

(19) 대한민국특허청(KR)
(12) 공개특허공보(A)(11) 공개번호 10-2023-0109446
(43) 공개일자 2023년07월20일

(51) 국제특허분류(Int. Cl.)

G06F 16/34 (2019.01) G06F 16/26 (2019.01)
G06F 16/33 (2019.01) G06F 40/284 (2020.01)
G06N 20/00 (2019.01)

(52) CPC특허분류

G06F 16/34 (2019.01)
G06F 16/26 (2019.01)

(21) 출원번호 10-2022-0005433

(22) 출원일자 2022년01월13일

심사청구일자 없음

(71) 출원인

고려대학교 산학협력단

서울특별시 성북구 안암로 145, 고려대학교 (안암동5가)

(72) 발명자

장동식

서울특별시 성북구 안암로9길 30, 2동 1102호

김영호

서울특별시 성북구 동소문로 211, A동 602호

(뒷면에 계속)

(74) 대리인

김준석, 박민욱

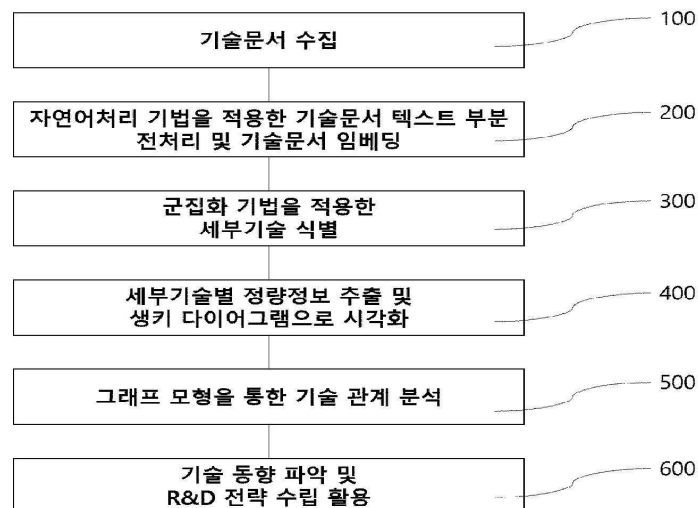
전체 청구항 수 : 총 4 항

(54) 발명의 명칭 기술문서 정보 시각화 및 기술 관계 분석 시스템

(57) 요약

특허 빅데이터 분석결과 전달을 위해 다양한 시각화 기법이 사용된다. 기존에 특허 빅데이터 시각화 기법들은 데이터 간의 상대적 비교에 유용하다. 그러나 비교 대상 데이터의 증가 또는 기술 흐름 파악에는 한계점이 존재한다 이를 보완하기 위해 일부 연구들은 생키 다이어그램을 사용하였으나, 세부기술의 식별과 기술 관계를 고려하지 못한다. 본 발명에서는 특허 빅데이터에 생키 다이어그램과 그래프 모형을 사용하여 세부기술을 식별, 기술개발 흐름 및 기술 관계를 시각화한다. 실험 데이터는 헬스케어 기술 관련 특허로 한국, 미국, 일본, 유럽에서 출원된 것이다. 실험결과로는 운동 부하 모니터링에서 스마트폰 연동 생체계측 기술로 개발이 진행되고 있다. 또한, 통신 관련 기술의 영향력이 큰 것으로 도출되었다.

대표도 - 도1



(52) CPC특허분류

G06F 16/3347 (2019.01)

G06F 40/284 (2020.01)

G06N 20/00 (2021.08)

(72) 발명자

강지호

서울특별시 용산구 효창원로37가길 7, 203호

이주현

서울특별시 성북구 개운사2길 13-39, 205호

이 발명을 지원한 국가연구개발사업

과제고유번호 1415177481

과제번호 P0017311

부처명 산업통상자원부

과제관리(전문)기관명 한국산업기술진흥원

연구사업명 산업혁신인재성장지원(R&D)

연구과제명 로봇산업 혁신을 위한 글로벌인재 성장지원 프로젝트

기 여 율 100/100

과제수행기관명 고려대학교 산학협력단

연구기간 2021.03.01 ~ 2021.12.31

명세서

청구범위

청구항 1

기술문서 정보 시각화 및 기술 관계 분석 시스템을 실시하는데 있어서,

분석 대상 기술분야 관련 기술문서를 수집하는 단계;

상기 기술문서를 수집하고, 수집한 기술문서 집합으로부터 텍스트 부분을 추출하여 자연어처리(Natural language processing) 기법 적용을 통해 불용어(Stopwords), 구두점, 숫자 제거 및 단어를 토큰화(Tokenization)하는 전처리 단계;

상기 전처리된 텍스트에 Doc2Vec을 적용하여 기술문서를 특정 벡터공간(Vector space)에 임베딩(Embedding)하고 임베딩 행렬(Embedding matrix)을 구축하는 단계;

상기 구성된 임베딩 행렬에 군집화(Clustering) 기법을 적용하여 세부기술(Elementary technology) 군집을 형성하고 그래프 모형(Graph model)을 적용하여 세부기술을 식별하는 단계;

상기 식별된 세부기술별로 정량정보를 추출하고 생키 다이어그램(Sankey diagram)으로 기술개발 흐름을 시각화하는 단계;

상기 세부기술 식별에 사용한 단어 목록으로 기술문서 집합에서 전체 동시발생 행렬(Co-occurrence matrix)을 구축 및 그래프 모형을 적용하여 기술 관계를 분석하는 단계;

를 포함하는 것을 특징으로 하는 기술문서 정보 시각화 및 기술 관계 분석 시스템.

청구항 2

제 1항에 있어서,

기술문서 집합으로부터 세부기술을 식별하기 위해 사용된 방법은,

텍스트 부분을 추출하고 자연어처리 기법을 적용하여 전처리, Doc2Vec을 통한 기술 문서 임베딩 및 임베딩 행렬 구축, 임베딩 행렬에 군집화 기법을 적용하여 다수개의 세부기술 군집을 형성, 군집별 단어 동시발생 행렬 구축, 군집별 단어 동시발생 행렬에 그래프 모형을 적용하여 상위 값을 가지는 중심단어를 도출하고, 이를 통해 세부기술을 정의하는 것을 특징으로 하는 기술문서 정보 시각화 및 기술 관계 분석 시스템.

청구항 3

제 1항에 있어서,

기술개발 흐름을 시각화하기 위해 사용된 방법은,

세부기술별로 기술문서 건수를 확인 후, 기술문서의 정량정보(ex. 출원인, 출원연도, 출원국가 등)를 추출하는 단계를 거쳐 생키 다이어그램으로 기술개발 흐름을 시각화하는 것을 특징으로 하는 기술문서 정보 시각화 및 기술 관계 분석 시스템.

청구항 4

제 1항에 있어서,

기술 관계 식별을 위한 그래프 모형 적용은,

상기 세부기술 식별을 위해 활용한 상위 값을 가지는 중심단어 목록을 활용하여 전체 기술문서 집합으로부터 단어 동시발생 행렬을 구축하고, 구축된 동시발생 행렬에 그래프 모형을 적용하여 기술 관계 그래프를 구축 및 시각화한 후, 기술 관계를 정량적으로 식별하는 것을 특징으로 하는 기술문서 정보 시각화 및 기술 관계 분석 시스템.

발명의 설명

기술 분야

[0001] 본 발명은 기계학습, 생키 다이어그램(Sankey diagram)과 그래프 모형(Graph model) 기법을 활용하여 기술문서 집합으로부터 추출한 정보를 시각화하고 기술 관계를 정량적으로 분석하는 시스템에 관한 것으로, 자연어처리 기법을 적용하여 기술문서 집합으로부터 세부기술을 식별하고 정량정보와 함께 생키 다이어그램으로 시각화하여 기술개발 흐름을 파악함과 동시에 기술문서에 등장한 단어들에 그래프 모형을 적용하여 기술 간의 관계를 분석하여 기업 및 연구기관에서 기술개발 흐름 파악, R&D 전략을 효과적으로 수립할 수 있도록 하는 기술문서 정보 시각화 및 기술 관계 분석 시스템에 관한 것이다.

배경 기술

[0002] 특허는 개발된 기술에 대한 독점적인 권리를 부여한다. 따라서 특허에는 개발된 기술의 상세한 정보가 다양한 형태로 포함된다. 최근 많은 기업과 기관에서 특허의 중요성을 인식하고 특허출원에 박차를 가하고 있다. 이러한 이유로 글로벌 특허출원 건수는 매년 증가하는 추세이며, 특허를 빅데이터로 인식하고 다양한 기법을 적용하는 연구가 활발히 진행되고 있다.

[0003] 특허 빅데이터는 기업·기관의 R&D, 시장 동향 등의 정보가 포함되어 있다. 그러므로 특허 빅데이터를 분석하여 다양한 가치를 창출하고 경영정보로써 활용하는 사례가 점차적으로 증가하고 있다.

[0004] 특허 빅데이터 분석은 크게 정성적 방법과 정량적 방법으로 구분되나 분석의 효율성 및 결과의 재현 가능성 제고를 위해 정량적 분석을 주로 수행한다. 정량적 방법은 통계 및 기계학습 알고리즘을 사용하여 결과를 도출하고, 이를 효과적으로 전달하기 위한 다양한 시각화 기법을 포함한다.

[0005] 최근 특허를 포함한 기술문서가 빅데이터화됨에 따라, 이를 정량적으로 분석하고 시각화하여 효과적으로 정보를 식별 및 추출하는 기법이 필요하다. 기술문서는 텍스트, 숫자, 그림 등 다양한 형태로 정보를 포함하고 있으므로, 이를 적절한 기법을 적용하여 전처리하고 활용 목적에 맞게 시각화하는 것이 중요하다.

[0006] 기존에 특허 빅데이터 시각화 기법들은 주요 시장국, 출원인, 기술 관계 등 단일 목적을 표현하는 것에 초점이 맞추어져 있다. 이와 같은 시각화 기법들은 데이터 간의 상대적 비교를 통한 결과 식별에 유용하나, 비교 대상 데이터의 증가 또는 기술 흐름 파악에서는 한계점이 존재한다. 이러한 한계점을 보완하기 위해 일부 선행연구들에서는 데이터 흐름의 양을 비율적으로 표현하여 효과적으로 전달하는 생키 다이어그램(Sankey diagram)을 사용한다. 그러나 생키 다이어그램을 사용한 선행연구들에서는 특허 빅데이터에 적용한 사례가 적으며, 수집된 특허에서 세부기술 식별 및 기술 간의 관계를 고려하지 못한다. 특허 빅데이터에서 주요 시장국 및 출원인 R&D 동향을 효과적으로 파악하기 위해서는 대상 기술의 세부기술을 식별하고 개발 흐름을 시각화할 필요가 있다.

[0007] 특허 빅데이터를 분석하여 다양한 가치를 창출해내는 것과 더불어 사용자 또는 의사 결정자에게 효과적으로 분석결과를 전달하는 방법은 매우 중요하다. 이런 관점에서 특허 빅데이터 시각화는 단순히 데이터의 표현뿐만 아니라, 정보 전달 및 설득의 목적을 포함하고 있다.

[0008] 다양한 선행연구들이 특허 빅데이터 분석결과를 시각화하는 방법을 제안하였다. Chen “Design Patent Map Visualization Display” (2009)은 차량의 디자인 특허 분석을 위해 상위 출원인 및 주요 시장국을 막대 그래프로 표현하였다. 또한, Multi-dimensional scale(MDS)를 사용하여 2차원 공간에 특허를 매핑하였다. Moehrle and Passing “Applying an Anchor Based Patent Mapping Approach: Basic Conception and the Case of Carbon Fiber Reinforcements” (2016)은 특허의 융합패턴 분석을 위해 텍스트 정보를 RadViz diagram에 매핑하였다. Moehrle and Passing이 제안한 시각화 방법은 기술 간의 유사성 식별에 유용하다. Kim et al “An Integrated Social Network Mining for Product-based Technology analysis of Apple” (2017)은 Apple의 제품들에 기반한 특허들을 사회네트워크 분석(Social network analysis, SNA)을 사용하여 그래프로 시각화하였다. SNA는 기술 간의 관계 확인에 효과적이다. 이와 같은 선행연구들은 시각화한 데이터 간의 상대적 비교를 통해 효과적으로 정보를 전달할 수 있다. 그러나 데이터의 양이 증가하거나 거시적 관점에서 기술개발 흐름을 파악하는 것에는 적합하지 않다.

[0009] 생키 다이어그램은 선의 너비로 데이터 양을 비율적으로 표현하여 프로세스에서의 전체적인 흐름 파악에 유용하다. 이러한 이유로, K Soundararajan, H K Ho and B Su, “Sankey Diagram Framework for Energy and Exergy Flows” (2014) 등 생키 다이어그램은 주로 에너지 등의 흐름 파악을 위한 시각화 기법으로 주로 사용되었다.

특허 빅데이터에 생키다이어그램을 활용하여 거시적 관점에서 기술개발 흐름을 시각화할 수 있다. Spreafico et al. “A State-of-the-art Review of FMEA/FMECA Including patents,” (2017)은 기술개발 동향을 분석하기 위해 관련 논문과 특허를 수집하고 해결 목적에 따라 분류 및 생키 다이어그램으로 시각화하였다. 이들은 저자가 명시적으로 선언한 문제를 기준으로 기술 분야를 직접 분류하였다. 이와 같은 방법은 기술분류에 정확도를 높일 수 있으나, 데이터의 양이 증가할 경우 효율성이 저하된다. Baskici et al. “Forecasting of Innovation in the Light of Semantic Networks” (2019)는 건강 관련 논문을 수집하고 생키 다이어그램으로 주요 연구 분야를 식별하였다. 이들은 논문에 텍스트마이닝을 적용하고 Bigram을 통해 도출된 단어로써 연구 분야를 식별하였다. Baskici et al.의 연구 분야 식별 방법은 정량적 방법으로써 효율성이 높으나, 단어의 출현 빈도를 기반으로 하여 적은 출현 빈도를 가지면서 영향력이 높은 기술의 식별이 어렵다.

[0010] 분석 대상 기술과 관련된 검색식 작성을 통해 특허 빅데이터를 수집하나, 융·복합 기술의 특성으로 인해 데이터에는 다수개의 기술이 포함되어 있다. 이러한 이유로, 세부기술을 식별하고 분석을 수행할 필요가 있다. 세부 기술 식별은 주로 특허의 텍스트 정보를 추출하여 정형화하고 군집화 적용, 기술 정의 순으로 수행된다. Jun et al. “Technology Forecasting Using Matrix Map and Patent Clustering” (2012), Jun et al. “Document Clustering Method Using Dimension Reduction and Support Vector Clustering to Overcome Sparseness” (2014)는 특허의 텍스트를 문서-단어행렬(Document-Term matrix, DTM)로 구성하고 주성분분석(Principal component analysis, PCA)을 사용하여 차원을 축소하였다. 또한, K-Medoids와 K-Means 군집화를 각 사용하고 군집별 상위 출현 단어로 세부기술을 식별하였다. Suh et al. “Patent Clustering and Network Analyses to Explore Nuclear Waste Management Technologies” (2020)은 DTM에 특이값 분해(Singular value decomposition, SVD)를 적용하여 차원을 축소하고 문서 유사도를 통해 기술군집을 형성하였다. 상기의 연구들에서 사용한 DTM은 Bag-of-words 기반 방법으로 문맥 정보를 반영하지 못한다.

발명의 내용

해결하려는 과제

[0011] 본 발명에서는 특허 빅데이터를 활용하여 기술개발 흐름을 효과적으로 시각화하기 위해 생키 다이어그램을 활용한다. 생키 다이어그램을 적용할 때, 세부기술을 식별하기 위한 방법으로 Doc2Vec, 군집화, Textrank를 사용한다. 또한, 세부기술별로 도출된 주요 단어들로 상관행렬을 구축하고 그래프 모형으로 표현하여 기술 간의 관계를 시각화한다.

[0012] 자연어처리 기법은 개발된 기술 내용이 포함된 기술문의 텍스트 부분을 정량적으로 분석하기 위해 전처리를 수행할 수 있다. 또한, 전처리된 텍스트를 군집화를 통해 기술문서 집합으로부터 세부기술을 식별할 수 있다. 생키 다이어그램은 정보의 흐름을 효과적으로 시각화하는 기법으로써, 이와같이 식별된 세부기술별로 개발 흐름을 시각화 및 파악하는 것에 도움을 준다. 추가적으로, 그래프 모형을 사용한다면 기술 간의 관계를 정량적으로 시각화 및 분석할 수 있다.

과제의 해결 수단

[0013] 본 발명에 따른 기술문서 정보 시각화 및 기술 관계 분석 시스템은 기술동향 파악 및 R&D 전략을 효과적으로 수립할 수 있도록 기술문서를 수집하여 텍스트 부분을 추출하고 자연어처리 기법을 적용하여 전처리하는 수단, 전처리된 텍스트에 Doc2Vec을 적용하여 기술문서 임베딩 행렬을 구축하는 수단, 그리고 임베딩 행렬에 군집화 기법을 적용하여 세부기술 군집을 형성하는 수단, 세부기술 군집별 단어 동시발생 행렬을 구축하고 그래프 모형을 적용하여 상위 단어를 추출 및 세부기술을 정의하는 수단, 그리고 식별된 세부기술별 정량정보를 추출하고 생키 다이어그램으로 시각화하여 기술개발 흐름을 파악하는 수단, 그리고 세부기술별 추출한 상위 단어들을 기술문서 집합에서 전체 동시발생 행렬을 구축하고 그래프 모형을 적용하여 기술 간의 관계를 식별하는 수단, 그리고 그 결과들을 바탕으로 기술동향 파악 및 R&D 전략 수립에 활용하는 수단을 포함한다.

[0014] 본 발명에 따른 방법은, 기술동향 파악 및 R&D 전략 수립을 보다 용이하게 실시하는데 있어서, 분석 대상이 되는 기술분야의 기술문서를 수집하는 단계; 수집된 기술문서 집합으로부터 텍스트 정보를 추출하고 자연어처리 기법을 통해 전처리 및 Doc2Vec을 적용하여 임베딩 행렬을 구축하는 단계; 구축된 임베딩 행렬에 군집화 기법을 적용하여 세부기술 군집을 형성하고 군집별 단어 동시발생 행렬을 구축 및 그래프 모형을 적용하여 상위 단어를 추출하고 세부기술을 식별하는 단계; 기술문서 집합으로부터 세부기술별로 정량정보를 추출하고 생키 다이어그램으로 이를 시각화하는 단계; 군집별 추출한 상위 단어 목록을 활용하여 기술문서 집합에서 전체 동시발생 행

렬을 구축하고 그래프 모형을 적용하여 기술 관계를 시각화 및 분석하는 단계; 분석 결과물을 활용하여 기술동향 파악 및 R&D 전략 수립에 활용하는 단계를 포함한다.

발명의 효과

[0015] 빅데이터화된 기술문서 집합으로부터 세부기술을 식별하고 기술별 특징 및 흐름을 시각화하여, 인사이트를 도출하는 것에는 상당한 노하우와 시간 및 비용이 소요된다. 기술문서 정보 시각화 및 기술 관계 분석 시스템은 이러한 문제를 해결하여 기술개발 흐름, 기술 관계를 시각화하고 분석할 수 있다.

[0016] 본 발명의 기술문서 정보 시각화 및 기술 관계 분석 시스템에 의하면, 기술문서 집합으로부터 텍스트 정보와 정량정보를 추출하여 자연어처리 기법으로 전처리 및 문서를 특정 공간에 임베딩하고 군집화를 적용하여 세부기술을 식별하며, 세부기술과 정량정보를 생키 다이어그램으로 시각화하여 기술개발 흐름을 효과적으로 파악하고, 전처리된 텍스트에 그래프 모형을 적용함으로써, 기술 간의 영향 관계를 시각화 및 분석하여 그 결과들을 바탕으로 기업 및 연구기관에서 기술동향 파악 및 R&D 전략 수립에 활용할 수 있다.

도면의 간단한 설명

[0017] 도1은 본 발명의 기술문서 정보 시각화 및 기술 관계 분석 시스템의 실시예 흐름도를 나타낸 것이다.
 도2는 본 발명에서 자연어처리 기법을 적용한 기술문서 텍스트 부분 전처리 및 기술문서 임베딩의 흐름도를 나타낸 것이다.
 도3은 본 발명에서 군집화 기법을 적용한 세부기술 식별의 흐름도를 나타낸 것이다.
 도4는 본 발명에서 세부기술별 정량정보 추출 및 생키 다이어그램으로 시각화하는 흐름도를 나타낸 것이다.
 도5는 본 발명에서 그래프 모형을 통한 기술 관계 분석의 흐름도를 나타낸 것이다.
 도 6은 본 발명에서 단어별 Textrank 값을 산출하는 식을 나타낸 것이다.
 도 7은 본 발명에서 특허 빅데이터의 세부기술을 식별하고 기술개발 흐름을 시각화하는 방법을 나타낸 것이다.
 도 8은 본 발명에서 헬스케어 관련 특허 빅데이터의 상위 20개 출원인을 나타낸 것이다.
 도 9는 본 발명에서 최적의 K를 위한 Elbow method 결과를 나타낸 것이다.
 도 10은 본 발명에서 세부기술 식별 결과를 나타낸 것이다.
 도 11은 본 발명에서 헬스케어 기술의 생키다이어그램을 나타낸 것이다.
 도 12는 본 발명에서 군집별 상위 키워드를 활용하여 기술 간의 관계를 그래프로 나타낸 것이다.

발명을 실시하기 위한 구체적인 내용

[0018] 본 발명에서는 수집된 특허 빅데이터에 자연어 처리 및 그래프 모형을 사용하여 세부기술을 식별한다. 또한, 특허 정량정보와 결합하여 생키 다이어그램으로 시각화한다. 마지막으로, 세부기술에서 도출된 주요 키워드로 상관행렬(Correlation matrix)을 구축하고 기술 간의 관계 식별을 위한 그래프 모형을 도출한다. 실험 데이터로는 헬스케어 기술과 관련된 것으로써 한국(KR), 미국(US), 일본(JP), 유럽(EP)에서 출원된 특허를 수집하여 사용한다.

[0019] 또한, 본 발명에서는 문맥 정보를 보존하면서 효과적으로 문서를 임베딩할 수 있는 Doc2Vec을 사용한다. 또한, 임베딩된 특허에 K-Means 군집화를 사용하여 기술군집을 형성한다. 기술군집별로 Textrank를 사용하여 빈도 기반이 아닌, 영향력이 높은 순으로 단어를 도출하고 세부기술을 식별한다. Textrank는 그래프 기반 모형으로 단어 집합 내에서 영향력이 높은 키워드를 식별하는 것에 효과적이다. 단어별 Textrank 값은 아래 도 5의 식과 같이 산출된다.

$$TR(V_i) = (1-d) + d \times \sum_{V_j \in IN(V_i)} \frac{W_{ji}}{\sum_{V_k \in OUT(V_i)} W_{jk}} TR(V_j)$$

[0020] [0021] 식에서 $TR(V_i)$ 는 i 번째 단어(V)에 대한 Textrank 값을 나타내며, W_{ji} 는 단어 i 와 j 사이의 가중치를 나타낸다. d 는 Dumping factor로써 0.85로 설정되며 $IN(V_i)$ 는 단어 i 와 연결된 모든 단어의 집합, $OUT(V_j)$ 는 단어 j 와 연

결된 모든 단어 집합이다.

- [0022] 도1은 본 발명의 기술문서 정보 시각화 및 기술 관계 분석 시스템의 실시예 흐름도이다.
- [0023] 도1에 도시된 바와 같이, 본 발명에 다른 기술문서 정보 시각화 및 기술 관계 분석 시스템은, 기술문서를 수집하는 단계(100)와, 자연어처리 기법을 적용한 기술문서 텍스트 부분 전처리 및 기술문서 임베딩을 수행하는 단계(200), 상기 수행한 기술문서 임베딩에 군집화 기법을 적용하여 세부기술을 식별하는 단계(300), 상기 식별된 세부기술별로 정량정보를 추출하고 생키 다이어그램으로 시각화하는 단계(400), 상기 세부기술 식별시 활용한 상위 단어 목록으로 기술문서 집합으로부터 전체 동시발생 행렬을 구축하고 그래프 모형을 통해 기술 관계를 분석하는 단계(500), 그 결과를 바탕으로 기술동향 파악 및 R&D 전략 수립에 활용한 단계(600)를 포함한다.
- [0024] 상기 자연어처리 기법을 적용한 기술문서 텍스트 부분 전처리 및 기술문서 임베딩 단계(200)는 자연어처리 기법으로써, 기술문서 집합으로부터 텍스트 부분을 추출하여 말뭉치를 구성하는 단계(201), 구성된 말뭉치에서 불용어, 구두점, 숫자, 제거 및 단어를 토큰화하여 전처리하는 단계(202), 마지막으로 전처리된 텍스트에 Do2Vec을 적용하여 기술문서를 임베딩하고 행렬을 구축하는 단계(203)로 구성한다.
- [0025] 상기 구성한 기술문서 임베딩 행렬에 군집화 기법을 적용하여 세부기술을 식별하는 단계(300)는 다양한 척도를 통해 세부기술 군집 개수를 설정하는 단계(301), 임베딩 행렬에 설정한 군집 개수에 따라 군집화 기법을 적용하여 세부기술 군집을 형성하는 단계(302), 형성된 세부기술 군집별로 단어 동시발생 행렬을 구축하는 단계(303), 군집별 그래프 모형을 적용하여 상위 단어들을 추출하고 이를 바탕으로 세부기술을 정의하는 단계(304)로 구성한다.
- [0026] 상기 식별된 세부기술별로 정량정보를 추출하고 생키 다이어그램으로 시각화하는 단계(400)는 세부기술별로 기술문서 건수를 식별하는 단계(401), 그리고 세부기술별로 시각화할 수 있는 정량정보를 추출하는 단계(402), 생키 다이어그램을 적용하여 기술문서 정보를 시각화하는 단계(403)로 구성한다.
- [0027] 상기 그래프 모형을 통한 기술 관계 분석 단계(500)는 앞서 식별된 세부군집별 상위 단어 목록을 활용하여 기술문서 집합으로부터 전체 동시발생 행렬을 구축하는 단계(501), 상기 구축한 전체 동시발생 행렬에 그래프 모형을 적용하여 기술 관계 그래프를 구축하는 단계(502), 상기 구축된 기술 관계 그래프를 바탕으로 기술 간의 관계를 식별하는 단계(503)을 포함한다.
- [0028] 실시예 1. 특허 빅데이터의 세부기술을 식별하고 기술개발 흐름을 시각화하는 방법
- [0029] 특허 빅데이터의 세부기술을 식별하고 기술개발 흐름을 시각화하는 방법을 나타낸 도 7의 단계는 다음과 같다. 먼저 수집된 특허 빅데이터에서 요약, 출원 국가, 출원인, 출원연도를 추출한다. 추출된 요약에는 텍스트 마이닝 기법을 적용하여 전처리하며, 다른 정량정보의 경우 상위 출원인 식별 및 출원 기간을 구분한다. 다음으로, 전처리된 특허 요약에 Doc2Vec을 적용하여 특정 공간에 특허를 임베딩한다. 또한, K-Means 군집화를 사용하여 세부 기술군집을 형성한다. 이때, 최적의 K값은 Elbow method를 통해 도출한다. 각 군집별로 Textrank를 수행하여 영향력이 높은 키워드를 추출하고 세부기술을 식별한다. 세부기술과 정량정보를 결합하여 생키 다이어그램으로 시각화하고 기술개발 흐름을 파악한다. 마지막으로, Textrank를 통해 도출된 영향력 높은 키워드들의 동시발생(Co-occurrence) 빈도를 이용하여 상관행렬을 구축하고 그래프로 표현한다. 실험 데이터는 헬스케어 기술과 관련된 특허로 KR, US, JP, EP에 출원된 것이다. 총 8,085건이며 출원 기간은 2005년부터 2018년까지이다.
- [0030] 헬스케어 관련 특허 빅데이터에서 출원 국가, 출원인, 출원연도를 먼저 추출하고 상위 20개의 출원인을 아래의 도 8과 같이 도출하였다. 또한, 출원연도를 2005-2009, 2010-2014, 2015-2018로 구분하였다.
- [0031] 요약을 추출하여 전처리하고 Doc2Vec의 Paragraph vector with distributed memory(PV-DM) 모형을 사용하여 200차원의 공간에 특허를 임베딩하였다. 임베딩 값에 K-Means 군집화를 사용하였으며 이때, 최적의 K값은 아래의 도 9와 같이 Elbow method를 통해 도출하였다.
- [0032] 도 9에서와 같이 도출된 K값을 사용하여 4개의 기술군집을 형성하고 군집별로 Textrank를 수행한다. 이때, Bigram으로 도출한 군집별 단어의 동시 발생 빈도를 사용하여 그래프를 구축한 후에 Textrank를 수행하였다.
- [0033] 도 9에서와 같이 군집별로 Textrank 값을 기준으로 상위 5개의 단어를 도출하였으며, 이를 통해 세부기술을 식별하였다. 세부기술로는 Cluster 1: 운동부하 모니터링 기술(Exercise tolerance monitoring technology), Cluster 2: 호흡상태 센싱 및 통신 기술(Breath sensing and communication technology), Cluster 3: 환자 혈압/혈당 모니터링 기술(Patient blood pressure/glucose monitoring technology), Cluster 4: 스마트폰 연동

생체계측 기술(Smartphone-linked biometric technologo)이다.

- [0034] 식별된 4개의 세부기술과 앞서 추출한 정량정보를 결합하여 기술개발 흐름을 파악하기 위한 생키 다이어그램을 도출한다. 도 11은 헬스케어 기술의 개발 흐름을 파악하기 위한 생키 다이어그램이다. 생키 다이어그램에서 최대 시장국은 미국이며, 최상위 출원인은 PHILIPS인 것을 확인할 수 있다. 또한, 헬스케어 기술은 최근 들어 기술 개발이 활발해지고 있으며 과거에는 Cluster 1인 운동부하 모니터링 기술이 주로 개발되었으나, 2010년 이후부터 Cluster 4인 스마트폰 연동 생체계측 기술이 많이 개발되고 있는 것으로 보인다. 이는 스마트폰이 대중화되기 시작한 시점과 일치한다. 상위 출원인들은 과거부터 현재까지 꾸준히 특허출원을 하고 있다. 이러한 이유로, 헬스케어 기술 분야는 스타트업 등의 신규 진입자들에게 기술장벽이 높을 것으로 사료된다.
- [0035] 도 11의 생키 다이어그램은 기술개발 흐름을 거시적인 관점에서 확인하기에 유용하나, 기술 간의 영향 정도를 파악하는 것에 한계점이 있다. 따라서 기술 간의 관계를 시각화하기 위해 도 8에서와 같이 도출된 군집별 상위 키워드를 활용한다. 이는 중복된 것을 제외한 총 15개의 키워드로써, 동시 발생 빈도를 통해 상관행렬을 구축한다. 구축된 상관행렬은 헬스케어 기술에서 영향력이 높은 15개 키워드들의 관계를 나타낸다. 이를 그래프로 구축하여 중심이 되는 기술을 식별한다. 도 12는 기술 간의 관계를 그래프로 나타낸 것이다.
- [0036] 도 12에서 다른 기술들과 연결이 많은 키워드는 communic.unit이다. 이는 통신 관련 기술이 헬스케어 기술 분야에서 핵심적인 역할을 하는 것으로 볼 수 있다. 또한 smart.phone은 센싱 및 통신 관련 키워드와 연결이 많다.
- [0037] 최근 특허 빅데이터를 분석하여 다양한 가치를 창출하는 사례가 증가하고 있다. 이와 더불어 분석 결과를 효과적으로 전달하기 위한 특허 빅데이터 시각화 기법이 각광 받고 있으나, 기존의 방법들은 단일 목적 표현에 초점이 맞추어져 있다.
- [0038] 본 발명은 특허 빅데이터에서 세부기술을 정량적으로 식별하고 기술개발 흐름을 거시적인 관점에서 파악할 수 있는 시각화 방법을 제안하였다. 제안된 방
- [0039] 법은 텍스트 마이닝, 특허 임베딩, 군집화, Textrank를 사용하여 세부기술을 식별하고 정량정보와 함께 생키 다이어그램으로 기술개발 흐름을 표현하였다. 또한, 기술 간의 관계를 파악하기 위해 주요 키워드들의 상관행렬을 구축하고 이를 그래프 모형으로 표현하였다. 실시예로써, 헬스케어 기술 분야에서 총 4개의 세부기술을 식별하였으며 최근 스마트폰 연동 생체계측 기술이 주로 개발되고 있는 것을 확인하였다. 추가적으로 통신 관련 기술이 헬스케어 기술에서 핵심적인 역할을 수행하고 있는 것을 도출하였다.
- [0040] 이상에서 설명한 본 발명은, 본 발명이 속하는 기술분야에서 통상의 지식을 가진 자에 있어 본 발명의 기술적 사상을 벗어나지 않는 범위 내에서 여러 가지 치환, 변형 및 변경이 가능하므로, 전술한 실시예 및 첨부된 도면에 한정되는 것이 아니다.

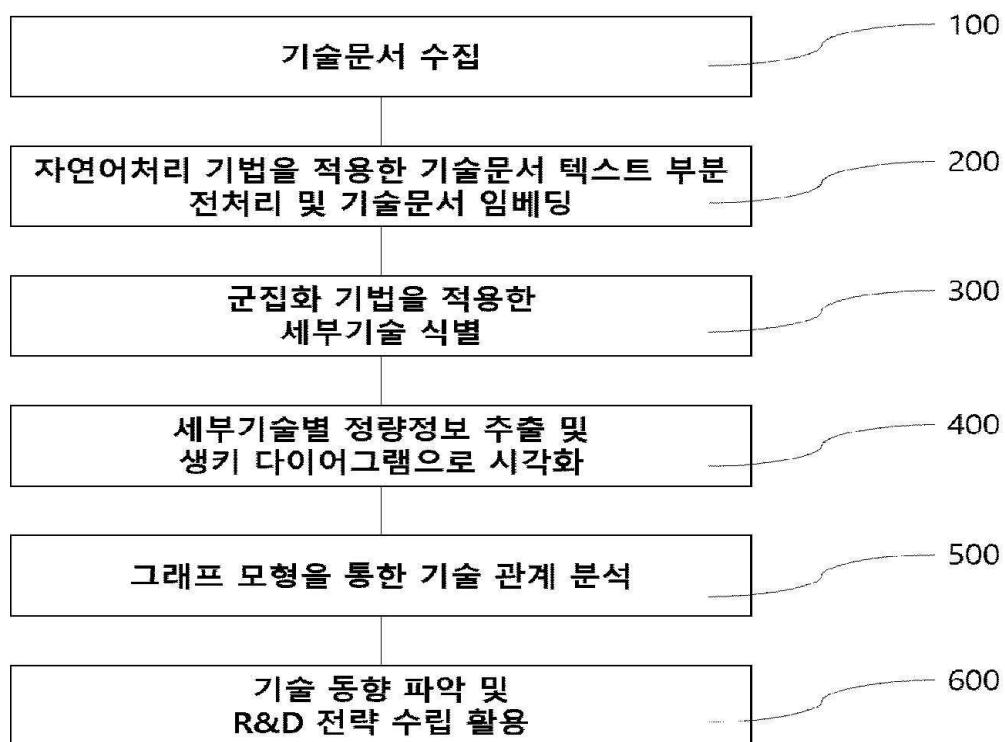
부호의 설명

- [0041] 100: 기술문서 수집부
- 200: 자연어처리 기법을 적용한 기술문서 텍스트 부분 전처리 및 기술문서 임베딩부
- 201: 말뭉치 구성부
- 202: 불용어, 구두점, 숫자 제거 및 단어 토큰화부
- 203: Doc2Vec으로 기술 문서 임베딩 및 임베딩 행렬 구축부
- 300: 군집화 기법을 적용한 세부기술 식별부
- 301: 세부기술 군집 개수 설정부
- 302: 임베딩 행렬에 군집화를 적용하여 세부기술 군집 식별부
- 303: 군집별 단어 동시발생 행렬 구축부
- 304: 군집별 그래프 모형 적용을 통한 상위 단어 추출 및 세부기술 정의부
- 400: 세부기술별 정량정보 추출 및 생키 다이어그램으로 시각화부
- 401: 세부기술별 기술문서 건수 식별부

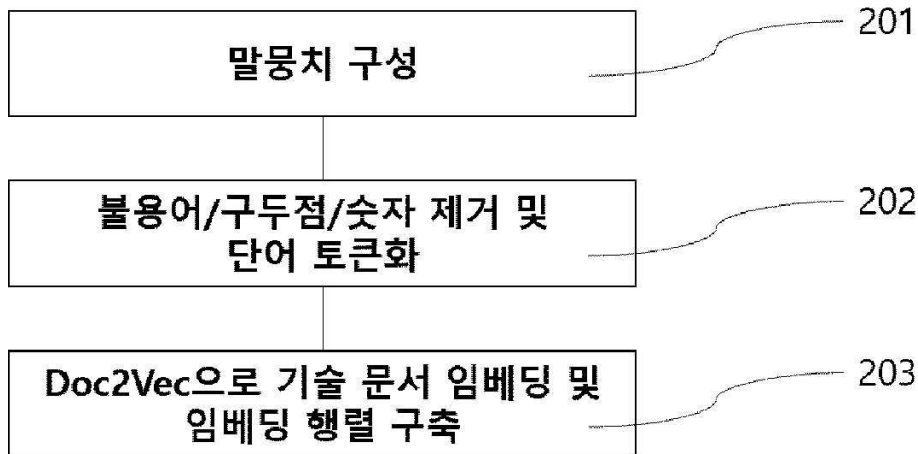
- 402: 세부기술별 정량정보 추출부
- 403: 생키 다이어그램을 적용한 기술문서 정보 시각화부
- 500: 그래프 모형을 통한 기술 관계 분석부
- 501: 군집별 상위 단어들의 기술문서 집합 전체 동시발생 행렬 구축부
- 502: 그래프 모형을 적용한 기술 관계 그래프 구축부
- 503: 기술 관계 식별부
- 600: 기술 동향 파악 및 R&D 전략 수립 활용부

도면

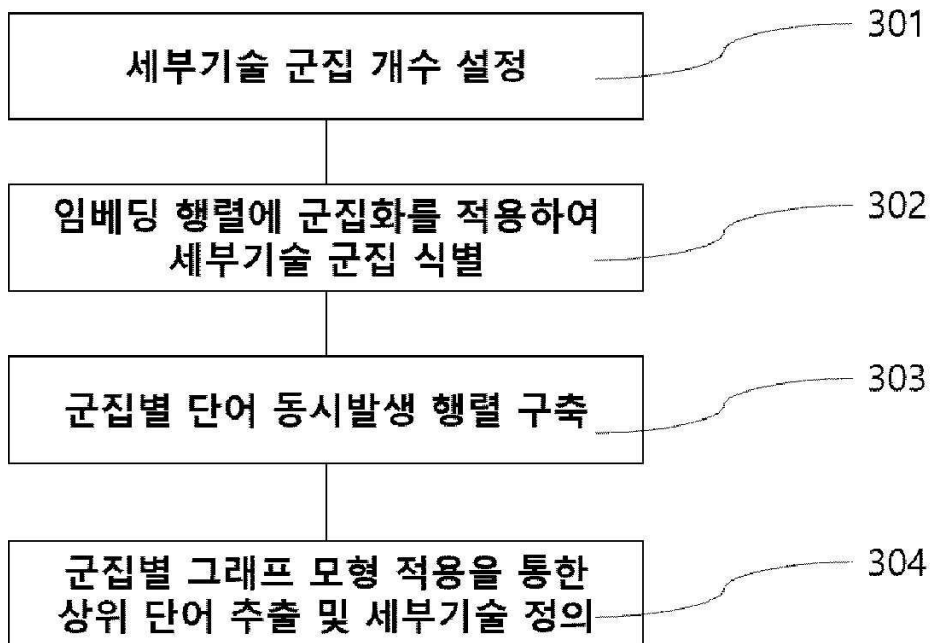
도면1



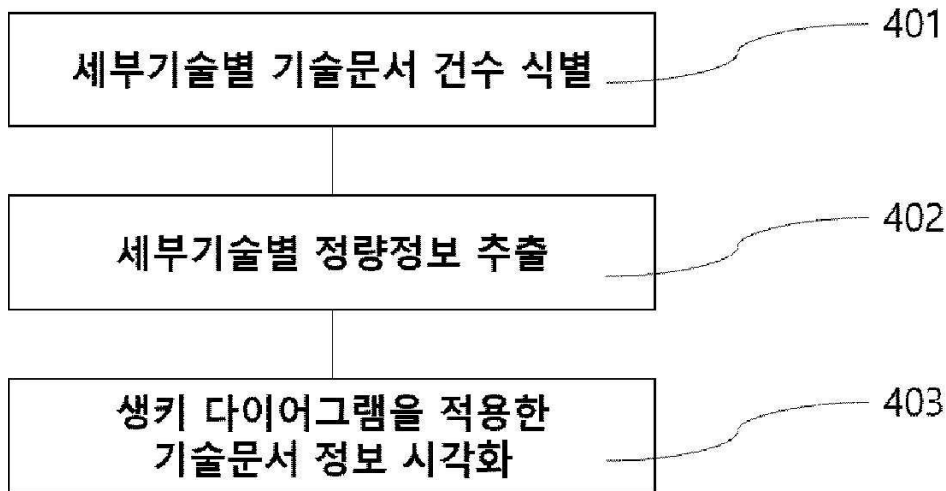
도면2



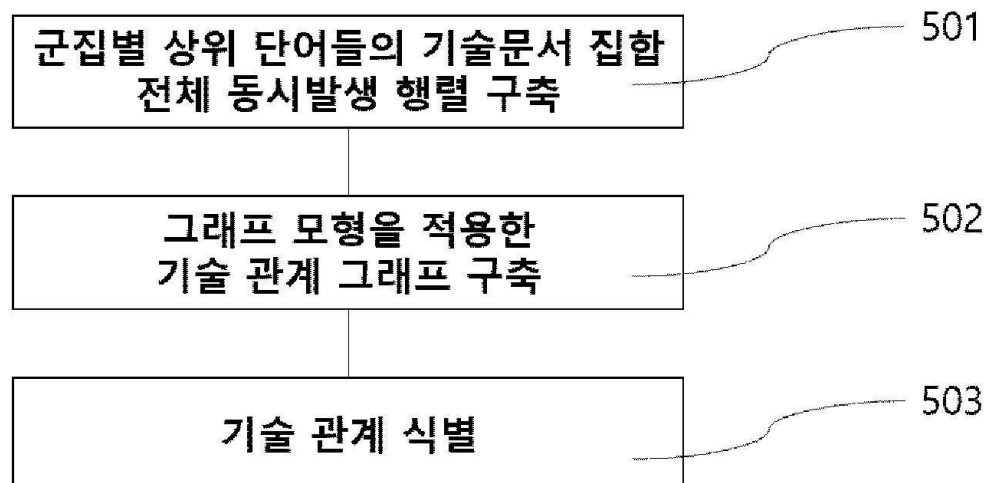
도면3



도면4



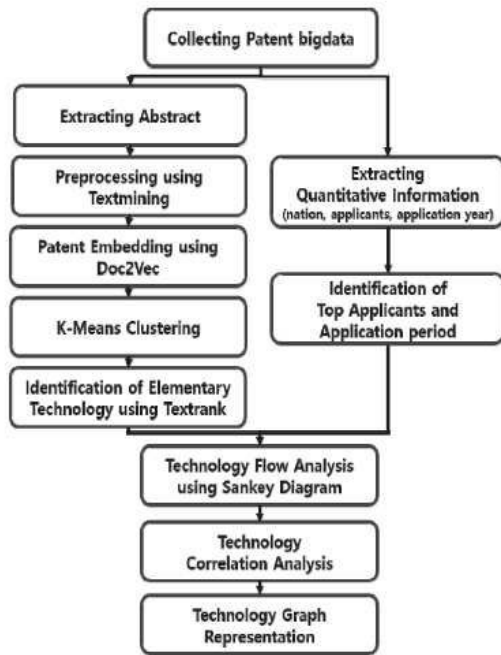
도면5



도면6

$$TR(V_i) = (1-d) + d \times \sum_{V_j \in IN(V_i)} \frac{W_{ji}}{\sum_{V_k \in OUT(V_i)} W_{jk}} TR(V_j)$$

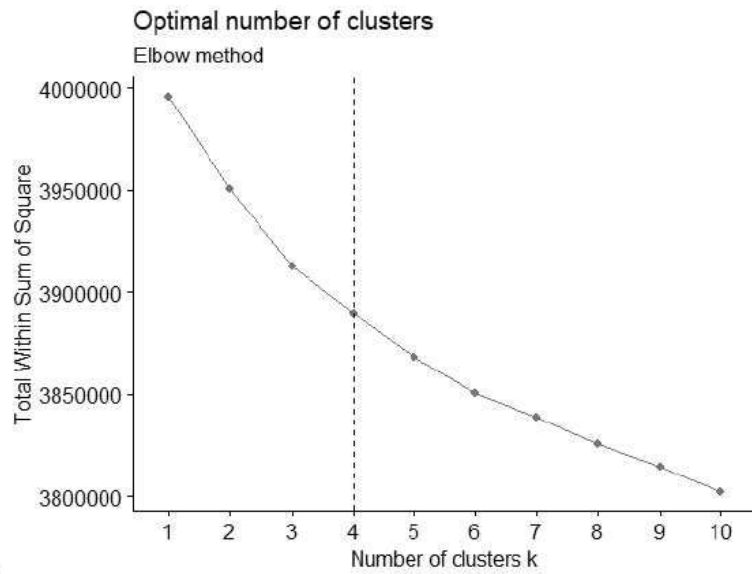
도면7



도면8

Rank	Applicant	Number of patent
1	PHILIPS	431
2	NIKE	191
3	SAMSUNG ELECTRONICS	139
4	FITBIT	115
5	ADIDAS	108
6	VALENCELL	85
7	BODYMEDIA	74
8	SOTERA WIRELESS	66
9	GENERAL ELECTRIC	64
10	HEALTH HERO NETWORK	60
11	ABBOTT	57
12	CARDIAC PACEMAKERS	57
13	APPLE	55
14	HILL-ROM SERVICES	48
15	HITACHI	43
16	DEXCOM	40
17	OMRON	40
18	PANASONIC	38
19	QUALCOMM	38
20	ROBERT BOSCH	38
-	Others	6,298
Total	8,085	

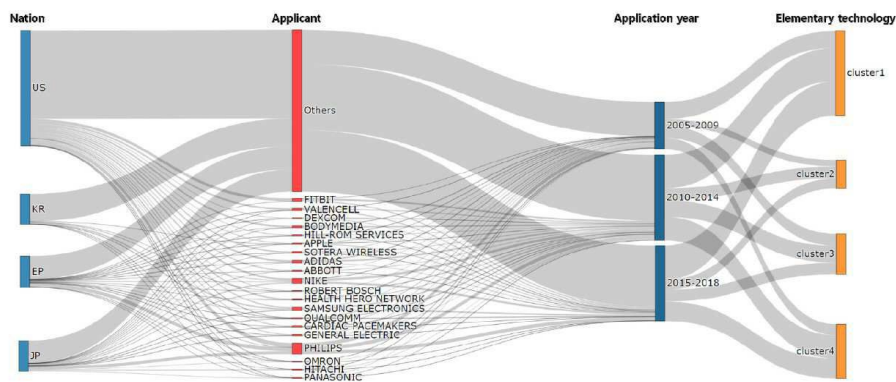
도면9



도면10

Cluster 1	Cluster 2	Cluster 3	Cluster 4
heart rate	heart rate	blood pressur	bodi temperatur
physic activ	sensor unit	patient monitor	smart phone
motion sensor	communic unit	monitor patient	heart rate
remot monitor	puls wave	patient medic	blood pressur
blood pressur	control unit	blood glucos	puls wave
Elementary technology identification			
Exercise tolerance monitoring technology	Breath sensing and communication technology	Patient blood pressure/glu cose monitoring technology	Smartphone - linked biometric technology

도면11



도면12

