◆ 13년 10월 고3 A형 28~30번

[28 ~ 30] 다음 글을 읽고 물음에 답하시오.

빅 데이터(Big Data)란 기존의 일반적인 기술로는 관리하기 곤란한 대량의 데이터를 가리키는 것으로, 그 특성은 데이터의 방대한 양과 다양성 및 데이터 발생의 높은 빈도로 요약된다. 이전과 달리 특수 학문 분야가 아닌 일상생활과 밀접한 환경에서도 엄청난 분량의 데이터가 만들어지게 되었고, 소프트웨어 기술의 발달로 이전보다 적은 시간과 비용으로 대량의 데이터 분석이 가능해졌다. 또한 이를 분석하여 유용한 규칙이나 패턴을 발견하고 다양한 예측에 활용하는 사례가 늘어나면서 빅 데이터 처리 기술의 중요성이 부각되고 있다.

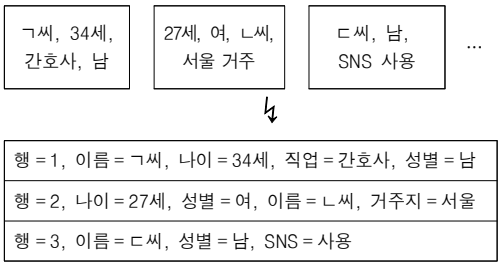
이러한 빅 데이터의 처리 및 분류와 관계된 기술에는 NoSQL 데이터베이스 시스템에 의한 데이터 처리 기술이 있다. 이를 이해하기 위해서는 기존의 관계형 데이터베이스 관리 시스템(RDBMS)에 대한 이해가 필요하다. RDBMS에서는 특정 기준이 제시된 데이터 테이블을 구성하고 이 기준을 속성으로 갖는 정형적 데이터를 다룬다. 고정성이 중요한 시스템이므로 상호 합의된 데이터 테이블의 기준을 자의적으로 추가, 삭제하거나 변용하는 것이 쉽지 않다. 또한 데이터 간의 일관성과 정합성*이 유지될 것을 요구하므로 데이터의 변동 사항은 즉각적으로 반영되어야 한다. <그림 1>은 RDBMS를 기반으로 은행들 간의 상호 연동되는 데이터를 정리하기 위해 사용하는 데이터 테이블의 가상 사례이다.

한에금 씨의 A 은행 거래내역

	거래일자	입금액	출금액	잔액	거래내용	기록사항	거래점
㉠ ····							
㉡ ····	2013.10.08.	30,000		61,217	이체	나저축	B 은행
㉢ ····	2013.10.09.		55,000	6,217	자동납부	전화료	A 은행
㉣ ····							

<그림 1> RDBMS에 의해 구성된 데이터 테이블의 예

NoSQL 데이터베이스 시스템은 특정 기준을 적용하기 어려운 비정형적 데이터를 효율적으로 처리할 수 있도록 설계되었다. 이 시스템에서는 선형으로 데이터의 특성을 나열하여 정리하는 방식을 통해 데이터의 속성을 모두 반영하여 처리한다. <그림 2>는 NoSQL 데이터베이스 시스템으로 자료를 다루는 방식을 나타낸 것이다.



<그림 2> NoSQL 데이터베이스 시스템에 의한 데이터 처리의 예

<그림 2>에서는 '이름=', '나이=', '직업='과 같이 데이터의 속성을 표시하는 기준을 같은 행 안에 포함시킴으로써 데이터의 다양한 속성을 빠짐없이 기록하고, 처리된 데이터를 쉽게 활용할 수 있도록 하고 있다. 또한 이 시스템은 데이터와 관련된 정보의 변용이 상대적으로 자유로우며, 이러한 변화가 즉각적으로 반영되지 않는다는 특성을 지닌다.

* 정합성 : 논리적 모순이 없는 성질이나 상태.

28. 윗글의 설명 방식으로 가장 적절한 것은?

① 시간의 흐름에 따른 빅 데이터 개념의 변화를 설명하고 있다.
② 빅 데이터를 다루는 기술을 기존 기술과 비교하여 설명하고 있다.
③ 빅 데이터가 활용되는 유형을 기준에 따라 구분하여 제시하고 있다.
④ 다양한 사례를 들어 빅 데이터의 특성을 구체적으로 설명하고 있다.
⑤ 빅 데이터의 중요성을 개인적 차원과 사회적 차원으로 나누어 설명하고 있다.

29. ㉠ ~ ㉣에 대한 설명으로 적절하지 <u>않은</u> 것은? [3점]

① ㉠행에 제시된 것은 은행 거래 데이터를 처리하기 위한 기준이다.
② ㉠행의 각 항목은 'A 은행'의 개별 지점에서 임의로 변경하기 어렵다.
③ ㉡행의 거래로 인해 발생한 데이터는 '나저축'의 'B 은행 데이터베이스'에도 즉시 반영된다.
④ ㉡행과 ㉢행의 데이터는 특정 기준을 속성으로 갖는 정형적 데이터이다.
⑤ ㉣행에 기준과 다른 항목을 지닌 데이터가 올 경우 ㉠행의 기준을 즉시 변경하여 데이터를 처리한다.

30. [A]에 'ㄱ씨의 취미는 독서이다.'라는 정보를 추가하고자 한다. 윗글에 비추어 그 방법에 대한 설명으로 가장 적절한 것은?

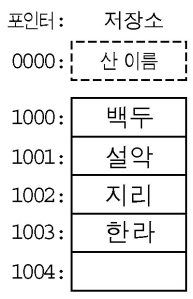
① 1행의 '성별 = 남' 다음에 '취미 = 독서'를 기록한다.
② 1행과 2행 사이에 행을 삽입하여 '취미 = 독서'를 기록한다.
③ 3행 다음에 행을 추가하여 '행 = 4, 이름 = ㄱ씨, 취미 = 독서'를 기록한다.
④ 기준에 맞는 데이터 테이블을 구성하여 해당란에 '독서'를 기록한다.
⑤ 1행에 '독서'라는 속성을 반영할 수 있는 기준이 없으므로 기록 자체가 불가능하다.

- 출전: 시로타 마코토 저. 《빅 데이터의 충격》
- 정답: 28.② 29.⑤ 30.①

[25~26] 다음 글을 읽고 물음에 답하시오.

소프트웨어 개발에서 자료 관리를 위한 구조로는 '배열'과 '연결 리스트'가 흔히 사용된다. 이 구조를 가진 저장소가 실제 컴퓨터 메모리에 구현된 위치를 '포인터'라고 한다.

㉠ 배열은 물리적으로 연속된 저장소들을 사용한다. 배열에서는 흔히 <그림 1>과 같이 자료의 논리적 순서와 실제 저장 순서가 일치하도록 자료가 저장된다. 이때 원하는 자료의 논리적인 순서만 알면 해당 포인터 값을 계산할 수 있으므로, 바로 접근하여 읽기와 쓰기를 할 수 있다. 그런데 <그림 1>에서 자료 '지리'를 삭제하려면 '한라'를 한 칸 당겨야 하고, 가나다순에 따라 '소백'을 삽입하려면 '지리'부터 한 칸씩 밀어야 한다. 따라서 삽입하거나 삭제하는 자료의 순번이 빠를수록 나머지 자료의 재정렬 시간이 늘어난다.



<그림 1> 배열

<그림 2> 연결 리스트

㉡ 연결 리스트는 저장될 자료와 다음에 올 자료의 포인터인 '다음 포인터'를 한 저장소에 함께 저장한다. 이 구조에서는 <그림 2>와 같이 '다음 포인터'의 정보를 담을 공간이 더 필요하지만, 이 정보에 의해 물리적 저장 위치에 상관없이 자료의 논리적 순서를 유지할 수 있다. 또한 자료의 삽입과 삭제는 '다음 포인터'의 내용 변경으로 가능하므로 상대적으로 간단하다. 예를 들어 <그림 2>에서 '소백'을 삽입하려면 빈 저장소의 ⓐ에 '소백'을 쓰고 ⓑ와 ⓒ에 논리적 순서에 따라 다음에 올 포인터 값인 '1004'와 '1002'를 각각 써 주면 된다. 하지만 특정 자료를 읽으려면 접근을 시작하는 포인터부터 그 자료까지 저장소들을 차례로 읽어야 하므로 자료의 논리적 순서에 따라 접근 시간에 차이가 있다.

한편 '다음 포인터'뿐만 아니라 논리순으로 앞에 연결된 저장소의 포인터를 하나 더 저장하는 ㉢'이중 연결 리스트'도 있다. 이 구조에서는 현재 포인터에서부터 앞뒤 어느 방향으로도 연결된 자료에 접근할 수 있어 연결 리스트보다 자료 접근이 용이하다.

25. 위 글을 통해 알 수 있는 사실로 옳지 않은 것은?

① 저장된 자료에 접근할 때는 포인터를 이용한다.
② 자료 접근 과정은 사용하는 자료 관리 구조에 따라 달라진다.
③ '배열'에서는 자료의 논리적 순서에 따라 자료 접근 시간이 달라진다.
④ '연결 리스트'는 저장되는 전체 자료의 개수가 자주 변할 때 편리하다.
⑤ '이중 연결 리스트'의 한 저장소에는 세 가지 다른 정보가 저장된다.

26. ㉠~㉢에 대해 <보기>의 실험을 한 후 얻은 결과로 옳은 것은? [3점]

<보 기>

동일 수의 자료를 논리순이 유지되도록 메모리에 저장한 다음 읽기, 삽입, 삭제를 동일 횟수만큼 차례로 실행하였다.

* 단, 충분히 많은 양의 자료로 충분한 횟수만큼 실험을 하되, 자료를 무작위로 선택하고 자료의 논리순이 유지되도록 함.

① ㉠은 ㉡에 비해 삭제 실험에 걸리는 총시간이 길었다.
② ㉠은 ㉢에 비해 저장 실험의 메모리 사용량이 많았다.
③ ㉡은 ㉠에 비해 삽입 실험에 걸리는 총시간이 길었다.
④ ㉡은 ㉢에 비해 저장 실험의 메모리 사용량이 많았다.
⑤ ㉢은 ㉡에 비해 읽기 실험에 걸리는 총시간이 길었다.

- 정답: 25.③ 26.①

◆ 25 LEET 언어이해 25~27번

[25~27] 다음 글을 읽고 물음에 답하시오.

최근 빅데이터, 소셜 네트워크 서비스 등 대용량 웹서비스를 제공하기 위해 비관계형 데이터베이스가 도입되고 있지만, 정형 데이터를 안정적으로 처리하기 위해서 가장 많이 활용되고 있는 것은 관계형 데이터베이스이다. 관계형 데이터베이스 및 정보 시스템 개발 과정에서 데이터베이스의 체계적 관리를 위한 소프트웨어인 DBMS가 어느 것인지에 상관없이, 데이터를 관리할 수 있도록 표준 질의언어인 SQL이 활용되고 있다.

데이터베이스 트랜잭션은 계좌이체, 주문 처리 등과 같이 한꺼번에 처리해야 하는 논리적 업무 단위를 말한다. 트랜잭션에는 SQL의 조회.삽입.삭제.갱신 등의 작업이 포함된다. 조회작업으로만 구성된 트랜잭션은 데이터베이스 내용을 변화시키지 않는다. 트랜잭션의 개념은 데이터베이스의 안전성을 유지하는 데 필수적이다. 예를 들어 계좌이체의 경우, 도중에 오류가 발생하여 출금 계좌에서 돈이 빠져나갔지만 입금 계좌에는 돈이 안 들어온 상황이 발생해서는 안 된다. 입출금 작업이 모두 성공적으로 종료되어야 이를 완전한 거래로 승인하여 '완료'하고, 일부라도 오류가 발생했을 때는 거래를 아예 진행하지 않은 상태로 '롤백'하여 거래의 안전을 확보해야 하는 것이다.

트랜잭션이 반드시 충족해야 하는 특성으로 원자성.일관성.격리성 등이 있다. 원자성은 계좌이체의 예에서 설명한 바와 같이 트랜잭션의 모든 작업이 성공적으로 완료되거나 아예 아무것도 실행되지 않아야 한다는 특성을 말한다. 일관성은 트랜잭션의 실행 전과 후 모두 데이터베이스에 정의된 무결성 제약조건을 충족하여 논리적으로 일관된 상태를 유지해야 함을 의미한다. 격리성은 둘 이상의 트랜잭션을 동시에 실행할 때 상호 간섭에 의한 문제를 일으키지 않는 성질로, 이를 만족한다면 트랜잭션의 동시 실행의 결과는 트랜잭션을 순차적으로 실행하였을 때의 결과와 같다.

㉠ 트랜잭션의 동시성 제어는 다중 사용자 환경에서 트랜잭션의 일관성과 격리성을 보장하기 위해 DBMS가 제공하는 기능이다. 동시성 제어를 하지 않으면 트랜잭션이 서로 충돌하여 갱신 분실 문제와 모순된 읽기 문제가 발생할 수 있다. 두 트랜잭션이 동일 데이터를 동시에 갱신할 때 한 트랜잭션의 갱신이 다른 트랜잭션이 갱신한 내용을 덮어 쓸 수 있는데, 이를 갱신 분실이라 한다. 모순된 읽기에는 오염된 읽기.반복 불가능한 읽기.팬텀 읽기가 있다. 오염된 읽기는 두 트랜잭션이 동시에 같은 데이터에 접근할 때 한 트랜잭션이 데이터를 갱신한 후 이를 완료하기 전에 다른 트랜잭션이 이 데이터를 읽었으나 이후 데이터 갱신작업을 롤백할 경우 발생하는 문제이다. 반복 불가능한 읽기는 한 트랜잭션 내에서 같은 데이터를 여러 번 조회하는 도중에 다른 트랜잭션이 해당 데이터값을 갱신한 후 완료하면 같은 질의의 결과가 서로 달라지는 문제를 말한다. 팬텀 읽기는 한 트랜잭션에서 질의를 통해 레코드 세트를 읽었지만 다른 트랜잭션이 레코드를 삽입한 후 같은 질의를 반복할 때, 이전과 다른 레코드 세트를 조회하는 현상을 말한다.

한편 SQL에서는 트랜잭션의 동시성 제어를 위한 네 단계의 격리성 수준을 정의한다. 가장 낮은 단계인 미완료 읽기는 완료되지 않은 데이터도 읽을 수 있어 모든 유형의 모순된 읽기가 발생할 수 있다. 다음으로 완료 읽기는 미완료 데이터를 읽지

못하도록 하여 오염된 읽기를 막을 수 있다. 세 번째 단계인 반복 가능 조회는 한 트랜잭션에서 하나의 스냅숏만 사용하도록 하여 오염된 읽기와 반복 불가능한 읽기는 발생하지 않으나, 팬텀 읽기를 막을 수는 없다. 마지막 단계인 직렬화 가능 실행은 2단계 잠금과 같은 기법을 사용하여 트랜잭션의 순차적 실행을 보장함으로써 최고 수준의 격리성을 제공한다. 잠금의 기본 원리는 한 트랜잭션이 자신이 먼저 접근한 데이터를 잠가 다른 트랜잭션의 접근을 막고, 작업을 마치면 이를 풀어 다른 트랜잭션이 사용할 수 있도록 하는 것이다. 이러한 기본 방식의 잠금은 데이터의 독점적 사용으로 인해 동시성을 현저히 저해하며, 또한 트랜잭션의 직렬화 가능 실행을 보장하지 못한다. 이 두 문제를 해결하기 위해 등장한 2단계 잠금은 항상 직렬화 가능 트랜잭션 실행을 보장한다. 일반적으로 격리성 수준이 높을수록 트랜잭션의 독립성이 강해지지만, 성능 및 동시성은 저하된다.

25. 윗글의 내용과 일치하는 것은?

① 조회작업으로 구성된 두 트랜잭션이 동시에 진행되면 모순된 읽기는 발생하지 않는다.

② 트랜잭션의 격리성 수준을 완료 읽기로 설정하면 트랜잭션의 원자성을 충족할 수 있다.

③ SQL 표준을 사용하여 형태가 정해지지 않은 대용량 데이터를 체계적으로 관리할 수 있다.

④ DBMS는 트랜잭션의 원자성을 보장하기 위해 제약조건을 위배하는 트랜잭션을 거부해야 한다.

⑤ 두 트랜잭션이 동일 데이터 영역을 넘나들며 진행되어도 모순된 읽기 문제는 발생하지 않는다.

26. ㉠에 대한 추론으로 적절하지 <u>않은</u> 것은?

① 격리성 수준을 가장 높게 설정하면 갱신 분실 문제가 발생하지 않는다.

② 격리성 수준 중 동시성이 가장 높은 단계는 모순된 읽기를 방지할 수 없다.

③ 격리성 수준을 직렬화 가능 실행에서 미완료 읽기로 변경하면 독립성이 약해진다.

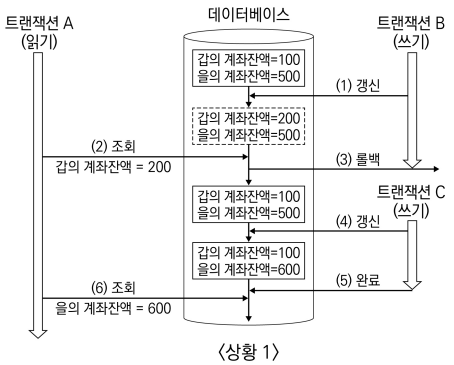
④ 갱신작업으로만 구성된 두 트랜잭션이 동시에 진행할 경우 팬텀 읽기는 발생하지 않는다.

⑤ 데이터를 독점적으로 사용하는 잠금 기법을 적용함으로써 완전한 격리성을 보장할 수 있다.

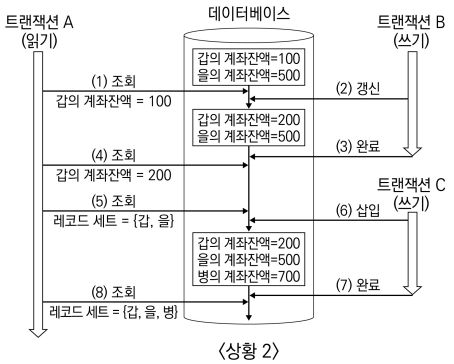
27. 윗글의 내용을 바탕으로 <보기>를 이해할 때, 적절하지 <u>않은</u> 것은?

<보 기>

다음 그림의 각 상황에서 트랜잭션 A는 읽기작업을, 트랜잭션 B와 C는 쓰기작업을 표시된 번호 순서대로 수행한다.



〈상황 1〉



〈상황 2〉

① <상황 1>에서 트랜잭션 A가 조회한 갑의 계좌잔액은 오염된 값이나, 을의 계좌잔액은 오염된 값이 아니다.

② <상황 1>의 모순성을 방지하려면 트랜잭션 A가 미완료 데이터를 조회하는 것을 허용해서는 안 된다.

③ <상황 1>, <상황 2>에서 확인할 수 있는 모순된 읽기의 유형은 모두 3가지이다.

④ <상황 2>는 세 트랜잭션을 순차적으로 실행하여 발생한 모순된 읽기를 보여 준다.

⑤ <상황 2>의 모순성을 방지할 수 있도록 격리성 수준을 설정하면 <상황 1>의 모순성도 발생하지 않는다.

- 정답: 25.① 26.⑤ 27.④

◆ 24 수능 8~11번

[8~11] 다음 글을 읽고 물음에 답하시오.

데이터를 처리할 때 데이터의 정확성은 매우 중요하다. 그런데 데이터에 결측치와 이상치가 포함되면 데이터의 특징을 제대로 ⓐ 나타내기 어렵다.

결측치는 데이터 값이 ⓑ 빠져 있는 것이다. 결측치를 처리하는 방법 중 하나인 대체는 다른 값으로 결측치를 채우는 것인데, 대체하는 값으로는 평균, 중앙값, 최빈값을 많이 사용한다. 중앙값은 데이터를 크기순으로 정렬했을 때 중앙에 위치한 값이다. 크기가 같은 값이 복수일 경우에도 순위를 매겨 중앙값을 찾고, 데이터의 개수가 짝수이면 중앙에 있는 두 값의 평균이 중앙값이다. 또 최빈값은 데이터에 가장 많이 나타나는 값을 이른다. 일반적으로 데이터 값이 연속적인 수치이면 평균으로, 석차처럼 순위가 있는 값에는 중앙값으로, 직업과 같이 문자인 경우에는 최빈값으로 결측치를 대체한다.

이상치는 데이터의 다른 값에 비해 유달리 크거나 작은 값으로, 데이터를 수집할 때 측정 오류 등에 의해 주로 ⓒ 생긴다. 그러나 정상적인 데이터라도 데이터의 특징을 왜곡하는 데이터 값이 있을 수 있다. 예를 들어, 데이터가 어떤 프로 선수들의 연봉이고 그중 한 명의 연봉이 유달리 많다면, 이상치가 포함된 데이터에 해당한다. 이런 데이터의 특징을 하나의 수치로 나타내려는 경우 ㉠ 대푯값으로 평균보다 중앙값을 주로 사용한다.

평면상에 있는 점들의 위치를 나타내는 데이터에서도 이상치를 발견할 수 있다. 대부분의 점들이 가상의 직선 주위에 모여 있다면 이 직선은 데이터의 특징을 잘 나타낸다고 할 수 있다. 이 직선을 직선 L이라고 하자. 그런데 직선 L로부터 멀리 떨어진 위치에도 몇 개의 점이 있다. 이 점들이 이상치이다.

㉡ 이상치를 포함하는 데이터에서 직선 L을 찾는다고 하자. 이때 사용할 수 있는 기법의 하나인 A 기법은 두 점을 무작위로 골라 정상치 집합으로 가정하고, 이 두 점을 ⓓ 지나는 후보 직선을 그어 나머지 점들과 후보 직선 사이의 거리를 구한다. 이 거리가 허용 범위 이내인 점들을 정상치 집합에 추가한다. 정상치 집합의 점의 개수가 미리 정해 둔 기준, 즉 문턱값보다 많으면 후보 직선을 최종 후보군에 넣는다. 반대로 점의 개수가 문턱값보다 적으면 후보 직선을 버린다. 만약 처음에 고른 점이 이상치이면, 대부분의 점들은 해당 후보 직선과의 거리가 너무 ⓔ 멀어 이 직선은 최종 후보군에서 제외되는 것이다. 이 과정을 반복하여 최종 후보군을 구하고, 최종 후보군에 포함된 직선 중에서 정상치 집합의 데이터 개수가 최대인 직선을 직선 L로 선택한다. 이 기법은 이상치가 있어도 직선 L을 찾을 가능성이 높다.

8. 윗글을 이해한 내용으로 적절하지 않은 것은?

① 데이터가 수치로 구성되지 않아도 최빈값을 구할 수 있다.
② 데이터의 특징이 언제나 하나의 수치로 나타나는 것은 아니다.
③ 데이터가 정상적으로 수집되었다면 이상치가 존재하지 않는다.
④ 데이터에 동일한 수치가 여러 개 있어도 중앙값으로 결측치를 대체할 수 있다.
⑤ 데이터를 수집하는 과정에서 측정 오류가 발생한 값이라도 이상치가 아닐 수 있다.

9. 윗글을 참고할 때, ㉠의 이유로 가장 적절한 것은?

① 중앙값은 극단에 있는 이상치의 영향을 덜 받기 때문이다.
② 중앙값을 찾기 위해 데이터를 나열할 때 이상치는 제외되기 때문이다.
③ 데이터의 개수가 많아질수록 이상치도 많아지고 평균을 구하기 어렵기 때문이다.
④ 이상치가 포함되면 평균을 구하는 것이 중앙값을 찾는 것보다 복잡하기 때문이다.
⑤ 이상치가 포함되면 평균은 데이터에 포함되지 않는 값일 가능성이 큰 반면 중앙값은 항상 데이터에 포함된 값이기 때문이다.

10. ㉡과 관련하여 윗글의 A 기법과 <보기>의 B 기법을 설명한 내용으로 가장 적절한 것은? [3점]

<보 기>

다음과 같은 방법으로 직선 L을 찾는 B 기법을 가정해 보자. 후보 직선을 임의로 여러 개 가정한 뒤에 모든 점에서 각 후보 직선들과의 거리를 구하여 점들과 가장 가까운 직선을 선택한다. 그러나 이렇게 찾은 직선은 직선 L로 적합한 직선이 아니다. 이상치를 포함해서 찾다 보니 대부분 최적의 직선과 이상치 사이에 위치한 직선을 선택하게 된다.

① A 기법과 B 기법 모두 최적의 직선을 찾기 위해 최대한 많은 점을 지나는 후보 직선을 가정한다.
② A 기법은 이상치를 제외하고 후보 직선을 가정하지만 B 기법은 이상치를 제외하는 과정이 없다.
③ A 기법에서 최종적으로 선택한 직선은 이상치를 지나지 않지만 B 기법에서 선택한 직선은 이상치를 지난다.
④ A 기법은 이상치의 개수가 문턱값보다 적으면 후보 직선을 버리지만 B 기법은 선택한 직선이 이상치를 포함할 수 있다.
⑤ A 기법에서 후보 직선의 정상치 집합에는 이상치가 포함될 수 있고 B 기법에서 후보 직선은 이상치를 지날 수 있다.

11. 문맥상 ⓐ~ⓔ와 바꿔 쓰기에 가장 적절한 것은?

① ⓐ: 형성(形成)하기
② ⓑ: 누락(漏落)되어
③ ⓒ: 도래(到來)한다
④ ⓓ: 투과(透過)하는
⑤ ⓔ: 소원(疏遠)하여

- 정답: 8.③ 9.① 10.⑤ 11.②

◆ 28 수능 예시문항 14~17번

[14~17] 다음 글을 읽고 물음에 답하시오.

정보 시스템에 대한 '접근'이란 시스템 자원을 사용하기 위해 시스템과 상호 작용하는 작업을 의미한다. 이때 정보의 '객체'는 접근의 대상이 되는 시스템 또는 시스템 자원을, 정보의 '주체'는 접근을 통해 특정 목적을 달성하고자 하는 사람 또는 프로그램 등을 의미한다. '접근제어'는 적절한 권한을 가진 정보 주체만이 정보 객체에 접근할 수 있도록 통제하는 기술이다.

접근제어에서는 보안등급에 따라 접근 권한이 관리되는데, 이때 '보안등급'은 정보 주체와 객체에 부여된 중요도 또는 신뢰도를 나타낸다. 인터넷 카페에서 등급에 따라 읽기 또는 쓰기 권한을 주는 것은 이러한 예에 해당한다. 접근제어에서 관리하는 권한은 접근제어행렬, 접근제어목록 등으로 표현될 수 있다. '접근제어행렬'은 정보 주체를 행으로, 정보 객체를 열로 구성한 테이블로서, 객체에 대한 주체의 접근 권한은 해당 주체의 행과 해당 객체의 열이 만나는 셀에 기록된다. '접근제어목록'은 특정 객체에 대한 접근 권한을 갖는 주체가 나열된 목록이다.

접근제어에는 임의적 접근제어, 강제적 접근제어 등이 있다. ㉠ '임의적 접근제어'에서는 정보 객체의 소유자가 해당 객체에 대한 보안등급을 부여한다. 또한 객체에 대한 주체의 접근 권한 역시 해당 정보 객체의 소유자가 결정한다. 따라서 임의적 접근제어에서 접근 권한을 표현할 때는 접근제어목록이 주로 사용된다. 임의적 접근제어는 구현이 쉽고 권한 관리가 유연한 방식이지만, 정보 객체의 소유자가 접근 권한을 임의로 변경할 수 있어서 접근 권한의 일률적 통제가 어렵다는 문제가 있다. ㉡ '강제적 접근제어'에서는 보안등급 부여와 접근 권한의 관리가 중앙화된 방식으로 수행된다. 따라서 접근 권한을 일률적으로 통제할 수 있다는 장점이 있다. 강제적 접근제어에는 벨라파둘라 모델과 비바 모델 등이 있는데, ㉮ 벨라파둘라 모델은 기밀 정보의 유출 방지에 적합하고, 비바 모델은 정보의 신뢰도 유지에 적합하다.

정보 객체가 문서이고 정보 주체가 객체에 대한 읽기와 쓰기 권한을 갖는다고 가정했을 때, 벨라파둘라 모델에서 정보 주체는 자신보다 높은 등급의 문서를 읽는 것이 금지되지만, 등급이 같거나 낮은 문서에 대해서는 읽는 것이 가능하다. 또한 정보 주체는 자신보다 낮은 등급의 문서에 쓰는 것은 금지되지만, 등급이 같거나 높은 문서에 쓰는 것은 허용된다. 비바 모델에서 정보 주체는 자신보다 높은 등급의 문서에 대해서는 쓰기 권한이 없지만, 등급이 같거나 낮은 문서에 대해서는 쓰기가 가능하다. 또한 정보 주체는 자신보다 낮은 등급의 문서에 대해서는 읽기 권한이 없지만, 등급이 같거나 높은 문서를 읽는 것이 허용된다. 정보 주체는 자신보다 낮은 등급의 문서에 포함된 신뢰도가 낮은 정보를 참조함으로써 자신이 보유한 정보의 신뢰도를 떨어뜨릴 수 있는데, 비바 모델에서는 이를 방지할 수 있다.

14. 윗글의 내용과 일치하지 않는 것은?

① 접근제어행렬은 접근 권한을 나타내는 테이블이다.
② 임의적 접근제어의 접근 권한 표현에는 접근제어목록이 주로 사용된다.
③ 접근은 시스템과의 상호 작용을 통해 시스템 자원을 사용하는 것을 목적으로 한다.
④ 접근제어에서는 정보 주체와 정보 객체에 부여된 중요도나 신뢰도에 따라 접근 권한이 관리된다.
⑤ 접근제어목록은 특정 정보 주체가 접근할 수 있는 정보 객체를 목록화하여 관리하기 위해 사용된다.

15. ㉠과 ㉡에 대한 이해로 적절하지 <u>않은</u> 것은?

① ㉠과 달리 ㉡은 중앙화된 방식으로 접근 권한을 통제하기 때문에 일률적인 권한 관리가 가능하다는 특징이 있다.
② ㉠과 달리 ㉡은 정보 객체의 소유자 외의 정보 주체가 해당 객체를 변경하는 것을 방지하기 위해 사용되는 방식이다.
③ ㉡과 달리 ㉠은 정보 객체의 소유자가 접근 권한을 관리하기 때문에 권한 관리가 유연한 방식이다.
④ ㉠과 ㉡은 모두 권한을 부여하고 관리하기 위해 사용된다.
⑤ ㉠과 ㉡은 모두 접근제어행렬을 이용한 접근 권한 표현이 가능한 방식이다.

16. ㉮의 이유로 가장 적절한 것은?

① 정보 객체의 정보가, 같은 등급의 정보 주체로 전달되지 않기 때문이다.
② 정보 주체와 정보 객체의 보안등급이 중앙화된 방식으로 관리되기 때문이다.
③ 정보 주체가 자신보다 낮은 등급의 정보 객체에 쓰는 것이 금지되기 때문이다.
④ 정보 주체가 자신보다 높은 등급의 정보 객체에 쓰는 것이 가능하기 때문이다.
⑤ 정보 주체와 정보 객체를 중요도에 따라 분류하고 이를 테이블을 이용해서 관리하기 때문이다.

17. 윗글을 바탕으로 〈보기〉를 이해한 내용으로 적절하지 <u>않은</u> 것은? [3점]

〈보 기〉

다음은 비바 모델 접근제어를 사용하는 ○○ 회사의 접근제어행렬이다. 이 회사에는 갑, 을, 병이라는 정보 주체와 A, B, C라는 정보 객체가 있다. 이 회사는 모든 정보 주체 및 객체를 1등급, 2등급, 3등급의 보안등급으로 분류하고 있다. 테이블에서 r은 읽기 권한을, w는 쓰기 권한을 의미한다.

객체 / 주체	A	B	C
갑	[ ]	rw	r
을	rw	w	r
병	w	w	rw

① 모든 주체가 B에 대한 쓰기 권한을, C에 대한 읽기 권한을 가지고 있음을 고려할 때, 갑은 A에 대한 읽기 권한을 가지고 있겠군.
② 을은 병에 비해 읽기 권한이 많다는 점을 고려할 때, 보안등급은 을이 병보다 높겠군.
③ 을은 A에 대한 읽기 권한과 쓰기 권한을 모두 가지고 있음을 고려할 때, 을과 A의 보안등급은 같겠군.
④ 을은 C에 대한 읽기 권한이 있으므로 C보다 보안등급이 낮은 을에게 C의 중요 정보가 유출될 수 있겠군.
⑤ 병이 A와 B에 대한 읽기 권한이 없는 것은 병이 보유한 정보의 신뢰도 하락을 막기 위한 것이겠군.

- 정답: 14.⑤ 15.② 16.③ 17.②

◆ 19년 10월 고3 30~35번

[30 ~ 35] 다음 글을 읽고 물음에 답하시오.

현대 사회는 정보 통신 기술의 발달로 매일 엄청난 양의 자료가 생성·축적되고 있다. 이러한 많은 양의 자료에서 유용한 정보를 찾아 활용하기 위해 다양한 분석 기법이 쓰이는데, 그중 정책 수립, 기업 관리, 의학 분야 연구, 마케팅 등에 널리 쓰이는 것이 연관성 분석이다. 마케팅 분야를 예로 든다면, 연관성 분석은 수집한 자료 안에 존재하는 품목 간의 연관 규칙을 발견하는 과정을 말하며, 연관 규칙은 '고객이 X를 사면 Y도 산다.'의 형태를 띤다. 이때 '고객이 X를 산다.'는 조건이 되고 '고객이 Y를 산다.'는 결과가 된다. 연관 규칙은 'X→Y'와 같이 조건과 결과를 기호로 표현하는 것이 일반적이며, 통계학의 확률을 기반으로 한다.

연관성 분석을 통해 유용한 연관 규칙을 찾기 위해서는 대상 품목들이 어느 정도의 연관성이 있는지를 측정해야 한다. 연관성 측도의 기본은 발생 빈도로, 이와 관련한 주요 측도에는 지지도, 신뢰도, 향상도가 있다. 먼저 지지도는 전체 거래에 대해서 조건과 결과에 있는 품목들이 함께 구매되는 경향을 나타낸다. 'X→Y'의 지지도는 X와 Y를 모두 구매하는 거래의 수를 전체 거래의 수로 나눈 값으로, 지지도가 높다는 것은 동시 구매가 많이 일어난다는 것을 의미한다. <표>는 다섯 가지의 품목만 취급하는 편의점에서 다섯 명의 고객이 한 번씩만 거래했다고 가정한 것이다. <표>에서 생수와 빵을 모두 산 경우는 다섯 번의 거래 중 두 번이므로, '생수→빵'의 지지도는 2/5(40%)이다. '빵→생수'의 지지도도 2/5이므로 'X→Y'와 'Y→X'의 지지도는 같다.

고객	품목
1	빵, 생수, 우유
2	빵, 휴지, 우유
3	빵, 세제, 우유
4	빵, 생수, 세제
5	생수, 휴지, 우유

<표>

신뢰도는 조건의 구매가 발생하였을 때 결과의 구매가 일어날 확률이다. 즉 'X→Y'의 신뢰도는 X와 Y를 모두 구매하는 거래의 수를 X를 구매하는 거래의 수로 나눈 값이다. 따라서 신뢰도가 높다는 것은 조건의 구매가 발생한 경우에 결과의 구매가 많이 일어남을 의미한다. <표>에서 생수를 구매한 세 번의 거래 중에서 두 번만 빵을 샀으므로, '생수→빵'은 2/3(약 66.7%)의 신뢰도를 갖는다. 그런데 '빵→생수'의 신뢰도는 2/4(50%)이다. 이처럼 'X→Y'와 'Y→X'의 신뢰도는 같지 않을 수 있다.

[A] 향상도는 어떤 연관 규칙에 대하여 조건 없이 결과가 일어날 확률보다, 조건이 일어났을 때 결과가 일어날 확률이 얼마나 더 향상되는지를 알려 주는 측도이다. 향상도는 신뢰도를 기대 신뢰도로 나눈 값이다. 기대 신뢰도란 'X→Y'에서 Y를 포함하는 거래의 수를 전체 거래의 수로 나눈 값이다. 'X→Y'에서 향상도가 1이라는 것은 X와 Y의 구매가 서로 독립적이라는 의미이다. 그리고 'X→Y'에서 향상도가 1보다 크다는 것은 X를 구매했을 때 Y를 구매할 확률이, 전체 거래에서 Y를 구매할 확률보다 크다는 것이다. 따라서 이 연관 규칙은 결과를 예측하는 데 있어서 우연적 기회보다 우수하여 마케팅 전략을 ⓐ 세우는 데 유용하게 활용된다. 반면에 'X→Y'에서 향상도가 1보다 작다는 것은 X를 구매했을 때 Y를 구매할 확률이, 전체 거래에서 Y를 구매할 확률보다 작다는 것이므로 이 연관 규칙을 마케팅 전략에 바로 적용하기는 어렵다. 그래서 향상도가 1보다 작은 경우에는 음의 연관 규칙을 만들어 유용하게 쓰일 수 있도록 하기도 한다. 음의 연관 규칙은 결과에 '이다' 대신에 '아니다'를 쓴다는 것을 제외하고는 연관 규칙과 유사하다. 예컨대 'X→Y'의 신뢰도가 30%이고, 'X→Y'의 기대 신뢰도가 40%라고 가정해 보자. 이 경우 'X→Y'의 향상도는 3/4으로 1보다 작다. 따라서 이를 음의 연관 규칙, 곧 'X를 사면 Y를 사지 않는다.'로 전환하면, 신뢰도는 70%(100% − 30%)가 되고, 기대 신뢰도는 60%(100% − 40%)가 되므로 향상도는 7/6로 1보다 커지게 되어 유용하게 쓰일 수 있다.

이와 같은 연관성 분석은 결과가 명확하기 때문에 이해하기 쉽고, 유용한 연관 규칙의 형태로 주어지므로 마케팅 전략에 적용하기도 좋다. 그러나 분석하려는 품목의 수가 늘어나면 연관 규칙이 기하급수적으로 늘어난다는 문제가 발생하는데, 이 문제를 해결하기 위한 보편적 방법으로 거래가 충분히 이루어지지 않은 품목을 제거하는 최소지지도 가지치기가 있다. 이는 지지도가 낮은 품목을 분석 대상에서 삭제하거나, 하위 품목을 상위 품목으로 일반화하여 품목들이 분석자가 임의로 설정한 최소지지도를 넘게 하는 것이다.

지금까지 살펴본 연관성 분석은 사건들의 발생 순서는 분석의 고려 대상으로 삼지 않았다. 그런데 순차적으로 일어나는 사건들을 나열한 시계열 자료를 분석하여 선후 사건들 사이의 연관성을 추론할 수도 있다. 이를 ㉠ 시차 연관성 분석이라고 한다. 시간의 흐름에 따라 어떤 사건들이 일어났는지를 분석하여 사건들 간의 연관성을 발견하면, 이러한 연관성을 토대로 미래의 사건을 예측하거나 사건들 사이의 인과 관계를 추론하는 등 다양하게 활용할 수 있다. 이와 같은 시차 연관성 분석을 하기 위해서는 사건이 일어난 시간이나 순서를 알려 주는 정보가 필요하다. 또한 다른 시간대에 일어난 사건이 동일한 분석 대상에서 일어났다는 것을 알려 주는 분석 대상의 식별 정보도 필요하다.

30. 윗글에 대한 설명으로 적절하지 않은 것은?

① 연관성 분석에 쓰이는 측도들을 예를 들어 설명하고 있다.
② 시차 연관성 분석의 특징과 분석에 필요한 요소들을 밝히고 있다.
③ 연관성 분석이 시대에 따라 변천하게 된 과정을 설명하고 있다.
④ 연관성 분석에서 발생할 수 있는 문제를 해결하기 위한 방법을 제시하고 있다.
⑤ 다양한 분석 기법이 여러 분야에서 널리 쓰이게 된 사회적 배경을 소개하고 있다.

31. 윗글의 내용과 일치하지 않는 것은?

① 연관성 분석에서 분석하려는 품목을 상위 품목으로 일반화하면 연관 규칙의 수가 기하급수적으로 늘어난다.
② 최소지지도 가지치기에는 지지도가 낮은 품목을 분석 대상에서 삭제하는 방법이 있다.
③ 연관성 분석은 결과가 명확하고 유용한 연관 규칙의 형태로 주어지는 장점이 있다.
④ 향상도가 1이라는 것은 조건과 결과가 서로 독립적이라는 의미이다.
⑤ 연관성 측도에서 기본이 되는 것은 발생 빈도이다.

32. 윗글의 <표>에 대해 이해한 내용으로 적절하지 않은 것은?

① '빵→생수'가 '빵→휴지'의 지지도보다 높은 것은 '빵'을 '생수'와 함께 구매한 경우가 '빵'을 '휴지'와 함께 구매한 경우보다 많은 것을 의미한다.
② '휴지→우유'의 신뢰도가 100%인 것은 '우유'를 구매한 모든 경우에 '휴지'를 구매한 것을 의미한다.
③ '생수→빵'과 '생수→우유'는 '생수→휴지'보다 신뢰도가 높다.
④ '우유→생수'의 지지도와 '생수→우유'의 지지도는 같다.
⑤ '빵→세제'의 신뢰도와 '세제→빵'의 신뢰도는 다르다.

33. ㉠을 활용한 사례로 적절한 것만을 <보기>에서 있는 대로 고른 것은?

< 보 기 >

ㄱ. 어느 병원에서 □□ 질환을 앓은 환자들을 추적하여, 이들 가운데 이전에 ○○ 질환을 앓은 경우가 많다는 것을 밝혀냈다. 이후 ○○ 질환을 앓는 환자의 경우에는 □□ 질환에 대한 예방 치료도 하도록 하였다.
ㄴ. 대형 유통 업체에서 10월 한 달간 라면과 계란의 판매대를 붙여 놓았을 때와 멀리 떼어 놓았을 때의 판매량을 조사하여, 멀리 떼어 놓았을 때의 판매량이 높다는 결과를 얻었다. 그 결과를 토대로 두 상품의 판매대를 멀리 떼어 놓기로 결정했다.
ㄷ. 백화점에서 자사의 백화점 카드로 결제한 고객들의 소비 성향을 분석하여, TV를 산 고객들이 재방문하여 고성능 스피커를 구입하는 경향이 있음을 알아내었다. 이를 토대로 TV를 산 고객들에게 고성능 스피커에 대한 상품 안내서를 우편으로 보냈다.
ㄹ. 온라인 쇼핑몰 운영자가 회원들의 웹 페이지 방문 순서를 분석하여, 주로 'A 웹 페이지→B 웹 페이지→C 웹 페이지→ .....' 순으로 방문한다는 규칙을 발견하였다. 그래서 회원들이 편리하게 이 경로에 따라 방문할 수 있는 회원 전용 웹 페이지를 따로 만들었다.

① ㄱ, ㄴ, ㄹ ② ㄱ, ㄷ, ㄹ ③ ㄱ, ㄷ
④ ㄴ, ㄷ ⑤ ㄴ, ㄹ

34. [A]를 바탕으로 할 때, <보기>에 대해 보인 반응으로 가장 적절한 것은? [3점]

< 보 기 >

어느 매장에서 고객들이 팥빙수를 만들기 위해 구매한 팥(A), 인절미(B), 콩가루(C)의 전체 거래 정보에 대해 연관성 분석을 하였다. 다음은 이를 통해 발견한 연관 규칙의 일부이다.

연관 규칙 (X→Y)	기대 신뢰도	신뢰도	향상도	
A→B	42.5%	55.6%	1.308	..... ㉮
B→C	40.0%	35.3%	0.883	..... ㉯
C→A	45.0%	50.0%	1.111	..... ㉰
:	:	:	:	

① ㉮의 연관 규칙에서 B를 포함하는 거래의 수를 전체 거래의 수로 나눈 값은 ㉰의 연관 규칙에서 A를 포함하는 거래의 수를 전체 거래의 수로 나눈 값보다 크군.
② ㉯의 연관 규칙에서 B를 구매했을 때 C를 구매할 확률은 전체 거래에서 C를 구매할 확률보다 작군.
③ ㉯의 연관 규칙의 신뢰도는 ㉯의 음의 연관 규칙의 신뢰도보다 크군.
④ ㉯의 연관 규칙이 ㉮의 연관 규칙보다 마케팅 전략에 바로 적용하여 활용하기에 유용하겠군.
⑤ ㉰의 연관 규칙을 음의 연관 규칙인 'A→C'로 전환하면 더욱 유용하게 쓸 수 있겠군.

35. ⓐ의 문맥적 의미와 가장 유사한 것은?

① 변호사는 그를 증인으로 세웠다.
② 시험이 끝난 학생들이 방학 계획을 세웠다.
③ 과장은 회사의 실적을 올리는 데 공을 세웠다.
④ 목수는 목재를 잘 자르기 위해 톱날을 세웠다.
⑤ 우리 학교는 많은 노력을 기울여 전통을 세웠다.

- 출전: 김병수 외 공저. 《데이터마이닝 입문》
- 정답: 30.③ 31.① 32.② 33.② 34.② 35.②

[14~17] 다음 글을 읽고 물음에 답하시오.

인터넷 검색 엔진은 검색어를 포함하는 웹 페이지를 찾아 화면에 보여 준다. 웹 페이지가 화면에 나타나는 순서를 정하기 위해 검색 엔진은 수백 개가 ⓐ넘는 항목을 고려한 다양한 방식을 사용한다. 대표적인 항목으로 중요도와 적합도가 있다.

검색 엔진은 빠른 시간 내에 검색 결과를 보여 주기 위해 웹 페이지들의 데이터를 수집하여 인덱스를 미리 작성해 놓는다. 인덱스란 단어를 알파벳순으로 정리한 목록으로, 여기에는 각 단어가 등장하는 웹 페이지와 단어의 빈도수 등이 저장된다. 이때 각 웹 페이지의 중요도가 함께 기록된다.

㉠중요도는 웹 페이지의 중요성을 값으로 나타낸 것으로 링크 분석 기법으로 측정할 수 있다. 기본적인 링크 분석 기법에서 웹 페이지 A의 값은 A를 링크한 각 웹 페이지들로부터 받는 값의 합이다. 이렇게 받은 A의 값은 A가 링크한 다른 웹 페이지들에 균등하게 나눠진다. 즉 A의 값이 4이고 A가 두 개의 링크를 통해 다른 웹 페이지로 연결된다면, A의 값은 유지되면서 두 웹 페이지에는 각각 2가 보내진다.

하지만 두 웹 페이지가 실제로 받는 값은 2에 댐핑 인자를 곱한 값이다. 댐핑 인자는 사용자들이 웹 페이지를 읽다가 링크를 통해 다른 웹 페이지로 이동하지 않는 비율을 반영한 값으로 1 미만의 값을 가진다. 댐핑 인자는 모든 링크에 동일하게 적용된다. 가령 그 비율이 20%이면 댐핑 인자는 0.8이고 두 웹 페이지는 A로부터 각각 1.6을 받는다. 웹 페이지로 연결된 링크를 통해 받는 값을 모두 반영했을 때의 값이 각 웹 페이지의 중요도이다. 웹 페이지들을 연결하는 링크들은 변할 수 있기 때문에 검색 엔진은 주기적으로 웹 페이지의 중요도를 갱신한다.

사용자가 검색어를 입력하면 검색 엔진은 인덱스에서 검색어에 적합한 웹 페이지를 찾는다. ㉡적합도는 단어의 빈도, 단어가 포함된 웹 페이지의 수, 웹 페이지의 글자 수를 반영한 식을 통해 값이 정해진다. 해당 검색어가 많이 나올수록, 그 검색어를 포함하는 다른 웹 페이지의 수가 적을수록, 현재 웹 페이지의 글자 수가 전체 웹 페이지의 평균 글자 수에 비해 적을수록 적합도가 높아진다. 검색 엔진은 중요도와 적합도, 기타 항목들을 적절한 비율로 합산하여 화면에 나열되는 웹 페이지의 순서를 결정한다.

14. 윗글을 통해 알 수 있는 내용으로 가장 적절한 것은?

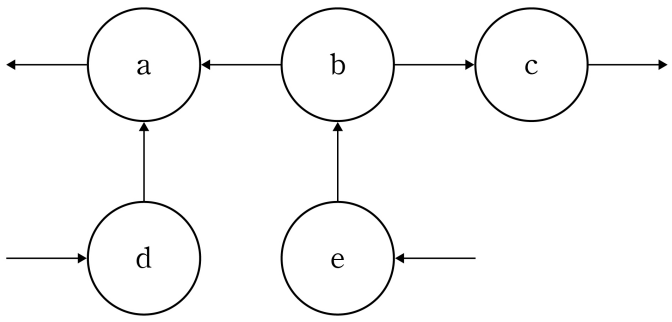
① 인덱스는 사용자가 검색어를 입력한 직후에 작성된다.
② 사용자가 링크를 따라 다른 웹 페이지로 이동하는 비율이 높을수록 댐핑 인자가 커진다.
③ 링크 분석 기법은 웹 페이지 사이의 링크를 분석하여 웹 페이지의 적합도를 값으로 나타낸다.
④ 웹 페이지의 중요도는 다른 웹 페이지에서 받는 값과 다른 웹 페이지에 나눠 주는 값의 합이다.
⑤ 사용자가 검색어를 입력하면 검색 엔진은 검색한 결과를 인덱스에 정렬된 순서대로 화면에 나타낸다.

15. ㉠, ㉡을 고려하여 검색 결과에서 웹 페이지의 순위를 높이기 위한 방안으로 가장 적절한 것은?

① 화제가 되고 있는 검색어들을 웹 페이지에 최대한 많이 나열하여 ㉠을 높인다.
② 사람들이 많이 접속하는 유명 검색 사이트로 연결하는 링크를 웹 페이지에 많이 포함시켜 ㉠을 높인다.
③ 알파벳순으로 앞 순서에 있는 단어들을 웹 페이지 첫 부분에 많이 포함시켜 ㉡을 높인다.
④ 다른 많은 웹 페이지들이 링크하도록 웹 페이지에서 여러 주제를 다루고 전체 글자 수를 많게 하여 ㉡을 높인다.
⑤ 다른 웹 페이지에서 흔히 다루지 않는 주제를 간략하게 설명하되 주제와 관련된 단어를 자주 사용하여 ㉡을 높인다.

16. <보기>는 웹 페이지들의 관계를 도식화한 것이다. 윗글을 바탕으로 <보기>를 이해한 내용으로 적절한 것은? [3점]

<보 기>



원은 웹 페이지이고, 화살표는 웹 페이지에서 링크를 통해 화살표 방향의 다른 웹 페이지로 연결됨을 뜻한다. 댐핑 인자는 0.5이고, d와 e의 중요도는 16으로 고정된 값이다.

(단, 링크와 댐핑 인자 외에 웹 페이지의 중요도에 영향을 주는 다른 요소는 고려하지 않음.)

① a의 중요도는 16이다.
② a가 b와 d로부터 각각 받는 값은 같다.
③ b에서 a로의 링크가 끊어지면 b와 c의 중요도는 같다.
④ e에서 a로의 링크가 추가되면 b의 중요도는 6이다.
⑤ e에서 c로의 링크가 추가되면 c의 중요도는 5이다.

17. 문맥상 ⓐ의 의미와 가장 가까운 것은?

① 공부를 하다 보니 시간은 자정이 넘었다.
② 그들은 큰 산을 넘어서 마을에 도착했다.
③ 철새들이 국경선을 넘어서 훨훨 날아갔다.
④ 선수들은 가까스로 어려운 고비를 넘었다.
⑤ 갑자기 냄비에서 물이 넘어서 좀 당황했다.

- 정답: 14.② 15.⑤ 16.⑤ 17.①

◆ 21 LEET 언어이해 1~3번

[1~3] 다음 글을 읽고 물음에 답하시오.

비즈니스 프로세스는 고객 가치 창출을 위해 기업 또는 조직에서 업무를 처리하는 과정을 말한다. 업무 처리 과정을 업무흐름도로 도식화하는 과정을 프로세스 모델링이라 하며, 그 결과물을 프로세스 모델이라고 한다. 프로세스 모델은 업무 처리 활동 및 활동들 간의 경로로 구성된다. 프로세스 모델이 효율적으로 작동하고 있는지를 확인, 분석, 수정·보완, 개선하는 작업이 필요한데, 프로세스 마이닝은 그중 한 기법이다. 프로세스 마이닝은, 시뮬레이션처럼 실제 이벤트 로그 수집 이전에 정립한 프로세스 모델 중심 분석기법과, 데이터 마이닝처럼 프로세스를 고려하지 않는 데이터 중심 분석기법을 연결하는 역할을 한다.

프로세스 마이닝은 정보시스템을 통해 확보한 이벤트 로그에서 프로세스에 관련된 가치 있는 정보를 추출하는 것이다. 이벤트 로그란 정보시스템에 축적된 비즈니스 프로세스 수행 기록인데, 이것이 프로세스 마이닝의 출발점이 된다. 이벤트 로그는 행과 열로 표현되는 이차원 표 형태이다. 업무 활동으로 발생한 이벤트는 행으로 추가되며, 각 열에는 이벤트의 속성들이 기록된다. 이때 기록되는 속성으로 필수적인 것은 사례 ID, 활동명, 발생 시점이며, 다양한 분석을 위해 그 외 속성들도 추가될 수 있다. 이벤트 로그는 사용자에게 도움이 되는 정보를 직접 제공할 수 없는 원데이터이므로, 그것을 우리가 사용할 수 있는 정보로 변환해 주어야 한다. 프로세스 마이닝에는 프로세스 발견, 적합성 검증, 프로세스 향상의 세 가지 유형이 있다.

프로세스 발견이란 프로세스 분석가가 알고리즘을 통해 이벤트 로그로부터 프로세스 모델을 도출하는 것을 말하는데, 이때 분석가는 별다른 업무 지식 없이도 작업을 수행할 수 있다. 만일 도출된 프로세스 모델이 복잡하여 유의미한 분석이 곤란할 경우, 퍼지 마이닝이나 클러스터링 기법을 활용할 수 있다. 퍼지 마이닝은 실행 빈도가 낮은 활동을 제거 또는 병합하거나, 그 활동들 간의 경로를 제거함으로써 프로세스 모델을 단순화해 주는 기법이다. 이때 프로세스 모델에 나타난 활동과 경로에 대한 임곗값을 설정하여 모델의 복잡도를 조절할 수 있다. 클러스터링은 특성이 유사한 사례들을 같은 그룹으로 묶어주는 기법이다. 전체 이벤트 로그를 대상으로 프로세스를 도출할 때 복잡한 프로세스 모델이 도출될 경우, 이 기법을 적용하여 이벤트 로그를 여러 개로 나눌 수 있다. 이렇게 세분화된 이벤트 로그에 프로세스 발견 기법을 적용하면, 프로세스 모델의 복잡도가 줄어든다.

적합성 검증이란 기존의 프로세스 모델과 이벤트 로그 분석에서 도출된 결과를 비교하여 어느 정도 일치하는지를 확인하는 것이다. 이때 기존의 프로세스 모델과 이벤트 로그에서 도출된 결과물이 불일치하는 경우가 발생하는데, 먼저 기존의 프로세스 모델이 적절함에도 불구하고 업무 담당자가 이를 준수하지 않는 경우를 들 수 있다. 이 경우에는 현실 세계의 실제 업무 수행 실태를 교정해야 한다. 이와 달리 이벤트 로그의 분석 결과물이 더 적절한 것으로 판단되는 경우에는 기존의 프로세스 모델을 수정할 필요가 있다.

프로세스 향상에는 두 유형이 있다. 하나는 기존의 프로세스 모델을 '수정'하는 것이며, 다른 하나는 업무 수행 시간 및 담당자 등 이벤트 로그 분석에서 얻은 부가적 정보를 추가하여 발견된 프로세스 모델을 '확장'하는 것이다. 확장의 예로는 이벤트 로그로부터 도출된 프로세스 모델에 프로세스 내 병목지점과 재작업 흐름을 시각화하는 것을 들 수 있다.

프로세스 마이닝은 데이터 과학에 근거를 두고 프로세스 분석가가 업무 전문가와 협업하여 기업이 수행하는 비즈니스 프로세스에 대한 문제점을 진단하고 개선 방안을 도출하는 데 기여할 수 있다.

1. 윗글과 일치하는 것은?

① 이벤트 로그는 프로세스 마이닝의 출발점이지만 그 자체로는 유용한 정보라 할 수 없다.

② 업무 전문가의 충분한 지식 없이 이벤트 로그로부터 프로세스 모델을 도출하기는 어렵다.

③ 프로세스 발견은 프로세스에 내재된 업무 관련 규정을 이벤트 로그로부터 도출하는 것이다.

④ 클러스터링은 복잡한 프로세스 모델을 여러 개의 세부 프로세스 모델로 구분해 주는 기법이다.

⑤ 이벤트 로그에서 업무 담당자를 파악하여 기존의 프로세스 모델에 활동과 경로를 추가하는 것은 프로세스 수정이다.

2. '프로세스 마이닝'에 대해 추론한 것으로 적절하지 <u>않은</u> 것은?

① 프로세스 마이닝을 도입하면 내부 규정의 준수 여부에 대한 감독이 용이해진다.

② 프로세스 마이닝을 통해 기존의 프로세스 모델이 실제로 어떻게 수행되는가를 파악할 수 있다.

③ 프로세스 마이닝은 판에 박힌 단순한 업무뿐 아니라 비정형적인 업무 처리 과정의 분석에도 활용된다.

④ 프로세스 마이닝은 예상된 이벤트 로그에 적용할 프로세스 모델 중심의 업무 성과 분석 및 개선 기법이다.

⑤ 프로세스 마이닝은 기존의 프로세스 모델뿐 아니라 발견으로 도출된 프로세스 모델을 향상하는 데에도 활용된다.

3. <보기>의 사례에 프로세스 마이닝을 적용할 때 가장 적절한 것은?

<보 기>

○○병원에서는 외래 환자의 과도한 대기 시간을 줄이고 의료 서비스의 품질을 개선하기 위해 외래 환자 진료 프로세스를 분석하고자 한다. 이 병원에서는 질환별로 진행해야 하는 표준 진료 프로세스를 임상진료 지침으로 수립해 두고 있다. 프로세스 마이닝 도구를 사용하여 프로세스 모델을 도출하였더니 지나치게 복잡한 프로세스 모델이 도출되어 분석이 곤란한 상황이다. 또한 환자의 민감한 개인 의료정보가 저장된 이벤트 로그를 프로세스 분석가에게 제공할 경우 정보 보호 및 프라이버시 이슈가 존재하고, 병원의 기밀이 유출될 우려가 제기되어 이를 해결하고자 한다.

① 복잡도 문제를 해결하기 위해 연령 및 질환을 기준으로 이벤트 로그의 사례를 클러스터링 하려면 필수적 속성만 이벤트 로그에 있어도 된다.

② 적합성 검증 결과 기존의 프로세스 모델과 이벤트 로그 분석 결과가 불일치하면 의료진에 대한 제재 조치나 지침 재교육이 필수적이다.

③ 이벤트 속성의 임곗값을 조절하여 빈번하게 수행되는 진료 프로세스 수행 패턴을 파악할 수 있다.

④ 환자의 개인정보 보호를 위해 사례 ID를 제외하고 이벤트 로그를 작성해야 한다.

⑤ 외래 환자의 대기 시간 분석을 위해서는 프로세스 확장이 필요하다.

- 정답: 1.① 2.④ 3.⑤

[16~18] 다음 글을 읽고 물음에 답하시오.

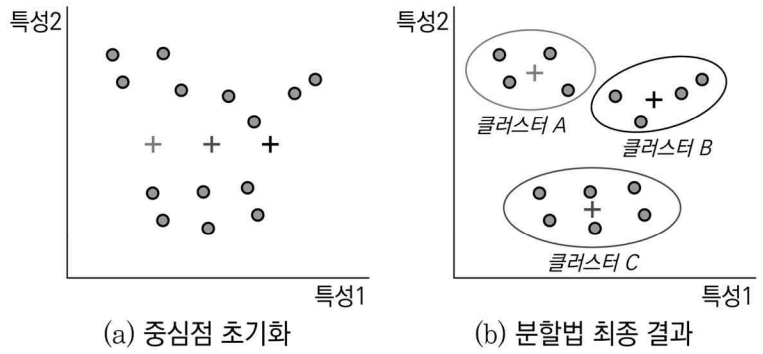
대규모 데이터를 분석하여 데이터 속에 숨어 있는 유용한 패턴을 찾아내기 위해 다양한 기계학습 기법이 활용되고 있다. 기계학습을 위한 입력 자료를 데이터 세트라고 하며, 이를 분석하여 유용하고 가치 있는 정보를 추출할 수 있다. 데이터 세트의 각 행에는 개체에 대한 구체적인 정보가 저장되며, 각 열에는 개체의 특성이 기록된다. 개체의 특성은 범주형과 수치형으로 구분되는데, 예를 들어 '성별'은 범주형이며, '체중'은 수치형이다.

기계학습 기법의 하나인 클러스터링은 데이터의 특성에 따라 유사한 개체들을 묶는 기법이다. 클러스터링은 분할법과 계층법으로 나뉘는데, 이 둘은 모두 거리 개념에 기초하고 있다. 가장 많이 사용되는 거리 개념은 기하학적 거리이며, 두 개체 사이의 거리는 n차원으로 표현된 공간에서 두 개체를 점으로 표시할 때 두 점 사이의 직선거리이다. 거리를 계산할 때 특성들의 단위가 서로 다른 경우가 많은데, 이런 경우 특성 값을 정규화할 필요가 있다. 예를 들어 특정 과목의 학점과 출석 횟수를 기준으로 학생들을 묶을 경우 두 특성의 단위가 다르므로 두 특성 값을 모두 0과 1 사이의 값으로 정규화하여 클러스터링을 수행한다. 또한 범주형 특성에 거리 개념을 적용하려면 이를 수치형 특성으로 변환해야 한다.

분할법은 전체 데이터 개체를 사전에 정한 개수의 클러스터로 구분하는 기법으로, 모든 개체는 생성된 클러스터 가운데 어느 하나에 속한다. <그림 1>에서 (b)는 (a)에 제시된 개체들을 분할법을 통해 세 개의 클러스터로 묶은 예이다. 분할법에서는 클러스터에 속한 개체들의 좌표 평균을 계산하여 클러스터 중심점을 구한다. 고전적인 분할법인 K-민즈 클러스터링(K-means clustering)에서는 거리 개념과 중심점에 기반하여 다음과 같은 과정으로 알고리즘이 진행된다.

1) 사전에 K개로 정한 클러스터 중심점을 임의의 위치에 배치하여 초기화한다.
2) 각 개체에 대해 K개의 중심점과의 거리를 계산한 후 가장 가까운 중심점에 해당 개체를 배정하여 클러스터를 구성한다.
3) 클러스터 별로 그에 속한 개체들의 좌표 평균을 계산하여 클러스터의 중심점을 다시 구한다.
4) 2)와 3)의 과정을 반복해서 수행하여 더 이상 변화가 없는 상태에 도달하면 알고리즘이 종료된다.

분할법에서는 이와 같이 개체와 중심점과의 거리를 계산하여 클러스터에 개체를 배정하므로 두 개체가 인접해 있더라도 가장 가까운 중심점이 서로 다르면 두 개체는 상이한 클러스터에 배정된다.

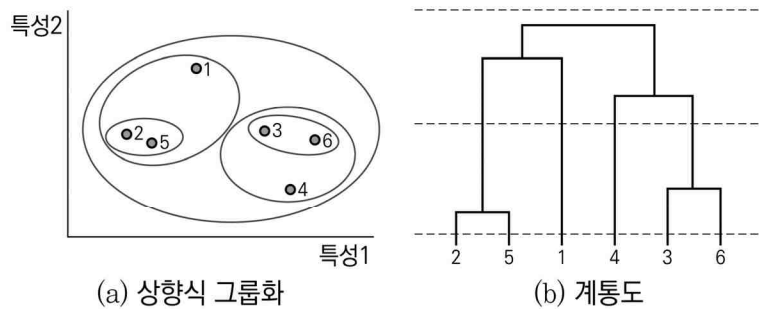


(a) 중심점 초기화 (b) 분할법 최종 결과

<그림 1> 분할법의 예

클러스터링이 잘 수행되었는지 확인하려면 클러스터링 결과를 평가하는 품질 지표가 필요하다. K-민즈 클러스터링의 경우 품질 지표는 개체와 그 개체가 해당하는 클러스터의 중심점 간 거리의 평균이다. K-민즈 클러스터링에서 K가 정해졌을 때 개체와 해당 중심점 간 거리의 평균을 최소화하는 '전체 최적해'는 확정적으로 보장되지 않는다. 알고리즘의 첫 번째 단계인 초기화를 어떻게 하느냐에 따라 클러스터링 결과가 달라질 수 있으며, 경우에 따라 좋은 결과를 찾는 데 실패할 수도 있다. 따라서 전체 최적해를 얻을 확률을 높이기 위해, 서로 다른 초기화를 시작으로 클러스터링 알고리즘을 여러 번 수행하여 나온 결과 중에 좋은 해를 찾는 방법이 흔히 사용된다. 그런데 K-민즈 클러스터링 알고리즘의 한 가지 문제는 클러스터의 개수인 K를 미리 정해야 한다는 것이다. K가 커질수록 각 개체와 해당 중심점 간 거리의 평균은 감소한다. 극단적으로 모든 개체를 클러스터로 구분할 경우 개체가 곧 중심점이므로 이들 사이의 거리의 평균값은 0으로 최소화되지만, 클러스터링의 목적에 부합하는 유용한 결과라고 보기 어렵다. 따라서 작은 수의 K로 알고리즘을 시작하여 클러스터링 결과를 구한 다음 K를 점차 증가시키면서 유의미한 품질 향상이 있는지 확인하는 방법이 자주 사용된다.

한편, 계층법은 클러스터 개수를 사전에 정하지 않아도 되는 장점이 있다. <그림 2>와 같이 개체들을 거리가 가까운 것들부터 차근차근 집단으로 묶어서 모든 개체가 하나로 묶일 때까지 추상화 수준을 높여가는 상향식으로 알고리즘이 진행되어 계통도를 산출한다. 따라서 계층법은 개체들 간에 위계 관계가 있는 경우에 효과적으로 적용될 수 있다. 계통도에서 점선으로 표시된 수평선을 아래위로 이동해 가면서 클러스터링의 추상화 수준을 변경할 수 있다.



(a) 상향식 그룹화 (b) 계통도

<그림 2> 계층법의 예

16. 윗글의 내용과 일치하는 것은?

① 클러스터링은 개체들을 묶어서 한 개의 클러스터로 생성하는 기법이다.

② 분할법에서는 클러스터링 수행자가 정확한 계산을 통해 초기 중심점을 찾아낸다.

③ 분할법은 하향식 클러스터링 기법이므로 한 개체가 여러 클러스터에 속할 수 있다.

④ 계층법으로 계통도를 산출할 때 클러스터 개수는 미리 정하지 않는다.

⑤ 계층법의 계통도에서 수평선을 아래로 내릴 경우 추상화 수준이 높아진다.

17. K-민즈 클러스터링에 대해 추론한 것으로 적절하지 않은 것은?

① 특성이 유사한 두 개체가 서로 다른 클러스터에 배치될 수 있다.

② 초기 중심점의 배치 위치에 따라 클러스터링의 품질이 달라질 수 있다.

③ 클러스터 개수를 감소시키면 클러스터링 결과의 품질 지표 값은 증가한다.

④ 초기화를 다르게 하면서 알고리즘을 여러 번 수행하면 전체 최적해가 결정된다.

⑤ K를 정하여 알고리즘을 진행하면 각 클러스터의 중심점은 결국 고정된 점에 도달한다.

18. <보기>의 사례에 클러스터링을 적용할 때 적절하지 않은 것은?

<보 기>

○○기업에서는 표적 시장을 선정하여 마케팅을 실행하기 위해 전체 시장을 세분화하고자 한다. 시장 세분화를 위해 특성이 유사한 고객을 묶는 기계학습 기법 도입을 검토 중이다. 이 기업에서는 고객의 거주지, 성별, 나이, 소득 수준 등 인구통계학적인 정보와 라이프 스타일에 관한 정보 등을 보유하고 있다.

① 고객 정보에는 수치형이 아닌 것도 있어 특성의 유형 변환이 요구된다.

② 고객 특성은 세분화 과정을 통해 계통도로 표현 가능하므로 계층법이 효과적이다.

③ K-민즈 클러스터링 알고리즘을 실행하려면 세분화할 시장의 개수를 먼저 정해야 한다.

④ 나이와 소득수준과 같이 단위가 다른 특성을 기준으로 시장을 세분화할 경우 정규화가 필요하다.

⑤ 모든 고객을 별도의 세분화된 시장들로 구분하여 1:1 마케팅을 할 경우 K-민즈 클러스터링의 품질 지표 값은 0이다.

- 정답: 16.④ 17.④ 18.④