



(19) 대한민국특허청(KR)
(12) 등록특허공보(B1)

(45) 공고일자 2017년05월30일
(11) 등록번호 10-1741408
(24) 등록일자 2017년05월24일

(51) 국제특허분류(Int. Cl.)
G06F 17/30 (2006.01)
(52) CPC특허분류
G06F 17/30713 (2013.01)
G06F 17/30716 (2013.01)
(21) 출원번호 10-2016-0134313
(22) 출원일자 2016년10월17일
심사청구일자 2016년10월17일
(56) 선행기술조사문헌
JP2016095684 A

(73) 특허권자
한국기술교육대학교 산학협력단
충남 천안시 동남구 병천면 충절로 1600, 내 (한국기술교육대학교)
(72) 발명자
한연희
대전광역시 유성구 대덕대로 596, 506호(로얄밸리)
허주성
충청남도 천안시 동남구 병천면 가전8길 94, A동 401호(미래빌리지)
(74) 대리인
김견수

전체 청구항 수 : 총 14 항

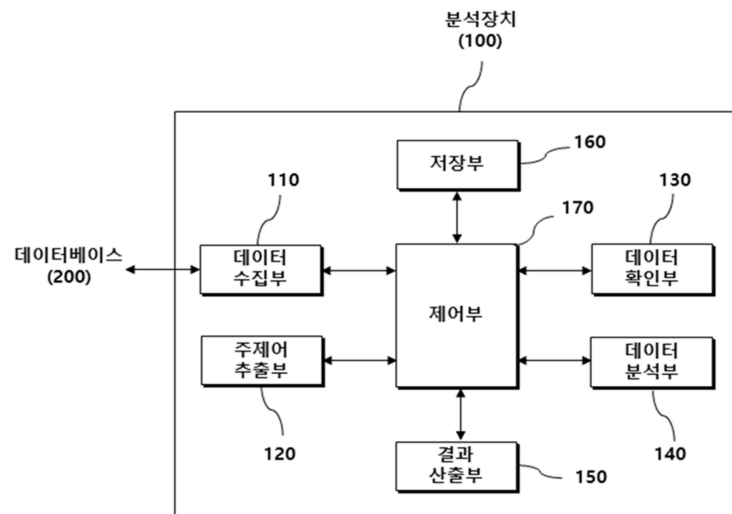
심사관 : 이복현

(54) 발명의 명칭 **대용량 학술논문정보 분석 시스템**

(57) 요약

본 발명은 특정 기술 분야에서 수집 년간 발행된 많은 수의 학술논문정보의 서지정보로부터 주제어를 도출하고, 도출된 주제어를 클러스터링 분석(예를 들어, K-평균 군집화(K-means clustering) 기법)을 수행하여 연대별 주제어 출현 패턴을 분석함으로써, 과거에서부터 현재까지의 연구현황 및 추이를 손쉽게 확인할 수 있도록 하는 대용량 학술논문정보 분석 시스템에 관한 것이다.

대표도 - 도2



(52) CPC특허분류
G06F 2216/03 (2013.01)

명세서

청구범위

청구항 1

특정 기술부분의 학술논문정보로부터 주제어를 추출하는 주제어 추출부;

상기 학술논문정보를 기 설정된 연대별로 구분하고, 연대별로 구분한 학술논문정보로부터 각각의 주제어에 대한 연대별 출현 빈도를 확인하는 데이터 확인부;

상기 각각의 주제어에 대한 연대별 출현 빈도의 클러스터링 분석을 수행하여 적어도 하나 이상의 그룹으로 군집화하는 데이터 분석부; 및

군집화한 적어도 하나 이상의 그룹을 특정 기술부분의 연구현황 및 추이와 관련된 적어도 하나 이상의 패턴으로 구분하여 결과를 출력하는 결과 산출부;를 포함하는 것을 특징으로 하는 대용량 학술논문정보 분석 시스템.

청구항 2

청구항 1에 있어서,

상기 주제어 추출부는,

저장부에 기 저장되어 있는 용어 테이블을 참조하여 상기 학술논문정보의 서지정보에 나타나는 용어가 기 설정되어 있는 임계값 이상이면 해당 용어를 주제어로 추출하는 것을 특징으로 하는 대용량 학술논문정보 분석 시스템.

청구항 3

청구항 1에 있어서,

상기 데이터 확인부는,

상기 학술논문정보를 기 설정된 연대별로 구분할 때, 연대별 학술논문정보의 수가 기 설정되어 있는 임계값 이하이면 해당 연대 및 해당 연대에 속하는 학술논문정보를 분석 대상에서 제외하는 것을 특징으로 하는 대용량 학술논문정보 분석 시스템.

청구항 4

청구항 1에 있어서,

상기 클러스터링 분석은,

n 개의 주제어에 대한 연대별 출현 빈도를 k 개의 그룹으로 군집화하는 K -평균 군집화 기법을 사용하는 것을 특징으로 하는 대용량 학술논문정보 분석 시스템.

청구항 5

청구항 4에 있어서,

상기 K -평균 군집화 기법은,

k 개의 초기 점(initial point)을 임의로 선택한 후 각각의 주제어를 상기 주제어와 가장 근접한 초기 점에 배정하여 k 개의 군집을 생성하고,

생성된 k 개의 군집에 속하는 각각의 주제어의 평균점을 계산한 후 상기 k 개의 초기 점을 새롭게 계산된 평균점으로 조정하며,

k 개의 군집에 속하는 각각의 주제어의 평균점이 더 이상 조정되지 않을 때까지 상기 평균점 조정을 반복하여 수행하는 것을 특징으로 하는 대용량 학술논문정보 분석 시스템.

청구항 6

청구항 1에 있어서,

상기 데이터 분석부는,

클러스터링 분석을 수행하기 이전에 각각의 주제어에 대한 연대별 출현 빈도의 정규화를 수행하는 것을 특징으로 하는 대용량 학술논문정보 분석 시스템.

청구항 7

청구항 1에 있어서,

상기 연구현황 및 추이와 관련된 패턴은,

최근 증가하고 있는 군집(steadily increase & until recently), 마지막 연대에 하강하고 있는 군집(steadily increase but recently decrease), 과거 특정 연대에서 급증한 군집(once popular), 꾸준히 하강하고 있는 군집(mostly decrease) 중 적어도 하나 이상을 포함하는 것을 특징으로 하는 대용량 학술논문정보 분석 시스템.

청구항 8

분석 장치에서, 특정 기술부분의 학술논문정보로부터 주제어를 추출하는 주제어 추출 단계;

상기 분석 장치에서, 상기 학술논문정보를 기 설정된 연대별로 구분하고, 연대별로 구분한 학술논문정보로부터 상기 주제어 추출 단계에서 추출한 각각의 주제어에 대한 연대별 출현 빈도를 확인하는 데이터 확인 단계;

상기 분석 장치에서, 상기 데이터 확인 단계에서 확인한 상기 각각의 주제어에 대한 연대별 출현 빈도의 클러스터링 분석을 수행하여 적어도 하나 이상의 그룹으로 군집화하는 데이터 분석 단계; 및

상기 분석 장치에서, 상기 데이터 분석 단계를 통해 군집화한 적어도 하나 이상의 그룹을 특정 기술부분의 연구현황 및 추이와 관련된 적어도 하나 이상의 패턴으로 구분하여 결과를 출력하는 결과 산출 단계;를 포함하는 것을 특징으로 하는 대용량 학술논문정보 분석 방법.

청구항 9

청구항 8에 있어서,

상기 주제어 추출 단계는,

저장부에 기 저장되어 있는 용어 테이블을 참조하여 상기 학술논문정보의 서지정보에 나타나는 용어가 기 설정되어 있는 임계값 이상이면 해당 용어를 주제어로 추출하는 것을 특징으로 하는 대용량 학술논문정보 분석 방법.

청구항 10

청구항 8에 있어서,

상기 데이터 확인 단계는,

상기 학술논문정보를 기 설정된 연대별로 구분할 때, 연대별 학술논문정보의 수가 기 설정되어 있는 임계값 이하이면 해당 연대 및 해당 연대에 속하는 학술논문정보를 분석 대상에서 제외하는 것을 특징으로 하는 대용량 학술논문정보 분석 방법.

청구항 11

청구항 8에 있어서,

상기 데이터 분석 단계에서 수행하는 상기 클러스터링 분석은,

n 개의 주제어에 대한 연대별 출현 빈도를 k 개의 그룹으로 군집화하는 K -평균 군집화 기법을 사용하는 것을 특징으로 하는 대용량 학술논문정보 분석 방법.

청구항 12

청구항 11에 있어서,

상기 K-평균 군집화 기법은,

k개의 초기 점을 임의로 선택한 후 각각의 주제어를 상기 주제어와 가장 근접한 초기 점에 배정하여 k개의 군집을 생성하는 군집 생성 단계;

상기 군집 생성 단계를 통해 생성된 k개의 군집에 속하는 각각의 주제어의 평균점을 계산한 후 상기 k개의 초기 점을 새롭게 계산된 평균점으로 조정하는 평균점 조정 단계; 및

k개의 군집에 속하는 각각의 주제어의 평균점이 더 이상 조정되지 않을 때까지 상기 평균점 조정 단계를 반복하여 수행하는 반복 수행 단계;를 포함하는 것을 특징으로 하는 대용량 학술논문정보 분석 방법.

청구항 13

청구항 8에 있어서,

상기 데이터 분석 단계는,

클러스터링 분석을 수행하기 이전에 각각의 주제어에 대한 연대별 출현 빈도의 정규화를 수행하는 것을 특징으로 하는 대용량 학술논문정보 분석 방법.

청구항 14

청구항 8에 있어서,

상기 결과 산출 단계에서의 연구현황 및 추이와 관련된 패턴은,

최근 증가하고 있는 군집, 마지막 연대에 하강하고 있는 군집, 과거 특정 연대에서 급증한 군집, 꾸준히 하강하고 있는 군집 중 적어도 하나 이상을 포함하는 것을 특징으로 하는 대용량 학술논문정보 분석 방법.

발명의 설명

기술 분야

[0001] 본 발명은 특정 기술 부분에서 수십 년간 발행된 많은 수의 학술논문정보의 서지정보로부터 주제어를 도출하고, 도출된 주제어를 클러스터링 분석(예를 들어, K-평균 군집화(k-means clustering) 기법)을 수행하여 연대별 주제어 출현 패턴을 분석함으로써, 과거에서부터 현재까지의 연구현황 및 추이를 손쉽게 확인할 수 있도록 하는 대용량 학술논문정보 분석 시스템에 관한 것이다.

배경 기술

[0002] 일반적으로 데이터 마이닝(Data Mining)은 방대하고 복잡한 데이터 내부에 존재하는 의미 있는 정보를 이끌어 내는 방법을 연구하는 학문으로서, 데이터베이스로부터 과거에는 알지 못했지만 데이터 속에서 유도된 새로운 데이터 모델을 발견하여 미래에 실행 가능한 정보를 추출해 내고 의사 결정에 이용하는 과정을 말한다. 즉 데이터에 숨겨진 패턴과 관계를 찾아내어 광맥을 찾아내듯이 정보를 발견해 내는 것이다.

[0003] 주로 숫자형태의 일정한 데이터 구조로 정형화된 데이터(structured data)를 분석해오던 데이터 마이닝의 연구자들은 최근 들어 텍스트, 이미지, 동영상, 음성 등과 같이 구조화되지 않은 비정형 데이터(unstructured data)에 관심을 기울이고 있다. 특히, 비정형 데이터 중에서 최근 인터넷 사용자의 폭발적인 증가에 힘입어, 웹 마이닝(Web Mining)과 텍스트 마이닝(Text Mining)의 중요성이 더욱 부각되고 있다.

[0004] 또한, 다양한 분야에서 대용량의 텍스트 데이터로부터 연구동향의 의미있는 패턴의 식별과 잠재적인 지식의 추출은 중요한 작업이다. 특히, 고속 인터넷의 출현은 다양한 형태의 대용량 텍스트 데이터를 생성한다. 하지만, 텍스트 데이터 구조 측면에서 불균일 및 복잡성이 존재하기 때문에 텍스트 데이터를 분석하기 대단히 어렵다.

[0005] 상기 텍스트 마이닝은 텍스트에서 유용하고 의미있는 정보를 도출하는 과정으로서, 정형화된 데이터 세트에서 패턴을 추출하는 기존의 데이터 마이닝 작업과는 달리 비정형화된 데이터 세트에 저장된 객체 간의 관계를 탐구하기 위한 것이다. 일반적으로 텍스트 마이닝은, (1) 비정형화된 텍스트 입력으로부터 정형화된 데이터베이스의 구축, (2) 정형화된 데이터로부터 패턴과 경향의 추출, 및 (3) 패턴과 경향의 평가 및 해석의 여러 단계를 포함

한다.

- [0006] 대용량 텍스트 데이터로부터 암시적 정보를 추출하기 위하여 많은 연구가 수행되었으며, 최근 일부 연구자들에 의해 대용량 특허문헌으로부터 새로운 기술 동향을 파악하거나 학술논문으로부터 연구 경향을 파악하는 것이 시도되고 있다.
- [0007] 예를 들어, 종래의 연구 중에서 논문에 자주 나타난 키워드의 의미있는 연상패턴을 찾기 위해 클러스터링 분석 뿐만 아니라 저자원의 임베딩 방법, 연관 규칙, 소셜 네트워크 분석을 적용하였으나, 기존의 연구는 단지 하나의 저널에 발표된 수천 건의 논문을 대상으로 수행된 것에 불과하여 분석대상의 범위가 넓어지는 경우 분석결과 의 정확도가 저하되고 이로 인해 분석결과를 신뢰하기 어려운 문제점이 있었다.
- [0008] 또한, 특허문헌 및 학술논문 이외에, 생물학의 유전자 정보, 기술경영 정보와 같은 대용량의 텍스트 데이터에 대한 분석 및 연구가 다양하게 시도되고는 있지만, 수십 년간 방대한 자료가 축적되어 있고 현재에도 동적으로 변화하고 있는 특성을 지닌 대용량의 텍스트 데이터를 짧은 시간 내에 분석하여 향후 기술현황 및 추이를 확인 하기 대단히 어려운 문제점이 있었다.
- [0009] 따라서 본 발명에서는 수십 년 동안 발표되어 누적되어 있는 대용량의 학술논문정보의 서지정보로부터 주제어를 도출하고, 도출한 주제어를 클러스터링 기법을 사용하여 분석함으로써, 과거에서부터 현재까지의 특정 기술 부 분의 연대별 주제어 출현 패턴에 따른 연구현황 및 추이를 일목요연하게 확인할 수 있는 방안을 제시하고자 한 다.
- [0010] 다음으로 본 발명의 기술 분야에 존재하는 선행기술에 대하여 간단하게 설명하고, 이어서 본 발명이 상기 선행 기술에 비해서 차별적으로 이루고자 하는 기술적 사항에 대해서 기술하고자 한다.
- [0011] 먼저 한국등록특허 제0598134호(2006.06.30.등록)는 모바일 환경에서 사용되는 벡터 형태의 데이터를 압축하는 방법 및 시스템에 관한 것으로, 특히 K-평균 군집화 및 사전기반 압축 기법을 통해 공간 데이터가 차지하는 저 장 공간의 크기를 효율적으로 줄일 수 있도록 하는 K-평균 군집화를 이용한 벡터 데이터 압축 방법 및 시스템에 관한 것이다.
- [0012] 상기 선행기술은 벡터 데이터에 대하여 최소의 연산을 통해 구분하기 어려운 정도의 위치 정확도 손실 차이를 가지면서 향상된 압축률을 얻을 수 있게 되며, 이에 따라 모바일 기기에서 차지하는 저장공간을 줄일 수 있는 효과를 지니고 있다.
- [0013] 하지만, 본 발명은 대용량 학술논문정보의 서지정보로부터 도출한 주제어를 클러스터링 기법을 통해 분석하여 특정 기술 부분의 연대별 주제어 출현 패턴에 따른 연구현황 및 추이를 확인하는 기술적 구성을 제시하고 있기 때문에, 상기 선행기술의 K-평균 군집화를 이용한 벡터 데이터 압축에 관련된 기술 구성과 비교해 볼 때 기술적 특징의 차이점이 분명하다.
- [0014] 또한, 한국공개특허 제2016-0113826호(2016.10.04.공개)는 빅데이터의 클러스터링 시간을 줄이고 클러스터 중심 의 정확도를 향상시키는 빅데이터 클러스터링을 위한 K-Means 초기 중심 선정 방법에 대한 기술이다.
- [0015] 상기 선행기술은 계통임의추출법을 적용한 K-Means 초기 중심 선정 방법을 제공함으로써, 빅데이터의 클러스터 링 시간을 줄이고 클러스터 중심의 정확도를 향상시킬 수 있는 효과를 제공한다.
- [0016] 하지만, 상기 선행기술은 빅데이터의 클러스터링 시간을 줄이고 중심의 정확도를 향상하기 위한 구성을 제시하 고 있을 뿐, 본 발명에서와 같이 대용량 학술논문정보의 서지정보로부터 주제어를 도출하는 구성, 및 도출된 주 제어를 클러스터링 기법을 사용하여 연대별 주제어 출현 패턴을 분석하는 기술적 구성을 전혀 제시하지 못하기 때문에 양자의 기술적 구성의 차이는 분명하다.
- [0017] 즉 상기 언급한 각각의 선행기술들은 K-평균 군집화 기법을 사용하여 데이터 처리를 수행하는 기술적 구성을 제 시하고는 있으나, 수십 년간 축적되어 있는 수백만 건의 대용량 학술논문정보의 서지정보로부터 도출한 주제어 를 클러스터링 기법(예를 들어, K-평균 군집화 기법)을 통해 연대별 주제어 출현 패턴을 분석하여 특정 기술 부 분의 연대별 주제어 출현 패턴에 따른 연구현황 및 추이를 확인하도록 하는 본 발명의 구성과는 기술적 차이점 이 분명한 것이다.

발명의 내용

해결하려는 과제

[0018] 본 발명은 상기와 같은 문제점을 해결하기 위해 창작된 것으로서, 특정 기술 부분에서 수십 년간 발행된 많은 수의 학술논문정보의 서지정보(예를 들어, 제목, 초록 등)로부터 주제어를 도출하여 고차원의 행렬을 구성하고, 이를 클러스터링 기법을 사용하여 연대별 주제어 출현 패턴을 분석함으로써, 과거에서부터 현재까지의 연구현황 및 추이를 손쉽게 확인할 수 있도록 하는 대용량 학술논문정보 분석 시스템을 제공하는 것을 목적으로 한다.

[0019] 또한, 본 발명은 비정형화된 데이터인 대용량의 학술논문정보의 서지정보를 활용하여 기술동향과 관련된 데이터 분석을 수행함으로써, 웹 마이닝과 텍스트 마이닝 분야의 연구를 활성화할 수 있도록 하는 대용량 학술논문정보 분석 시스템을 제공하는 것을 다른 목적으로 한다.

과제의 해결 수단

[0020] 본 발명의 일 실시예에 따른 대용량 학술논문정보 분석 시스템은, 특정 기술부분의 학술논문정보로부터 주제어를 추출하는 주제어 추출부; 상기 학술논문정보를 기 설정된 연대별로 구분하고, 연대별로 구분한 학술논문정보로부터 각각의 주제어에 대한 연대별 출현 빈도를 확인하는 데이터 확인부; 상기 각각의 주제어에 대한 연대별 출현 빈도의 클러스터링 분석을 수행하여 적어도 하나 이상의 그룹으로 군집화하는 데이터 분석부; 및 군집화한 적어도 하나 이상의 그룹을 특정 기술부분의 연구현황 및 추이와 관련된 적어도 하나 이상의 패턴으로 구분하여 결과를 출력하는 결과 산출부;를 포함하는 것을 특징으로 한다.

[0021] 그리고 상기 주제어 추출부는, 저장부에 기 저장되어 있는 용어 테이블을 참조하여 상기 학술논문정보의 서지정보에 나타나는 용어가 기 설정되어 있는 임계값 이상이면 해당 용어를 주제어로 추출하는 것을 특징으로 한다.

[0022] 그리고 상기 데이터 확인부는, 상기 학술논문정보를 기 설정된 연대별로 구분할 때, 연대별 학술논문정보의 수가 기 설정되어 있는 임계값 이하이면 해당 연대 및 해당 연대에 속하는 학술논문정보를 분석 대상에서 제외하는 것을 특징으로 한다.

[0023] 그리고 상기 클러스터링 분석은, n 개의 주제어에 대한 연대별 출현 빈도를 k 개의 그룹으로 군집화하는 K -평균 군집화 기법을 사용하는 것을 특징으로 한다.

[0024] 이때 상기 K -평균 군집화 기법은, k 개의 초기 점(initial point)을 임의로 선택한 후 각각의 주제어를 상기 주제어와 가장 근접한 초기 점에 배치하여 k 개의 군집을 생성하고, 생성된 k 개의 군집에 속하는 각각의 주제어의 평균점을 계산한 후 상기 k 개의 초기 점을 새롭게 계산된 평균점으로 조정하며, k 개의 군집에 속하는 각각의 주제어의 평균점이 더 이상 조정되지 않을 때까지 상기 평균점 조정을 반복하여 수행하는 것을 특징으로 한다.

[0025] 그리고 상기 데이터 분석부는, 클러스터링 분석을 수행하기 이전에 각각의 주제어에 대한 연대별 출현 빈도의 정규화를 수행하는 것을 특징으로 한다.

[0026] 그리고 상기 연구현황 및 추이와 관련된 패턴은, 최근 증가하고 있는 군집(steadily increase & until recently), 마지막 연대에 하강하고 있는 군집(steadily increase but recently decrease), 과거 특정 연대에서 급증한 군집(once popular), 꾸준히 하강하고 있는 군집(mostly decrease) 중 적어도 하나 이상을 포함하는 것을 특징으로 한다.

[0027] 또한, 본 발명의 일 실시예에 따른 대용량 학술논문정보 분석 방법은, 분석 장치에서, 특정 기술부분의 학술논문정보로부터 주제어를 추출하는 주제어 추출 단계; 상기 분석 장치에서, 상기 학술논문정보를 기 설정된 연대별로 구분하고, 연대별로 구분한 학술논문정보로부터 상기 주제어 추출 단계에서 추출한 각각의 주제어에 대한 연대별 출현 빈도를 확인하는 데이터 확인 단계; 상기 분석 장치에서, 상기 데이터 확인 단계에서 확인한 상기 각각의 주제어에 대한 연대별 출현 빈도의 클러스터링 분석을 수행하여 적어도 하나 이상의 그룹으로 군집화하는 데이터 분석 단계; 및 상기 분석 장치에서, 상기 데이터 분석 단계를 통해 군집화한 적어도 하나 이상의 그룹을 특정 기술부분의 연구현황 및 추이와 관련된 적어도 하나 이상의 패턴으로 구분하여 결과를 출력하는 결과 산출 단계;를 포함하는 것을 특징으로 한다.

[0028] 그리고 상기 주제어 추출 단계는, 저장부에 기 저장되어 있는 용어 테이블을 참조하여 상기 학술논문정보의 서지정보에 나타나는 용어가 기 설정되어 있는 임계값 이상이면 해당 용어를 주제어로 추출하는 것을 특징으로 한다.

[0029] 그리고 상기 데이터 확인 단계는, 상기 학술논문정보를 기 설정된 연대별로 구분할 때, 연대별 학술논문정보의 수가 기 설정되어 있는 임계값 이하이면 해당 연대 및 해당 연대에 속하는 학술논문정보를 분석 대상에서 제외하는 것을 특징으로 한다.

- [0030] 그리고 상기 데이터 분석 단계에서 수행하는 상기 클러스터링 분석은, n 개의 주제어에 대한 연대별 출현 빈도를 k 개의 그룹으로 군집화하는 K -평균 군집화 기법을 사용하는 것을 특징으로 한다.
- [0031] 이때 상기 K -평균 군집화 기법은, k 개의 초기 점을 임의로 선택한 후 각각의 주제어를 상기 주제어와 가장 근접한 초기 점에 배정하여 k 개의 군집을 생성하는 군집 생성 단계; 상기 군집 생성 단계를 통해 생성된 k 개의 군집에 속하는 각각의 주제어의 평균점을 계산한 후 상기 k 개의 초기 점을 새롭게 계산된 평균점으로 조정하는 평균점 조정 단계; 및 k 개의 군집에 속하는 각각의 주제어의 평균점이 더 이상 조정되지 않을 때까지 상기 평균점 조정 단계를 반복하여 수행하는 반복 수행 단계;를 포함하는 것을 특징으로 한다.
- [0032] 그리고 상기 데이터 분석 단계는, 클러스터링 분석을 수행하기 이전에 각각의 주제어에 대한 연대별 출현 빈도의 정규화를 수행하는 것을 특징으로 한다.
- [0033] 그리고 상기 결과 산출 단계에서의 연구현황 및 추이와 관련된 패턴은, 최근 증가하고 있는 군집, 마지막 연대에 하강하고 있는 군집, 과거 특정 연대에서 급증한 군집, 꾸준히 하강하고 있는 군집 중 적어도 하나 이상을 포함하는 것을 특징으로 한다.

발명의 효과

- [0034] 이상에서와 같이 본 발명의 대용량 학술논문정보 분석 시스템에 따르면, 수십 년간 발행된 대용량의 학술논문정보로부터 도출한 주제어를 클러스터링 기법(예를 들어, K -평균 군집화)을 통해 분석함으로써, 과거에서부터 현재까지의 특정 기술 분야의 연대별 주제어 출현 패턴에 따른 연구현황 및 추이를 손쉽게 확인할 수 있는 효과가 있다.
- [0035] 또한, 비정형화된 데이터인 학술논문정보를 활용한 기술현황 및 추이와 관련된 데이터 분석을 통해 웹 마이닝과 텍스트 마이닝 분야의 연구를 크게 활성화시킬 수 있는 효과가 있다.
- [0036] 또한, 지금까지 축적된 대량의 신문기사, 연애기사, 온라인 쇼핑정보 등에 포함된 주제어를 클러스터링 기법을 통해 분석할 수 있으므로 본 발명에서 제시한 특정 기술 분야의 학술논문정보의 연대별 연구현황 및 추이 이외에 다양한 분야에 폭넓게 적용할 수 있는 효과가 있다.

도면의 간단한 설명

- [0037] 도 1은 본 발명의 일 실시예에 따른 대용량 학술논문정보 분석 시스템의 구성을 개략적으로 나타낸 도면이다.
- 도 2는 도 1의 분석 장치의 구성을 보다 상세하게 나타낸 도면이다.
- 도 3은 본 발명의 분석 시스템에 사용된 IT 용어의 수와 추출한 주제어의 수를 나타낸 도면이다.
- 도 4는 도 3에서 추출한 주제어의 일부 리스트를 나타낸 도면이다.
- 도 5는 각 연대별 논문의 수와 각 주제어의 연대별 출현 빈도를 나타낸 데이터세트이다.
- 도 6은 K -평균 군집화의 결과를 이용하여 각 군집에 해당하는 주제어들의 평균 출현 정도를 연대별로 나타낸 도면이다.
- 도 7은 K -평균 군집화를 통해 구분된 8개의 군집을 4가지의 패턴으로 나눈 주제어의 목록을 나타낸 도면이다.
- 도 8은 본 발명의 일 실시예에 따른 대용량 학술논문정보 분석 방법의 동작과정을 상세하게 나타낸 순서도이다.
- 도 9는 도 8의 클러스터링 분석과정을 보다 상세하게 나타낸 순서도이다.

발명을 실시하기 위한 구체적인 내용

- [0038] 이하, 첨부된 도면을 참조하여 본 발명의 대용량 학술논문정보 분석 시스템을 상세하게 설명한다. 본 발명은 여러 가지 상이한 형태로 구현될 수 있으며, 여기에서 설명하는 실시예에 한정되지 않는다. 명세서 전체를 통하여 유사한 부분에 대해서는 동일한 도면 부호를 붙였다.
- [0039] 도 1은 본 발명의 일 실시예에 따른 대용량 학술논문정보 분석 시스템의 구성을 개략적으로 나타낸 도면이며, 도 2는 도 1의 분석 장치의 구성을 보다 상세하게 나타낸 도면이다.
- [0040] 도 1에 도시된 바와 같이 본 발명은, 분석 장치(100) 및 데이터베이스(200)를 포함한다. 이때 분석 장치(100)와

데이터베이스(200)는 유/무선 인터넷 등의 네트워크를 통해 통신 접속되어 있다.

- [0041] 분석 장치(100)는 특정 기술 부분에서 수십 년간 발행된 대용량의 학술논문정보를 수집, 분석하고, 분석 결과를 그래프 등의 형태로 도식화함으로써, 사용자가 과거에서부터 현재까지의 특정 기술 부분의 연구동향 및 추이를 손쉽게 확인할 수 있도록 하는 기능을 수행한다.
- [0042] 즉 상기 분석 장치(100)는 네트워크를 통해 하나 이상의 데이터베이스(200)로부터 특정 기술 부분의 학술논문정보를 수집하고, 수집한 특정 기술 부분의 학술논문정보의 제목 및 초록을 포함한 서지정보로부터 주제어를 도출한 후, 도출한 각각의 주제어에 대한 연대별 출현 빈도를 클러스터링 분석하여 소정의 그룹으로 군집화한다. 그리고 군집화한 소정의 그룹으로부터 몇 가지 패턴을 구분하여 결과를 산출하고, 산출 결과를 그래픽 형태로 표시해 줌으로써, 관련 종사자들이 산출 결과를 토대로 특정 기술 부분의 학술논문정보에 대한 현재까지의 연구동향이나 추이를 확인할 수 있도록 한다.
- [0043] 데이터베이스(200)는 특정 기술 부분에 관한 각종 학술논문정보를 누적, 저장하고 있으며, 분석 장치(100)로 지금까지 누적, 저장된 학술논문정보의 빅데이터를 제공한다.
- [0044] 상기 분석 장치(100)는 도 2에 도시된 바와 같이, 데이터 수집부(110), 주제어 추출부(120), 데이터 확인부(130), 데이터 분석부(140), 결과 산출부(150), 저장부(160), 제어부(170) 등으로 구성된다.
- [0045] 데이터 수집부(110)는 특정 기술 부분의 학술논문정보를 관리하는 데이터베이스(200)로부터 분석 대상이 되는 모든 학술논문정보(예를 들어, 제목, 초록 등의 서지정보)를 수집하는 기능을 수행한다.
- [0046] 주제어 추출부(120)는 데이터 수집부(110)를 통해 수집한 특정 기술부분의 학술논문정보의 서지정보로부터 주제어를 추출하는 기능을 수행한다.
- [0047] 이때 상기 주제어 추출부(120)는 저장부(160)에 기 저장되어 있는 용어 테이블을 참조하여 상기 학술논문정보의 서지정보에 나타나는 용어가 기 설정되어 있는 임계값 이상이면 해당 용어를 주제어로 추출한다.
- [0048] 데이터 확인부(130)는 데이터 수집부(110)를 통해 수집한 분석 대상의 모든 학술논문정보를 기 설정된 연대별(예를 들어, 년간, 월간 등의 단위로서, 본 발명에서는 10년 단위를 예로 설정하였음)로 구분하고, 연대별로 구분한 학술논문정보로부터 각각의 주제어에 대한 연대별 출현 빈도를 확인하여 분석을 위한 준비를 수행한다.
- [0049] 이때 상기 데이터 확인부(130)는 상기 학술논문정보를 기 설정된 연대별로 구분할 때, 각각의 연대별 학술논문정보의 수가 기 설정되어 있는 임계값 이하이면 해당 연대 및 해당 연대에 속하는 학술논문정보를 분석 대상에서 제외할 수 있다. 즉 유효한 분석 결과를 도출하기 어렵다고 판단하는 최소의 수량을 사전에 설정해 놓은 상태에서, 해당 부분의 로 데이터가 기 설정한 수량 이하인 경우 이를 분석대상에서 제외하여 분석 결과의 정확도를 높일 수 있도록 하는 것이다.
- [0050] 또한, 상기 데이터 확인부(130)는 분석 대상의 학술논문정보를 연대별로 구분하는 방식을 사용하였으나 이에 한정되는 것은 아니며, 그 이외에 저자별, 저널별 등의 다른 기준을 적용하여 구분할 수 있음은 물론이다.
- [0051] 데이터 분석부(140)는 상기 데이터 확인부(130)에서 확인한 각각의 주제어에 대한 연대별 출현 빈도의 클러스터링 분석을 수행하여 적어도 하나 이상의 그룹으로 군집화하는 기능을 수행한다.
- [0052] 이때 데이터 분석부(140)에서 수행하는 클러스터링 분석은, n개의 주제어에 대한 연대별 출현 빈도를 k개의 그룹으로 군집화하는 K-평균 군집화 기법을 사용하는 것이 가장 바람직하며, 그 이외에 공지된 모든 형태의 클러스터링 기법을 적용할 수 있다.
- [0053] 예를 들어, 클러스터링 분석에 K-평균 군집화 기법을 적용하는 경우, 본 발명에서는 다음의 순서로 진행하게 된다.
- [0054] 첫 번째, k개의 초기 점을 임의로 선택한 후 각각의 주제어를 상기 주제어와 가장 근접한 초기 점에 배정하여 k개의 군집을 생성한다. 즉 k개의 초기 점을 임의로 선택하고, 각각의 주제어와 k개의 초기 점과의 거리를 계산하며, 거리 계산 결과를 토대로 각각의 주제어를 가장 가까운 초기 점에 배정하여 k개의 군집을 생성하는 것이다.
- [0055] 두 번째, 생성된 k개의 군집에 속하는 각각의 주제어의 평균점을 계산한 후 상기 k개의 초기 점을 새롭게 계산된 평균점으로 조정한다.
- [0056] 세 번째, k개의 군집에 속하는 각각의 주제어의 평균점이 더 이상 조정되지 않을 때까지 상기 평균점 조정을 받

복하여 수행한다.

- [0057] 한편, 상기 데이터 분석부(140)는 클러스터링 분석을 수행하기 이전에 각각의 주제어에 대한 연대별 출현 빈도의 정규화를 수행한다. 즉 연대별로 구분한 학술논문정보의 수가 전부 다르므로 인해 발생할 수 있는 오차를 없애고 정확한 결과 값을 얻기 위하여, 클러스터링 분석을 수행하기 이전에 상기 각각의 주제어에 대한 연대별 출현 빈도의 정규화를 수행하는 것이다.
- [0058] 결과 산출부(150)는 데이터 분석부(140)에서의 클러스터링 분석을 통해 군집화한 적어도 하나 이상의 그룹을 특정 기술부분의 연구현황 및 추이와 관련된 적어도 하나 이상의 패턴으로 구분하여 결과를 도출한다.
- [0059] 이때 본 발명에서는 상기 연구현황 및 추이와 관련된 패턴으로, 최근 증가하고 있는 군집(즉 최근까지 지속적으로 상승되고 있는 주제어), 마지막 연대에 하강하고 있는 군집(즉 지속적으로 상승하고 있으나 최근에 하강하고 있는 주제어), 과거 특정 연대에서 급증한 군집(즉 한때 상승했었던 주제어), 꾸준히 하강하고 있는 군집(즉 지속적으로 하강하고 있는 주제어)의 4가지 패턴을 적용하였다. 물론, 상기 연구현황 및 추이와 관련된 패턴은 상술한 4가지 패턴 이외에 다른 형태의 패턴을 적용할 수 있다.
- [0060] 저장부(160)는 분석 장치(100)에서 사용하는 클러스터링 알고리즘을 포함한 각종 동작프로그램, 주제어 추출에 사용되는 용어 테이블을 저장하고 있다. 그리고 주제어 추출부(130)에서 추출한 주제어 정보, 데이터 확인부(130)에서 처리한 각각의 주제어에 대한 연대별 출현 빈도 정보, 데이터 분석부(140)에서 수행한 클러스터링 분석 정보, 결과 산출부(150)에서 처리한 결과 정보 등을 저장하고 있다.
- [0061] 제어부(170)는 분석 장치(100)의 동작을 총괄적으로 제어하는 기능을 수행하는 부분으로서, 데이터 수집부(110)의 분석 대상이 되는 학술논문정보 수집 제어, 주제어 추출부(120)의 주제어 추출 제어, 데이터 확인부(130)의 각각의 주제어에 대한 연대별 출현 빈도 처리 제어, 데이터 분석부(140)의 클러스터링 분석 제어, 결과 산출부(150)의 결과 산출 제어를 수행한다.
- [0062] 다음에는, 본 발명의 대용량 학술논문정보 분석 시스템을 이용한 컴퓨터공학 기술 부분의 학술논문정보 분석과정에 대한 구체적인 실시예를 도 3 내지 도 7을 참조하여 상세하게 설명한다.
- [0063] 도 3 내지 도 7은 발명의 분석 시스템을 이용한 컴퓨터공학 기술 부분의 학술논문정보 분석과정의 구체적인 실시예를 나타낸 도면이다.
- [0064] 먼저, 본 발명에서는 컴퓨터공학 기술 분야에 대한 학술논문의 연구경향 및 추이를 확인하기 위하여, 컴퓨터공학 기술 분야의 대표적인 데이터베이스 중 하나인 DBLP(<http://dblp.uni-trier.de>)로부터 약 3백만 건의 학술논문정보를 수집하여 분석을 수행하였다.
- [0065] 본 발명에서는 1935년부터 2015년까지 발행된 3백만 건의 학술논문정보의 제목 및 요약의 서지정보로부터 주제어를 추출하였다.
- [0066] 본 발명에서 추출한 주제어는 도 3에 도시된 바와 같이, 3천여 개의 IT 용어 사전(즉 2896개의 Gartner IT Term)을 기준으로 하여 컴퓨터공학 기술 분야의 모든 학술논문정보에 기록된 주제어 중에서 불용어, 범용어를 제외한 1397개의 주제어를 추출하였다. 예를 들어, 추출된 주제어는 도 4에 도시된 바와 같이, network, internet, web, openstack, nfv, deep learning, protocol, framework, tcp, bluetooth, programming language, synchronization, terminal, data center, mobile ip, database design, social network analysis, ftp, bandwidth, buffer, streaming, operating system, clustering, middle ware, html5, web crawler, microprocessor, artificial intelligence, ipv6, c++ 등이 있다.
- [0067] 그리고 본 발명에서는 DBLP 데이터베이스로부터 수집한 약 3백만 건의 학술논문정보를 1935-1945, 1946-1955, 1956-1965(D1960), 1966-1975(D1970), 1976-1985(D1980), 1986-1995(D1990), 1996-2005(D2000), 2006-2015(D2010)의 8개 연대로 나누었다. 하지만, 1935-1945 및 1946-1955의 2개의 연대는 도 5에 도시된 바와 같이 다른 연대에 발행된 학술논문에 비하여 훨씬 적은 113건 및 398건의 학술논문이 발행되었기 때문에 본 발명에서는 1935-1945 및 1946-1955의 2개의 연도는 삭제하고 6개의 연대를 기준으로 분석하였다.
- [0068] 예를 들어, 도 5의 각 주제어 중 'network'의 경우 D1960에서 69번, D1970에서 337번 등 1960년대부터 2010년대까지 꾸준히 관찰됨을 알 수 있다. 하지만, 'openstack'의 경우 2010년대에만 160번 관찰되는 것을 확인할 수 있다.
- [0069] 이처럼 도 5에 나타난 각 주제어의 연도별 출현 빈도는 K-평균 군집화 기법을 통해 분석된다.

- [0070] 본 발명에서 사용되는 K-평균 군집화는 n개의 주제를 k개의 그룹으로 군집화하는 것으로서, K-평균 군집화 방법은 다음과 같다.
- [0071] 먼저, k개의 초기 점을 임의로 선택한 후 각각의 주제를 상기 주제와 가장 근접한 초기 점에 배정하여 k개의 군집을 생성한다.
- [0072] 그리고 생성된 k개의 군집에 속하는 각각의 주제의 평균점을 계산한 후 상기 k개의 초기 점을 새롭게 계산된 평균점으로 조정한다.
- [0073] 그리고 k개의 군집에 속하는 각각의 주제의 평균점이 더 이상 조정되지 않을 때까지 상기 평균점 조정을 반복하여 수행한다.
- [0074] K-평균 군집화의 결과는 클러스터(k) 또는 거리 항목의 수에 따라 달라진다. 본 발명에서는 주제어 w1과 w2 사이의 거리 d는 다음과 같이 정의된다.

$$d = \sum_{i=1}^6 |f_i^1 - f_i^2|$$

- [0075] 여기서, f_i^1 및 f_i^2 는 i번째 연대에서 주제어 w1, w2의 출현 빈도이다.
- [0076]

- [0077] 한편, 군집 수를 결정 하는 방법들은 기존의 몇 가지가 공지되어 있으나 어느 하나의 방식이 좋다고 말하기는 어려우며 보통 문제에 대한 배경지식을 기반으로 방법을 결정한다. 거리계산방식은 유클리드(Euclidean), 맨하튼(Manhattan), 상관관계(Correlation) 등 다양한 방식이 있으나 데이터의 특성과 분석 목적에 맞게 결정을 하며, 본 발명에서는 상관관계를 이용하였다.

- [0078] 상기 K-평균 군집화 기법을 통해 컴퓨터공학 기술 분야의 각 학술논문정보에 대한 각 주제의 연도별 출현 빈도의 분석 결과, 8개의 군집으로 분석했을 때 가장 의미 있는 결과를 얻을 수 있었으며, 연대별 논문의 개수가 모두 다르기 때문에 정확한 결과를 얻기 위해 연대별로 주제어 출현 빈도에 대한 정규화를 K-평균 군집화에 앞서 수행하였다.

- [0079] 도 6은 K-평균 군집화의 결과를 이용하여 각 군집에 해당하는 주제어들의 평균 출현 정도를 연대별로 나타낸 도면으로서, 8개의 군집은 크게 최근 증가하고 있는 군집, 마지막 연대에 하강하고 있는 군집, 과거 특정 연대에 서 급증한 군집, 꾸준히 하강하고 있는 군집의 4가지 패턴으로 설명할 수 있다.

- [0080] 도 7은 K-평균 군집화를 통해 구분된 8개의 군집을 4가지의 패턴으로 나눈 주제어의 목록을 나타낸 도면으로서, 그룹 1에 속한 주제어들은 빈도수가 꾸준히 증가 하다가 최근에는 급격히 증가하는 것을 볼 수 있으며, 그룹 2에 속해 있는 주제어들은 2000년대 들어 빈도수가 점차 감소하고 있음을 보여주고 있다.

- [0081] 또한, 그룹 3에 속한 주제어들은 과거 1970년대 및 1980년대에 빈도수가 가장 높았다가 점차 감소하고 있는 추세를 보이고 있으며, 마지막으로 그룹 4에 속한 주제어들은 꾸준히 빈도수가 하강 하고 있는 것을 보여주고 있다. 논문의 초록이 연구의 요약정보를 담고 있다는 점을 감안했을 때, 도 7은 연구자들에게 꾸준한 사랑을 받아 온 주제어들, 또는 서서히 관심을 잃어가거나 최근 뜨겁게 탐구되는 연구 분야를 보여주고 있다.

- [0082] 이처럼 본 발명에서는 컴퓨터공학 기술 분야의 대용량 학술논문정보로부터 각 주제를 기반으로 연대별 출현 빈도를 클러스터링 분석함으로써, 과거에서부터 현재에 이르는 연구경향 및 추이를 확인할 수 있다. 즉 시간에 따른 주제어 군집 결과는 DBLP의 연구동향을 알아볼 수 있는 매우 흥미로운 자료로 사용될 수 있을 것이다.

- [0083] 다음에는, 이와 같이 구성된 본 발명에 따른 대용량 학술논문정보 분석 방법의 일 실시예를 도 8과 도 9를 참조하여 상세하게 설명한다. 이때 본 발명의 방법에 따른 각 단계는 사용 환경이나 당업자에 의해 순서가 변경될 수 있다.

- [0084] 도 8은 본 발명의 일 실시예에 따른 대용량 학술논문정보 분석 방법의 동작과정을 상세하게 나타낸 순서도이며, 도 9는 도 8의 클러스터링 분석과정을 보다 상세하게 나타낸 순서도이다.

- [0085] 우선, 분석 장치(100)는 네트워크를 통해 특정 기술 부분의 학술논문정보를 관리하는 데이터베이스(200)로부터 분석 대상이 되는 모든 학술논문정보를 수집한다(S100).

- [0086] S100 단계를 통해 분석 대상의 학술논문정보를 수집한 이후, 분석 장치(100)는 학술논문정보의 서지정보를 통해

주제어를 추출한다(S200).

- [0087] 상기 S200 단계를 통해 주제어를 추출할 때, 분석 장치(100)는 저장부(160)에 기 저장되어 있는 용어 테이블을 참조하여 상기 학술논문정보의 서지정보에 나타나는 용어가 기 설정되어 있는 임계값 이상일 경우 해당 용어를 주제어로 추출한다. 이때 임계값은 통상적으로 1 이상으로 설정하는 것이 바람직하다.

[0088] 상기 S200 단계를 통해 주제어를 추출한 이후, 분석 장치(100)는 상기 학술논문정보를 기 설정된 연대별로 구분하고, 연대별로 구분한 학술논문정보로부터 S200 단계에서 추출한 각각의 주제어에 대한 연대별 출현 빈도를 확인한다(S300). 이때 분석 장치(100)는 상기 학술논문정보를 기 설정된 연대별로 구분할 때, 연대별 학술논문정보의 수가 기 설정되어 있는 임계값 이하이면 해당 연대 및 해당 연대에 속하는 학술논문정보를 분석 대상에서 제외한다. 즉 유효한 분석 결과를 도출하기 어렵다고 판단되는 연대에 속하는 학술논문정보를 분석대상에서 제외하여 분석 결과의 정확도를 높이도록 한 것이다.

[0089] 그리고 분석 장치(200)는 S300 단계에서 확인한 상기 각각의 주제어에 대한 연대별 출현 빈도의 클러스터링 분석을 수행하여 적어도 하나 이상의 그룹으로 군집화한다(S400). 즉 S300 단계에서 확인한 n개의 주제어에 대한 연대별 출현 빈도를 k개의 그룹으로 군집화하기 위한 클러스터링 분석을 수행하는 것으로서, 본 발명에서는 K-평균 군집화 기법을 사용한다. 이때 분석 장치(100)는 클러스터링 분석을 수행하기 이전에, 연대별로 구분한 학술논문정보의 수가 다름으로 인해 발생될 수 있는 분석 오차를 줄이기 위해 각각의 주제어에 대한 연대별 출현 빈도의 정규화를 수행한다.

[0090] 상기 S400 단계의 K-평균 군집화를 도 9를 참조하여 보다 상세하게 설명하면, 우선 k개의 초기 점을 임의로 선택한 후 각각의 주제어를 상기 주제어와 가장 근접한 초기 점에 배정하여 k개의 군집을 생성하는 군집 생성 단계를 수행한다(S410).

[0091] 상기 S410 단계를 통해 생성된 k개의 군집에 속하는 각각의 주제어의 평균점을 계산한 후 상기 k개의 초기 점을 새롭게 계산된 평균점으로 조정하는 평균점 조정 단계를 수행한다(S420).

[0092] 그리고 k개의 군집에 속하는 각각의 주제어의 평균점이 더 이상 조정되지 않을 때까지 평균점 조정을 반복하여 수행한다(S430).

[0093] 이처럼, S400 단계를 통해 n개의 주제어에 대한 연대별 출현 빈도를 k개의 그룹으로 군집화한 이후, 분석 장치(100)는 군집화한 적어도 하나 이상의 그룹을 특정 기술부분의 연구현황 및 추이와 관련된 적어도 하나 이상의 패턴으로 구분한 후(S500), 결과 데이터를 텍스트 및 그래픽이 포함된 형태로 제공(예를 들어, 화면상에 표시하거나 파일로 제공할 수 있음)하여 관련 종사자가 확인할 수 있도록 한다(S600).

[0094] S500 단계를 통해 확인할 수 있는 연구현황 및 추이와 관련된 패턴은, 최근 증가하고 있는 군집, 마지막 연대에 하강하고 있는 군집, 과거 특정 연대에서 급증한 군집, 꾸준히 하강하고 있는 군집의 4가지 패턴을 기본적으로 사용한다. 하지만, 연구현황 및 추이와 관련된 패턴은 상술한 4가지 패턴에 한정할 필요가 없으며 다양한 형태로 변환하여 적용할 수 있다.

[0095] 이처럼, 본 발명은 수십 년간 축적된 많은 수의 학술논문정보에서 도출한 주제어를 K-평균 군집화와 같은 클러스터링 기법을 통해 분석함으로써, 특정 기술 부분의 연대별 주제어 출현 패턴에 따른 연구현황 및 추이를 손쉽게 확인할 수 있다.

[0096] 또한, 비정형화된 데이터인 학술논문정보를 활용한 기술현황 및 추이와 관련된 데이터 분석을 통해 웹 마이닝 및 텍스트 마이닝 분야의 연구를 활성화시킬 수 있다.

[0097] 또한, 본 발명에서 제시한 특정 기술 부분의 학술논문정보와 관련된 연구현황 및 추이 이외에도, 지금까지 축적된 대량의 신문기사, 연애기사, 온라인 쇼핑정보 등의 다양한 분야에 폭넓게 적용할 수 있다.

[0098] 여기에서, 상술한 본 발명에서는 바람직한 실시예를 참조하여 설명하였지만, 해당 기술분야의 숙련된 당업자는 하기의 특허청구범위에 기재된 본 발명의 사상 및 영역으로부터 벗어나지 않는 범위 내에서 본 발명을 다양하게 수정 및 변경할 수 있음을 이해할 수 있을 것이다.

부호의 설명

- [0099] 100 : 분석 장치 110 : 데이터 수집부
 120 : 제어어 추출부 130 : 데이터 확인부

- 140 : 데이터 분석부

150 : 결과 산출부

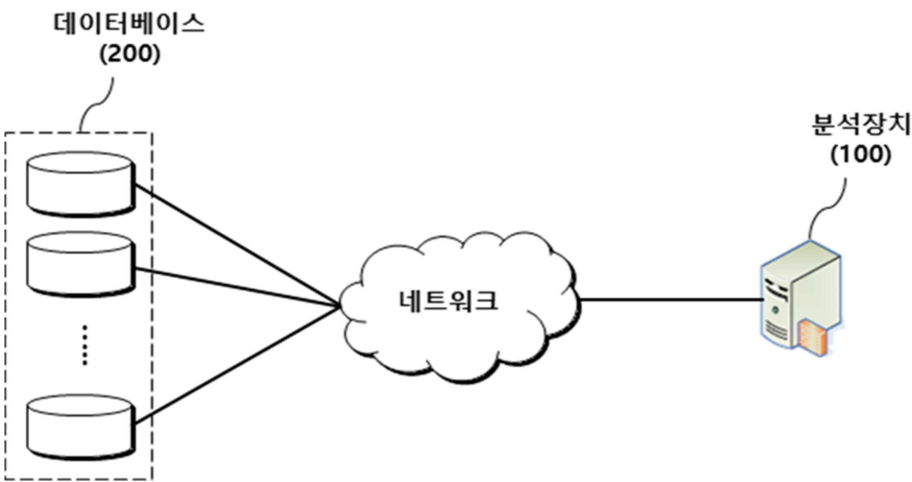
160 : 저장부

170 : 제어부

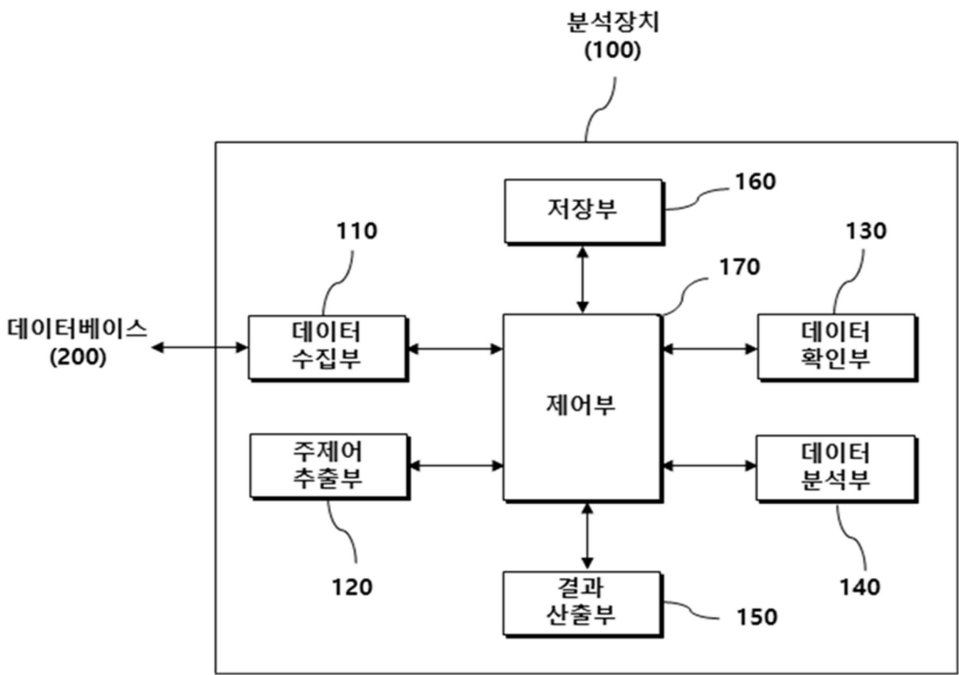
200 : 데이터베이스

도면

도면1



도면2



도면3

	IT 용어사전	추출 주제어
수량	2896	1397

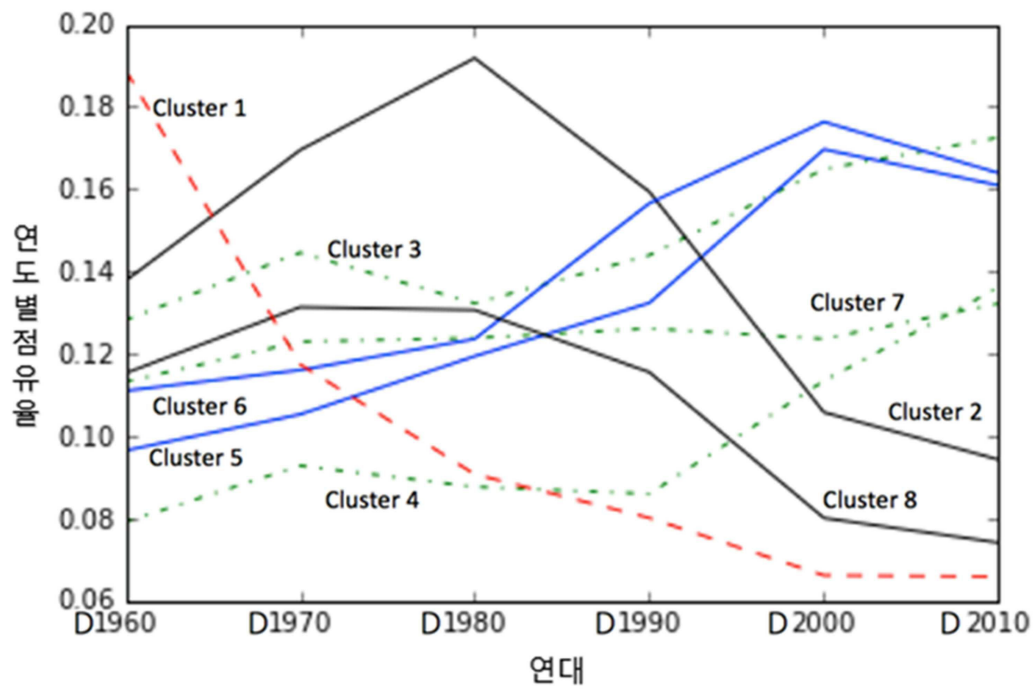
도면4

network, internet, web, openstack, nfv, deep learning, protocol, framework, tcp, bluetooth, programming language, synchronization, terminal, data center, mobile ip, database design, social network analysis, ftp, bandwidth, buffer, streaming, operating system, clustering, middle ware, html5, web crawler, microprocessor, artificial intelligence, ipv6, c++

도면5

구간	D1960	D1970	D1980	D1990	D2000	D2010
	1956~1965	1966~1975	1976~1985	1986~1995	1996~2005	2006~2015
논문 개수	5096	22827	66566	248179	795146	2113389
network	69	337	1763	9848	42465	151642
openstack	0	0	0	0	0	160
...	...	...	...	...	...	...
digital	199	481	1219	3243	15604	35176
data center	0	2	16	26	132	4043

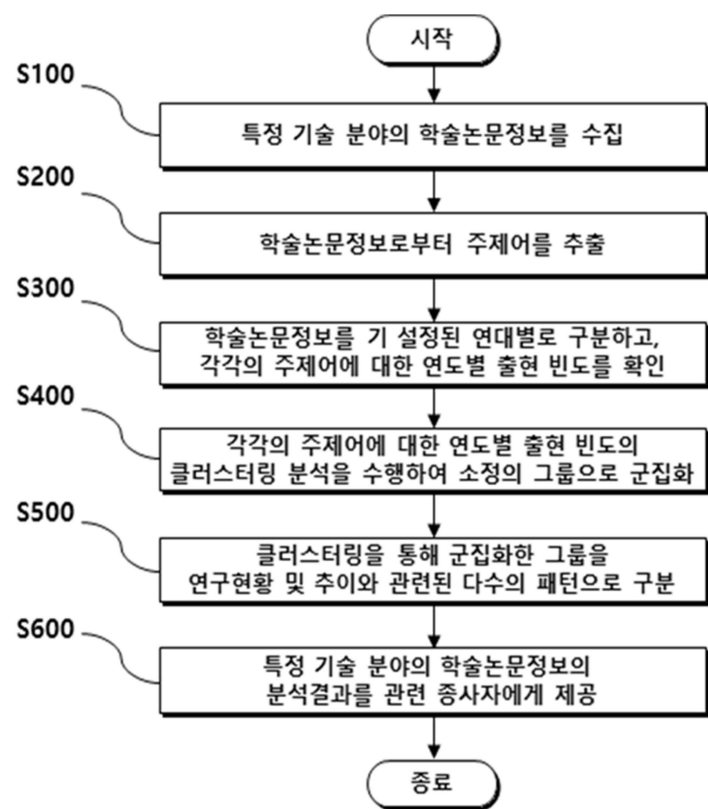
도면6



도면7

그룹	k-means clusters	주제어	비고
그룹1	Cluster 3 Cluster 4 Cluster 7	access point, active directory, adaptive learning, affective computing, analytic applications, analytics, application architecture, authentication technologies, autonomous system, autonomous vehicles, backbone, backbone network, bandwidth, base station, batch processing, benchmarking, bioinformatics, business analytics, content management, concurrent engineering, cognitive radio, clustering, cloud computing, circuit switching, checksum, code division multiple access, design thinking, deep learning, data scientist, data center, distributed data management, dynamic routing, eye tracking, femtocells, fraud detection, grid computing, html5, internet protocol, ipv6, job scheduling, media access control, mobile ip, mobile network, mobile social networks, multithreading, network intelligence, network management, network virtualization, openstack, p2p, protocol stack, proxy servers, saas, sdh/sonnet, session initiation protocol, social analytics, social network analysis, software defined networking, switched network, tcp/ip, text analytics, virtual reality, virtualization, vlan, vpn, web crawler, wireless broadband	최근 증가하고 있는 군집
그룹2	Cluster 5 Cluster 6	architecture, availability, broadcast, buffer, cache, clock, community, data mining, encryption, framework, fuzzy logic, information technology, java, load balancing, machine learning, mobile, protocol, qos, scalability, signature, streaming, tcp, throughput, visualization, web, web services, xml	마지막 연대에 하강하고 있는 군집
그룹3	Cluster 2 Cluster 8	artificial intelligence, attenuation, broadband, c++, channel capacity, circuit board, composition, cycle time, database design, database management system, distributed computing, distributed database, expert system, hypertext, information management, knowledge representation, microprocessor, operation system, packet switching, parallel processing, response time, software development, speech recognition, synchronization, terminal, transducer, workstations	과거 특정 연대에서 급증한 군집
그룹4	Cluster 1	binary code, modulation, redundancy, switch, control, application, information, processing, simulation	꾸준히 하강하고 있는 군집

도면8



도면9

