 진단용 문제를 구성할 필요가 있다.
- [37] 본 발명은 상기와 같은 문제를 해결하기 위한 것이다.
- [38] 본 발명의 실시예를 따르면, 새로 유입된 사용자를 분석하기 위한 진단용 문제를 효율적으로 추출할 수 있다. 보다 구체적으로, 데이터 분석 시스템의 문제 데이터베이스의 풀이 결과 데이터가 하나도 존재하지 않은 신규 사용자의 초기 벡터값을 임의의 신뢰도로 계산하기 위해 신규 사용자가 풀어야 할 문제 세트를 효율적으로 추출할 수 있다.
- [39] 이를 따르면, 사용자 진단을 위한 문제 세트가 효율적으로 구성될 수 있어, 사용자가 해당 시스템에서 많은 문제를 풀어보지 않고서도 신뢰성 있는 분석

결과를 제공할 수 있는 효과가 있다.

- [40] 한편 머신러닝을 적용하여 학습 데이터를 분석하는 경우, 머신러닝을 적용하여 분석된 결과값을 사람이 이해할 수 있는 방식으로 해석하기 위한 레이블링의 문제가 발생할 수 있다.
- [41] 사람의 개입 없이, 즉 별도의 레이블링 과정없이 머신러닝 프레임워크를 적용하여 학습 결과 데이터를 모델링하면, 모델링된 결과가 어떠한 피처를 포함하고 있는지 확인할 수 없는 문제가 발생한다. 나아가 사용자 또는 문제를 분류한 경우, 분류 기준이 확인되지 않기 때문에 분석 결과를 사람이 이해할 수 있도록 사후적으로 해석해야 하는 문제가 발생한다.
- [42] 예를 들어 특정 사용자가 제 1 분류, 제 2 분류, 제 3 분류의 속성을 가지는 것으로 분석된 경우, 제 1 분류는 동명사의 이해도가 낮고, 제 2 분류는 시제의 이해도가 높으며, 제 3 분류는 토익 파트 1의 정복률이 중간인 속성을 가지는 것으로 분류 기준을 사람이 이해할 수 있도록 해석할 수 있어야 해당 사용자의 학습 수준 및 취약점을 설명해줄 수 있을 것이다.
- [43] 그러나 일명 비지도 학습 방식의 머신러닝 프레임워크를 적용하여 데이터를 분석하면, 결과 값이 나온 경우에도 데이터의 어떠한 속성에 따라 분류된 것인지 확인하기 어렵다.
- [44] 본 발명은 상기와 같은 문제를 해결하기 위한 것이다.
- [45] 본 발명의 실시예를 따르면, 비지도 학습 기반의 머신 러닝으로 분석된 결과를 사람이 알 수 있는 상태로 해석하기 위해 사후적으로 레이블링하는 방법을 제공할 수 있다.
- [46] 이를 따르면 기계 학습 과정에서 사람의 주관을 배제할 수 있어 순수하게 데이터 기반으로 모델링한 결과를 추출할 수 있으며, 기계 학습과 구분하여 레이블을 지정할 수 있어 기계 학습된 결과를 효율적으로 해석할 수 있는 효과가 있다.
- [47]
- [48] 도 1은 본 발명의 실시예를 따라 사용자 진단용 문제 세트를 추출하는 방법을 설명하기 위한 순서도이다.
- [49] 단계 110 및 단계 115는 데이터 분석 시스템에서 신규 사용자 진단용 문제 세트를 추출하기 위한 전제가 되는 단계이다.
- [50] 본 발명의 실시예를 따르면 단계 110에서 전체 문제와 전체 사용자에 대해 풀이 결과 데이터가 수집될 수 있다.
- [51] 보다 구체적으로, 데이터 분석 서버는 문제 데이터베이스를 구성하고, 상기 문제 데이터베이스에 속하는 전체 문제들에 대한 전체 사용자의 풀이 결과 데이터를 수집할 수 있다.
- [52] 예를 들어 데이터 분석 서버는 시중에 나와 있는 각종 문제들에 대한 데이터베이스를 구축하고, 사용자가 해당 문제들을 푼 결과를 수집하는 방식으로 풀이 결과 데이터를 수집할 수 있다. 상기 문제 데이터베이스는 듣기

- 평가 문제를 포함하고, 텍스트, 이미지, 오디오, 및/또는 동영상 형태일 수 있다.
- [53] 이때 데이터 분석 서버는 수집된 문제 풀이 결과 데이터를 사용자, 문제, 결과에 대한 리스트 형태로 구성할 수 있다. 예를 들어 $Y(u, i)$ 는 사용자 u 가 문제 i 를 풀 결과를 의미하며, 정답인 경우 1, 오답인 경우 0의 값이 부여될 수 있다.
- [54] 나아가 본 발명의 실시예를 따르는 데이터 분석 서버는 사용자와 문제로 구성된 다차원 공간을 구성하고, 사용자가 문제를 맞았는지 틀렸는지를 기준으로 상기 다차원 공간에 값을 부여하여, 각각의 사용자 및 문제에 대한 벡터를 계산할 수 있다. (단계 115) 이때 상기 사용자 벡터와 문제 벡터가 포함하는 피처는 특정되지 않으며, 예를 들어 본 발명의 실시예를 따라 도 3에 대한 설명에서 후술하는 방법에 따라 해석될 수 있다.
- [55] 이후 데이터 분석 서버는 상기 사용자 벡터와 상기 문제 벡터를 이용하여 임의의 사용자가 임의의 문제를 맞출 확률, 즉 정답률을 추정할 수 있다. (단계 120)
- [56] 이때 상기 사용자 벡터와 상기 문제 벡터에 다양한 알고리즘을 적용하여 상기 정답률을 계산할 수 있으며, 본 발명을 해석함에 있어 정답률을 계산하기 위한 알고리즘은 제한되지 않는다.
- [57] 예를 들어 데이터 분석 서버는 상기 사용자의 벡터값 및 상기 문제의 벡터값에 정답률 추정을 위해 파라미터를 설정한 시그모이드 함수를 적용하여 사용자의 해당 문제에 대한 정답률을 계산할 수 있다.
- [58] 또 다른 예로 데이터 분석 서버는 상기 사용자의 벡터값 및 상기 문제의 벡터값을 이용하여 특정 사용자의 특정 문제에 대한 이해도를 추정하고, 상기 이해도를 이용하여 특정 사용자가 특정 문제를 맞출 확률을 추정할 수 있다.
- [59] 예를 들어 사용자 벡터의 1번째 행의 값이 $[0, 0, 1, 0.5, 1]$ 인 경우, 이는 제 1 사용자가 1, 2번째 개념은 전혀 이해하지 못하고, 3번째 및 5번째 개념은 완벽히 이해하고, 그리고 4번째 개념은 절반만큼 이해한 것으로 해석될 수 있다.
- [60] 나아가 문제 벡터의 1번째 행의 값이 $[0, 0.2, 0.5, 0.3, 0]$ 이라 할 때, 이는 제 1 문제가 1번 개념은 전혀 포함하고 있지 않고, 2번 개념이 20% 정도 포함, 3번 개념이 50% 정도 포함, 4번 개념이 30% 정도 포함된 것으로 해석될 수 있다.
- [61] 이때 제 1 사용자의 제 1 문제의 이해도를 추정하면, $0 \times 0 + 0 \times 0.2 + 1 \times 0.5 + 0.5 \times 0.5 + 1 \times 0 = 0.75$ 로 계산될 수 있다. 즉, 제 1 사용자는 제 1 문제를 75퍼센트 이해하는 것으로 추정될 수 있다.
- [62] 그러나 사용자의 특정 문제에 대한 이해도와 특정 문제를 맞출 확률은 동일하다고 할 수 없다. 위의 예에서 제 1 사용자가 제 1 문제를 75 퍼센트 이해한다면 제 1 문제를 실제로 풀었을 때 정답일 확률은 어느 정도인 것인가
- [63] 이를 위해 심리학, 인지과학, 교육학 등에서 사용되는 방법론을 도입하여 이해도와 정답률의 관계를 추정할 수 있다. 예를 들어 Reckase 및 McKinely가 고안한 M2PL (multidimensional two-parameter logistic) 잠재적 특성 이론 (Latent Trait Model) 등을 고려하여 이해도와 정답률을 추정할 수 있다.

- [64] 그러나 본 발명은 합리적인 방식으로 이해도와 정답률 관계를 추정할 수 있는 종래 기술을 적용하여 사용자의 문제에 대한 정답률을 계산할 수 있으면 족하며, 본 발명은 이해도와 정답률의 관계를 추정하는 방법론에 제한되어 해석될 수 없음을 유의해야 한다.
- [65] 이후 데이터 분석 서버는 신규 사용자에 대한 진단용 문제 세트를 구성하기 위해 문제 데이터베이스에서 적어도 하나 이상의 후보 문제를 임의로 추출할 수 있다. (단계 120)
- [66] 이후 데이터 분석 서버는 후보 문제에 대한 풀이 결과 데이터가 존재하는 사용자를 확인하고, 상기 사용자가 상기 후보 문제만 푼 것으로 가정하고 해당 사용자에 대한 가상 벡터값을 계산할 수 있다. 상기 가상 벡터값은 예를 들어 후보 문제의 풀이 결과 데이터만 존재하는 사용자의 문제 데이터베이스의 각각의 문제를 맞출 확률로 계산할 수 있다. (단계 130, 140) 상기 가상 벡터값은 단계 110에 대한 설명에서 전술한 방식은 물론 합리적인 종래 기술을 따라 계산될 수 있다.
- [67] 예를 들어 문제 데이터베이스에서 제 1 문제가 진단 후보 문제로 추출된 경우, 상기 제 1 문제를 풀이한 사용자가 전체 사용자 중, 사용자 1, 사용자 2, 사용자 3이며, 상기 제 1 문제에 대해 사용자 1은 정답, 사용자 2는 정답, 사용자 3은 오답인 경우, 데이터 분석 서버는 (사용자, 문제, val)의 입력값을 (1, 1, 1), (2, 1, 1) (3, 1, 0)로 확인하고, (1, 1, 1), (2, 1, 1) (3, 1, 0)의 입력값만 존재하는 것으로 가정하고 사용자 1, 2, 3이 다른 문제를 맞출 확률을 계산할 수 있다.
- [68] 이는 상기 사용자를 신규 사용자로 가정하고, 신규 사용자가 상기 후보 문제만 풀었을 때, 즉, 신규 사용자에 대한 데이터가 상기 후보 문제에 대한 풀이 결과만 존재할 때, 동일한 분석 프레임워크에서 다른 문제에 대한 정답 예측률이 실제 결과와 얼마나 일치하는지 확인하기 위한 것이다.
- [69] 다시 말하면, 해당 문제를 통해 추정한 다른 문제에 대한 정답 확률이 다른 문제를 실제 풀이한 결과와 정합되는 방향으로 진단 문제를 추출하기 위한 것이다.
- [70] 따라서 데이터 분석 서버는 후보 문제를 풀이한 사용자가 실제로 풀이한 다른 문제를 확인하고, 상기 가상 벡터값을 적용하여 상기 다른 문제의 정답률을 계산하고, 계산된 정답률과 실제 풀이결과를 비교할 수 있다. (단계 160, 170)
- [71] 앞의 예에서 사용자 1이 실제로 제 1 문제, 제 3 문제, 제 5 문제를 풀이했고, 제 1 문제를 맞고 (1, 1, 1), 제 3 문제를 틀리고 (1, 3, 0), 제 5 문제를 맞은 (1, 5, 1) 경우를 가정할 수 있다. 이때, (1, 1, 1)의 입력 값만으로 계산한 가상 사용자 u의 제 3 문제, 제 5문제에 대한 정답률, 즉 가상 벡터값을 적용하여 계산한 제 3 문제, 제 5문제에 대한 정답률이 0.4, 0.6라면, 실제 풀이 결과와 차이는 제 3 문제에 대해 0.6, 제 5 문제에 대해 0.4로 계산될 것이다.
- [72] 이후 단계 180에서 데이터 분석 서버는 후보 문제를 통해 추정한 다른 문제의 정답률과 실제 값의 차이를 평균화할 수 있다. 보다 구체적으로 데이터 분석

서버는 후보 문제에 대한 풀이 결과 데이터가 있는 다른 사용자 전체에 대해, 상기 다른 사용자가 실제 풀이한 문제에 대한 상기 차이를 평균화할 수 있다. 본 명세서에서 이는 진단 문제 후보의 평균 비교값으로 명칭할 수 있다.

- [73] 앞의 예에서 사용자 1이 실제 풀이한 문제가 제 1, 제 3, 제 5 문제이며, 사용자 2가 실제 풀이한 문제가 제 1, 제 2 문제이며, 사용자 3이 실제 풀이한 문제가 제 4, 제 5 문제인 경우, 입력값을 (1, 1, 1)만 존재하는 것으로 가정하여 제 3, 제 5 문제를 맞출 확률과 사용자 1이 실제 제 3, 제 5 문제를 풀이한 결과값의 차이, 입력값을 (2, 1, 1)만 존재하는 것으로 가정하여 제 2 문제를 맞출 확률과 사용자 2가 실제 제 2 문제를 풀이한 결과값의 차이, 입력값을 (3, 1, 0)만 존재하는 것으로 가정하여 제 4, 제 5 문제를 맞출 확률과 사용자 3이 실제 제 4, 제 5 문제를 풀이한 결과값의 차이를 본 발명의 실시예를 따르는 데이터 분석 서버는 계산할 수 있다.
- [74] 이후 데이터 분석 서버는 후보 문제인 제 1 문제에 대해 상기 결과값들의 차이를 문제 2, 3, 4, 5 각각에 대해 평균화할 것이다.
- [75] 데이터분석 서버는 이와 같은 방식으로 문제 데이터베이스에 존재하는 각각의 문제들을 진단 문제 후보들로 설정하여 해당 후보 문제의 평균 비교값을 계산하고, 상기 평균 비교값을 이용하여 진단용 문제를 구성할 수 있다. (단계 190)
- [76] 예를 들어 데이터 분석 서버는 문제 데이터베이스의 모든 문제를 하나씩 진단 문제 후보로 설정하고, 각각의 평균 비교값을 계산하여 평균 비교값이 적은 순서로 진단 문제 후보를 정렬하고, 상위에 정렬된 진단 문제 후보에서 임의의 세트를 추출하는 방식으로 진단 문제 세트를 생성할 수 있다.
- [77] 또 다른 예로 데이터 분석 서버는 문제 데이터베이스에서 미리 설정된 개수로 랜덤하게 추출된 복수의 문제를 진단 문제 후보 세트로 설정하고, 각 세트를 구성하는 각각의 진단 문제 후보의 평균 비교 값을 계산하여 상기 진단 문제 후보 세트의 대표 평균 비교값을 계산하고, 상기 대표 평균 비교값이 미리 설정된 범위 이내인 진단 문제 후보 세트를 최종적으로 진단 문제 세트로 결정할 수 있다.
- [78] 도 2은 본 발명의 실시예를 따라 기계학습 프레임워크를 적용하여 데이터를 분석한 결과를 해석하는 방법을 설명하기 위한 순서도이다.
- [79] 단계 310에서 데이터 분석 서버는 사용자의 문제 풀이 결과 데이터에 기계학습 프레임워크를 적용하여 사용자 및/또는 문제를 모델링할 수 있다.
- [80] 예를 들어 본 발명의 실시예를 따르는 데이터 분석 서버는 소위, 비지도 학습 (Unsupervised Learning) 기반의 기계 학습 프레임워크를 기반으로, 문제 또는 사용자에 대한 별도의 레이블링 없이 문제의 사용자의 풀이 결과만으로 모델링 벡터를 생성할 수 있다.
- [81] 나아가 데이터 분석 서버는 수집된 사용자의 문제 풀이 결과 데이터를 데이터 사이의 거리 기반 또는 확률 분포를 기반으로 유사도를 계산하고, 상기 유사도가

- 임계값 이내인 사용자 및/또는 문제를 분류할 수 있다
- [82] 또 다른 예로 본 발명의 실시예를 따르는 데이터분석 서버는 수집된 사용자의 문제 풀이 결과 데이터를 기반으로 전체 사용자 및 전체 문제에 각각에 대한 벡터를 생성하고, 적어도 하나 이상의 속성을 기준으로 사용자 또는 문제를 분류할 수 있다.
- [83] 그런데 이때 머신 러닝 프레임워크를 적용하여 생성한 사용자 벡터, 문제 벡터는 별도의 레이블이 달려 있지 않으며, 상기 벡터가 어떠한 속성을 포함하고 있는지, 또는 사용자와 문제를 어떠한 속성에 따라 분류한 것인지 해석하기 어려운 문제가 있다.
- [84] 따라서 본 발명의 실시예를 따르는 데이터 분석 프레임워크는 머신러닝을 통한 데이터 분석 결과를 사후적으로 레이블링하여 해석하는 방법을 제안하고자 한다. 본 발명의 실시예를 따르는 레이블링은 머신러닝 과정에서 적용되는 것이 아니라, 머신러닝이 종료되고 난 후, 즉, 머신러닝을 통해 분석된 결과를 해석하기 위해 부여되는 것임을 유의해야 한다.
- [85] 본 발명의 실시예를 따르는 데이터분석 프레임워크는 모델링 벡터로 표현된 문제 또는 사용자 데이터에서, 랜덤하게 적어도 하나의 문제 또는 사용자를 추출하고, 추출된 문제 또는 사용자를 해석하기 위한 적어도 하나의 레이블을 임의로 부여하고, (단계 220) 상기 레이블을 해당 문제 또는 사용자에게 인덱싱할 수 있다. (단계 230)
- [86] 상기 레이블은 예를 들어 특정 과목에 대한 개념 또는 주제를 트리 형식으로 구성한 메타데이터의 인덱싱 정보일 수 있다. 상기 개념 또는 주제는 전문가에 의해 부여될 수 있으나, 본 발명은 이에 한정되지 않는다.
- [87] 도 2에 별도로 도시된 것은 아니지만, 데이터 분석 서버는 레이블 생성을 위해 해당 과목의 학습 요소 및/또는 주제를 트리 구조로 나열하여 최소 학습 요소에 대한 메타데이터 세트를 생성하고, 상기 최소 학습 요소를 분석에 적합한 그룹 단위로 분류할 수 있다.
- [88] 예를 들어 특정 과목 A의 제 1 주제를 A1-A2-A3-A4-A5... 로 분류하고, 제 1 주제 A1의 세부 주제를 제 2 주제로 하여 A11-A12-A13-A14- A15 ...로 분류하고, 제 2 주제 A11의 세부 주제를 제 3 주제로 하여 A111- A112- A113- A114- A115 ...로 분류하고, 제 3주제 A111의 세부 주제를 제 4주제로 하여 동일한 방법으로 분류하는 경우, 해당 과목의 주제는 트리 구조로 나열될 수 있다.
- [89] 이러한 트리 구조의 최소학습 요소들은 사용자 및/또는 문제의 분석에 적합한 단위인 분석 그룹별로 관리될 수 있다. 사용자 및/또는 문제를 해석하기 위한 레이블을 학습 요소의 최소 단위로 설정하는 것보다 분석에 적합한 소정의 그룹 단위로 설정하는 것이 보다 적절하기 때문이다.
- [90] 예를 들어 영어 과목의 학습 요소를 트리 구조로 분류한 최소 단위를 {동사-시제, 동사-시제-과거완료진행, 동사-시제-현재완료진행, 동사-시제-미래완료진행, 동사-시제-과거완료, 동사-시제-현재완료,

동사-시제-미래완료, 동사-시제-과거진행, 동사-시제-현재진행, 동사-시제-미래진행, 동사-시제-과거, 동사-시제-현재, 동사-시제-미래} 으로 구성된 경우, 학습 요소의 최소 단위인 <동사-시제>, <동사-시제-과거완료진행>, <동사-시제-현재완료진행>, <동사-시제-미래완료진행> 각각에 대해 사용자의 취약점을 분석하면, 지나치게 세분화되어 유의미한 분석 결과를 도출하기 어렵다.

- [91] 학습은 특정 분류 아래서 종합적, 전체적으로 진행되기 때문에 과거 완료 진행을 모르는 학생이 현재 완료 진행을 안다고 할 수 없기 때문이다. 따라서 본 발명의 실시예에 따르면 학습 요소의 최소단위는 분석에 적합한 단위인 분석 그룹별로 관리될 수 있으며, 상기 분석 그룹에 대한 정보가 추출된 문제를 설명하기 위한 레이블로 활용될 수 있다.
- [92] 예를 들어 데이터 분석 서버는 클러스터에서 임의로 적어도 하나 이상의 문제를 추출하고, 상기 문제의 출제 의도를 설명할 수 있는 레이블을 추출된 문제에 부여할 수 있다.
- [93] 이후 데이터 분석 서버는 1차 추출된 문제에 부여된 제 1 레이블을 기준으로 전체 문제 데이터를 분류할 수 있다. (단계 230)
- [94] 예를 들어 최초로 추출된 제 1 문제에 제 1 레이블이 지정된 경우, 데이터 분석 서버는 상기 1 문제와 유사도를 기준으로 임계값 이내인 문제들과 임계값 이외인 문제들을 구분할 수 있다.
- [95] 나아가 데이터 분석 서버는 상기 제 1 문제와 유사도가 임계값 이내인 문제들에 상기 제 1 레이블을 부여할 수 있다.
- [96] 이후 데이터 분석 서버는 상기 제 1 문제와 유사도가 임계값 이외인 문제들 중 랜덤하게 적어도 하나의 문제를 추출하고 (단계 240), 2차 추출된 문제를 해석하기 위한 제 2 레이블을 선정하고, 2차 추출된 문제 및 상기 2차 추출된 문제와 유사도가 임계값 이내인 다른 문제들에 상기 제 2 레이블을 부여할 수 있다. (단계 250)
- [97] 이 경우, 1차 추출된 문제와 유사한 문제들에는 제 1 레이블, 2차 추출된 문제와 유사한 문제들에는 제 2 레이블이 부여되고, 1차 추출된 문제는 물론 2차 추출된 문제와 유사한 문제들에는 제 1 레이블과 제 2 레이블이 부여될 것이다.
- [98] 이러한 방식으로 문제들에 레이블 부여를 반복하면 전체 문제를 분류할 수 있다. (단계 260)
- [99] 예를 들어 특정 문제가 <동사-시제>에 대한 제 1 레이블, <동사의 형식>에 대한 제 2 레이블, <능동태와 수동태>에 대한 제 3 레이블이 부여되고, 그 비율이 각각 75%, 5%, 20%인 경우, 해당 문제는 제 1 레이블 및 제 3 레이블을 이용하여 해석될 수 있다.
- [100] 예를 들어 해당 문제는 <동사-시제>를 출제 의도로 하고, <능동태와 수동태>에 대한 오답 보기를 포함하는 것으로 해석될 수 있다.
- [101] 나아가 동일한 제 1 레이블, 제 2 레이블, 제 3 레이블이 사용자에게 부여된

경우, 해당 사용자는<동사-시제> 및 <능동태와 수동태>에 대한 이해도가 각각 75%, 20% 로 추정되는 것으로 해석될 수 있다.

[102]

[103] 본 명세서와 도면에 게시된 본 발명의 실시 예들은 본 발명의 기술 내용을 쉽게 설명하고 본 발명의 이해를 돕기 위해 특정 예를 제시한 것뿐이며, 본 발명의 범위를 한정하고자 하는 것은 아니다. 여기에 게시된 실시 예들 이외에도 본 발명의 기술적 사상에 바탕을 둔 다른 변형 예들이 실시 가능하다는 것은 본 발명이 속하는 기술 분야에서 통상의 지식을 가진 자에게 자명한 것이다.