



(19) 대한민국특허청(KR)
(12) 공개특허공보(A)

(11) 공개번호 10-2015-0102477

(43) 공개일자 2015년09월07일

(51) 국제특허분류(Int. Cl.)

G06F 19/00 (2011.01) G06F 17/30 (2006.01)

(21) 출원번호 10-2014-0024396

(22) 출원일자 2014년02월28일

심사청구일자 2014년02월28일

(71) 출원인

고려대학교 산학협력단

서울특별시 성북구 안암로 145, 고려대학교 (안암동5가)

(72) 발명자

장동식

서울 성북구 안암로9길 30, 2동 1102호 (안암동3가, 삼익아파트)

박상성

서울 동대문구 한천로58길 107, 104동 2105호 (이문동, 현대아파트)

(뒷면에 계속)

(74) 대리인

김동용, 문경진

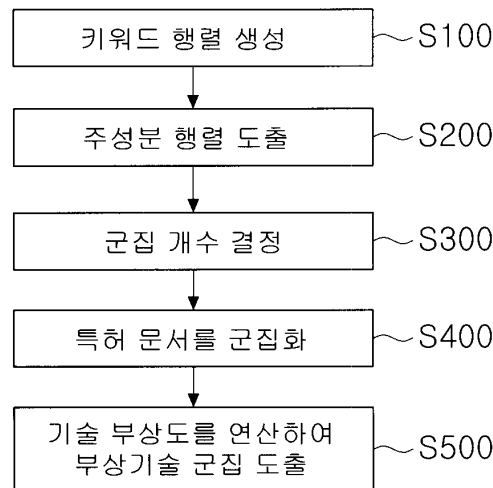
전체 청구항 수 : 총 8 항

(54) 발명의 명칭 부상기술 예측방법 및 장치

(57) 요약

본 발명의 일실시예에 따른 부상기술 예측방법은 다수의 특허문서들을 텍스트 마이닝하여 키워드를 도출한 후 키워드 행렬을 생성하는 단계, 생성된 키워드 행렬에 특이값 분해(SVD, Singular Value Decomposition)과정을 수행하고, 주성분 분석(PCA, Principal Component Analysis)을 통해 키워드 행렬로부터 주성분 행렬을 도출하는 단계, 다수의 특허문서들을 분류하기 위한 군집 개수를 결정하는 단계, 군집화 알고리즘을 이용하여 다수의 특허문서들 각각을 다수의 군집들 중 하나로 군집화하는 단계 및 다수의 군집들 각각의 평균 패밀리 특허규모, 평균 피인용수 및 출원후 평균 경과 기간을 기초로 다수의 군집들 중 부상기술(emerging technology)에 대응하는 군집을 결정하는 단계를 포함한다.

대표도 - 도1



(72) 발명자

김갑조

서울 성북구 안암로 145, 553호 공학관 (안암동5가, 고려대학교이과대학)

이준혁

서울 성북구 안암로 145, 산학협력단 자연계 553호 캠퍼스 공학관 (안암동5가, 고려대학교)

이 발명을 지원한 국가연구개발사업

과제고유번호 2013053572

부처명 교육부

연구관리전문기관 한국연구재단

연구사업명 (이공)일반연구자-기본(유형 II)

연구과제명 시맨틱 기술 기반 IP 마이닝 시스템 개발

기 여 율 1/2

주관기관 고려대학교 산학협력단

연구기간 2013.09.01 ~ 2014.08.31이 발명을 지원한 국가연구개발사업

과제고유번호 F13SN08T4605

부처명 교육부

연구관리전문기관 한국연구재단

연구사업명 제조 물류분야에서의 빅데이터 운용사업팀

연구과제명 BK21PLUS21

기 여 율 1/2

주관기관 고려대학교 산학협력단

연구기간 2013.09.01 ~ 2014.02.28

특허청구의 범위

청구항 1

다수의 특허문서들을 텍스트 마이닝하여 키워드를 도출한 후 키워드 행렬을 생성하는 단계;

생성된 상기 키워드 행렬에 특이값 분해(SVD, Singular Value Decomposition)과정을 수행하고, 주성분 분석(PCA, Principal Component Analysis)을 통해 상기 키워드 행렬로부터 주성분 행렬을 도출하는 단계;

상기 다수의 특허문서들을 분류하기 위한 군집 개수를 결정하는 단계;

군집화 알고리즘을 이용하여 상기 다수의 특허문서들 각각을 다수의 군집들 중 하나로 군집화하는 단계; 및

상기 다수의 군집들 각각의 평균 패밀리 특허규모, 평균 피인용수 및 출원후 평균 경과 기간을 기초로 상기 다수의 군집들 중 부상기술(emerging technology)에 대응하는 군집을 결정하는 단계를 포함하는 것을 특징으로 하는 부상기술 예측방법.

청구항 2

제1항에 있어서,

상기 군집화 개수를 결정하는 단계는 BIC(Bayesian Information Criterion), AIC(Akaike Information Criterion) 또는 Silhouette측도를 활용하여 결정하는 것을 특징으로 하는 부상기술 예측방법.

청구항 3

제1항에 있어서,

상기 군집화 알고리즘은 k-평균 알고리즘(k-means algorithm) 또는 k-메도이드 알고리즘(k-medoid algorithm)인 것을 특징으로 하는 부상기술 예측방법.

청구항 4

제1항에 있어서,

상기 부상기술에 대응하는 군집을 결정하는 단계는 상기 다수의 군집들 각각의 기술 부상도를 계산하는 단계를 포함하고,

상기 기술 부상도는 상기 평균 패밀리 특허 규모에 비례하고, 상기 평균 피인용수와 상기 출원 후 평균 경과 기간에 반비례하며,

가장 큰 기술 부상도를 갖는 군집이 상기 부상 기술에 대응하는 군집으로 결정되는 것을 특징으로 하는 부상기술 예측방법.

청구항 5

제1항 내지 제4항 중 어느 한 항의 방법을 실행하기 위한 프로그램이 기록되어 있는 것을 특징으로 하는 컴퓨터 판독 가능한 기록 매체.

청구항 6

다수의 특허문서들에서 텍스트 마이닝하여 키워드를 도출하는 키워드 도출부;

상기 키워드 도출부에서 도출된 키워드와 특허문서들을 매칭시켜 키워드 행렬을 생성하고, 상기 키워드 행렬에 특이값 분해를 수행하며, 주성분 분석(PCA)을 통해 주성분 행렬을 생성하는 행렬 생성부;

상기 다수의 특허문서들을 분류하기 위한 군집 개수를 결정하고, 상기 다수의 특허문서들 각각을 다수의 군집들 중 하나로 군집화하는 클러스터링부; 및

상기 다수의 군집들 중 부상기술에 대응하는 군집을 결정하는 군집결정부를 포함하는 것을 특징으로 하는 부상기술 예측장치.

청구항 7

제6항에 있어서,

상기 클러스터링부는 군집 개수를 결정하기 위해 BIC(Bayesian Information Criterion), AIC(Akaike Information Criterion) 또는 Silhouette측도를 활용하고, 군집화하기 위해 군집화 알고리즘인 k-평균 알고리즘(k-means algorithm) 또는 k-메도이드 알고리즘(k-medoid algorithm)을 활용하는 것을 특징으로 하는 부상기술 예측장치.

청구항 8

제6항에 있어서,

상기 군집결정부는 다수의 군집들 각각의 평균 패밀리 특허규모, 평균 피인용수 및 출원후 평균 경과 기간을 기초로 상기 다수의 군집들 중 부상기술에 대응하는 군집을 결정하기 위한 기술 부상도를 연산하고, 가장 큰 기술 부상도를 갖는 군집을 상기 부상 기술에 대응하는 군집으로 결정하며,

상기 기술 부상도는 상기 평균 패밀리 특허 규모에 비례하고, 상기 평균 피인용수와 상기 출원 후 평균 경과 기간에 반비례하는 것을 특징으로 하는 부상기술 예측장치.

명세서

기술분야

[0001] 본 발명은 부상기술 예측방법 및 장치에 관한 발명으로서 공개된 특허문서를 활용하여 부상기술을 예측하는 방법 및 그 장치에 관한 발명이다.

배경기술

[0002] 산업이 발전함으로서 지식재산에 대한 중요성이 한층 부각되고 있는 현실에서 산업 재산권의 확보는 기업의 경쟁력과 밀접한 관련이 있다. 산업 재산권은 개발한 기술에 대해서 노력과 창의성에 대한 보상으로 일정기간 독점적으로 사용할 수 있는 권리를 주는 것과 동시에 독점 사용기간이 끝나는 시점에는 모든 사람들이 자유롭게 사용하게 함으로서 기술의 발전을 도모하기 위해 생겨났다. 이와 같은 산업 재산권의 특성상 선행 특허권의 존재를 모른채 신제품 개발을 하였다가 특허분쟁에 휘말리거나 과도한 로열티를 지급하여 경영상의 타격을 입고 나아가 개발한 신제품을 포기해야 하는 경우가 발생할 수 있다.

[0003] 산업 재산권은 신사업을 개척하거나 신수종을 발굴, 신제품을 개발, 생산하기 위해서 반드시 조사해야하고, 연구해야할 대상이 되었다. 산업 재산권 좁게는 특허의 경우에는 기업의 성공과 생존의 중요한 요인으로 받아들여지고 있는 현 시점에서 특허는 오히려 기업이 신제품을 개발하기 위한 아이템을 제공하는 중요 판단기준으로 활용되고 있다. 즉 특허의 분석을 통해 유망기술을 발굴하고, 신제품을 개발하거나 생산하게 될 경우 주요 경쟁상대 기업은 어느 기업인지, 핵심기술이 무엇이며, 보유기업은 어느 기업인지 등 쉽게 접할 수 없는 기술 및 시장에 대한 동향을 파악할 수 있다. 특히 공백기술에 대한 결론도출이 가능할 수 있어, 기술개발을 통한 시장진입 가능성을 한층 높일 수 있다.

[0004] 특허 분석은 사업성공, 실패를 결정짓는 중요한 결정도구가 되어가고 있는 현실에서 특허 분석결과와 신뢰성을 담보할 수 없는 문제와 특허 분석을 통한 결론도출이 쉽지 않은 문제가 있다. 즉 기존 연구들은 공백기술을 도출하는데 단지 군집에 포함된 원소의 수만을 고려한다는 한계점이 있다. 이러한 방법은 공백기술을 도출하게 되면 별로 중요하지 않은 기술 군이기 때문에 특허수가 적은 경우도 공백기술 군으로 결정되는 문제가 있으며 출원된 특허의 수는 적지만 이미 충분히 개발된 기술군집도 공백기술로 정의되는 문제가 있다.

발명의 내용

해결하려는 과제

[0005] 본 발명은 위와 같은 문제를 해결하기 위해 안출된 것으로서 특허분석을 통해 부상기술을 예측하는데 있어서 신뢰성이 높은 분석결과를 도출하기 위한 부상기술 예측방법 및 장치를 제공하는데 목적이 있다.

과제의 해결 수단

- [0006] 본 발명의 일실시예에 따른 부상기술 예측방법은 다수의 특허문서들을 텍스트 마이닝하여 키워드를 도출한 후 키워드 행렬을 생성하는 단계, 생성된 키워드 행렬에 특이값 분해(SVD, Singular Value Decomposition)과정을 수행하고, 주성분 분석(PCA, Principal Component Analysis)을 통해 키워드 행렬로부터 주성분 행렬을 도출하는 단계, 다수의 특허문서들을 분류하기 위한 군집 개수를 결정하는 단계, 군집화 알고리즘을 이용하여 다수의 특허문서들 각각을 다수의 군집들 중 하나로 군집화하는 단계 및 다수의 군집들 각각의 평균 패밀리 특허규모, 평균 피인용수 및 출원후 평균 경과 기간을 기초로 다수의 군집들 중 부상기술(emerging technology)에 대응하는 군집을 결정하는 단계를 포함한다.
- [0007] 여기서, 군집화 개수를 결정하는 단계는 BIC(Bayesian Information Criterion), AIC(Akaike Information Criterion) 또는 Silhouette측도를 활용하여 결정할 수 있다.
- [0008] 여기서, 군집화 알고리즘은 k-평균 알고리즘(k-means algorithm) 또는 k-메도이드 알고리즘(k-medoid algorithm)일 수 있다.
- [0009] 여기서, 부상기술에 대응하는 군집을 결정하는 단계는 다수의 군집들 각각의 기술 부상도를 계산하는 단계를 포함하고, 기술 부상도는 평균 패밀리 특허 규모에 비례하고, 평균 피인용수와 출원 후 평균 경과 기간에 반비례하며, 가장 큰 기술 부상도를 갖는 군집이 부상기술에 대응하는 군집으로 결정될 수 있다.
- [0010] 본 발명의 일실시예에 따른 부상기술 예측장치는 다수의 특허문서들에서 텍스트 마이닝하여 키워드를 도출하는 키워드 도출부, 키워드 도출부에서 도출된 키워드와 특허문서들을 매칭시켜 키워드 행렬을 생성하고, 키워드 행렬에 특이값 분해를 수행하며, 주성분 분석(PCA)을 통해 주성분 행렬을 생성하는 행렬 생성부, 다수의 특허문서들을 분류하기 위한 군집 개수를 결정하고, 다수의 특허문서들 각각을 다수의 군집들 중 하나로 군집화하는 클러스터링부 및 다수의 군집들 중 부상기술에 대응하는 군집을 결정하는 군집결정부를 포함한다.
- [0011] 여기서, 클러스터링부는 군집 개수를 결정하기 위해 BIC(Bayesian Information Criterion), AIC(Akaike Information Criterion) 또는 Silhouette측도를 활용하고, 군집화하기 위해 군집화 알고리즘인 k-평균 알고리즘(k-means algorithm) 또는 k-메도이드 알고리즘(k-medoid algorithm)을 활용할 수 있다.
- [0012] 여기서, 군집결정부는 다수의 군집들 각각의 평균 패밀리 특허규모, 평균 피인용수 및 출원후 평균 경과 기간을 기초로 다수의 군집들 중 부상기술에 대응하는 군집을 결정하기 위한 기술 부상도를 연산하고, 가장 큰 기술 부상도를 갖는 군집을 부상 기술에 대응하는 군집으로 결정하며, 기술 부상도는 평균 패밀리 특허 규모에 비례하고, 평균 피인용수와 출원 후 평균 경과 기간에 반비례할 수 있다.

발명의 효과

- [0013] 본 발명의 일실시예에 따른 부상기술 예측방법 및 장치에 의하면 신뢰성이 높은 특허분석 결과를 제공함으로써 부상기술 예측이 가능하고, 공백기술에 대한 정확한 정보의 제공이 가능하다.

도면의 간단한 설명

- [0014] 도1은 본 발명인 부상기술 예측방법의 흐름도이다.
- 도2는 특허문서를 검색하는 시스템도이다.
- 도3은 특허검색 수행결과 연도별 특허출원수를 도시한 그래프이다.
- 도4는 실루엣 측도를 활용한 군집 개수 결정 구성도이다.
- 도5는 실루엣 측도를 활용한 군집 개수 결정의 결과 그래프이다.
- 도6은 k-medoids를 활용한 군집화 구성도이다.
- 도7은 본 발명의 일실시예에 따른 부상기술 예측장치의 구성도이다.

발명을 실시하기 위한 구체적인 내용

- [0015] 본 명세서에 개시되어 있는 본 발명의 개념에 따른 실시 예들에 대해서 특정한 구조적 또는 기능적 설명은 단지

본 발명의 개념에 따른 실시 예들을 설명하기 위한 목적으로 예시된 것으로서, 본 발명의 개념에 따른 실시 예들은 다양한 형태들로 실시될 수 있으며 본 명세서에 설명된 실시 예들에 한정되지 않는다.

[0016] 본 발명의 개념에 따른 실시 예들은 다양한 변경들을 가할 수 있고 여러 가지 형태들을 가질 수 있으므로 실시 예들을 도면에 예시하고 본 명세서에서 상세하게 설명하고자 한다. 그러나, 이는 본 발명의 개념에 따른 실시 예들을 특정한 개시 형태들에 대해 한정하려는 것이 아니며, 본 발명의 사상 및 기술 범위에 포함되는 모든 변경, 균등물, 또는 대체물을 포함한다.

[0017] 제1 또는 제2 등의 용어는 다양한 구성 요소들을 설명하는데 사용될 수 있지만, 상기 구성 요소들은 상기 용어들에 의해 한정되어서는 안 된다. 상기 용어들은 하나의 구성 요소를 다른 구성 요소로부터 구별하는 목적으로만, 예컨대 본 발명의 개념에 따른 권리 범위로부터 벗어나지 않은 채, 제1 구성 요소는 제2 구성 요소로 명명될 수 있고 유사하게 제2 구성 요소는 제1 구성 요소로도 명명될 수 있다.

[0018] 어떤 구성 요소가 다른 구성 요소에 "연결되어" 있다거나 "접속되어" 있다고 언급된 때에는, 그 다른 구성 요소에 직접적으로 연결되어 있거나 또는 접속되어 있을 수도 있지만, 중간에 다른 구성 요소가 존재할 수도 있다고 이해되어야 할 것이다. 반면에, 어떤 구성 요소가 다른 구성 요소에 "직접 연결되어" 있다거나 "직접 접속되어" 있다고 언급된 때에는 중간에 다른 구성 요소가 존재하지 않는 것으로 이해되어야 할 것이다. 구성 요소들 간의 관계를 설명하는 다른 표현들, 즉 "~사이에"와 "바로 ~사이에" 또는 "~에 이웃하는"과 "~에 직접 이웃하는" 등도 마찬가지로 해석되어야 한다.

[0019] 본 명세서에서 사용한 용어는 단지 특정한 실시 예를 설명하기 위해 사용된 것으로서, 본 발명을 한정하려는 의도가 아니다. 단수의 표현은 문맥상 명백하게 다르게 뜻하지 않는 한, 복수의 표현을 포함한다. 본 명세서에서, "포함하다" 또는 "가지다" 등의 용어는 본 명세서에 기재된 특징, 숫자, 단계, 동작, 구성 요소, 부분품 또는 이들을 조합한 것이 존재함을 지정하려는 것이지, 하나 또는 그 이상의 다른 특징들이나 숫자, 단계, 동작, 구성 요소, 부분품 또는 이들을 조합한 것들의 존재 또는 부가 가능성을 미리 배제하지 않는 것으로 이해되어야 한다.

[0020] 다르게 정의되지 않는 한, 기술적이거나 과학적인 용어를 포함해서 여기서 사용되는 모든 용어들은 본 발명이 속하는 기술 분야에서 통상의 지식을 가진 자에 의해 일반적으로 이해되는 것과 동일한 가진다. 일반적으로 사용되는 사전에 정의되어 있는 것과 같은 용어들은 관련 기술의 문맥상 가지는 의미와 일치하는 의미를 갖는 것으로 해석되어야 하며, 본 명세서에서 명백하게 정의하지 않는 한, 이상적이거나 과도하게 형식적인 의미로 해석되지 않는다.

[0021] 이하, 본 명세서에 첨부된 도면들을 참조하여 본 발명의 실시 예들을 상세히 설명한다.

[0022] 도1은 본 발명인 부상기술 예측방법의 흐름도이다.

[0023] 도1에 도시된 바와 같이 본 발명인 부상기술 예측방법은 다수의 특허문서들을 텍스트 마이닝하여 키워드를 도출한 후 키워드 행렬을 생성하는 단계(S100), 생성된 키워드 행렬에 특이값 분해(SVD, Singular Value Decomposition)과정을 수행하고, 주성분 분석(PCA, Principal Component Analysis)을 통해 키워드 행렬로부터 주성분 행렬을 도출하는 단계(S200), 다수의 특허문서들을 분류하기 위한 군집 개수를 결정하는 단계(S300), 군집화 알고리즘을 이용하여 다수의 특허문서들 각각을 다수의 군집들 중 하나로 군집화하는 단계(S400) 및 다수의 군집들 각각의 평균 패밀리 특허규모, 평균 피인용수 및 출원후 평균 경과 기간을 기초로 다수의 군집들 중 부상기술(emerging technology)에 대응하는 군집을 결정하는 단계(S500)를 포함한다.

[0024] 본 발명의 일실시예에 따른 부상기술 예측방법은 특허문서를 이용하는 예측방법으로 특허문서에 기재된 텍스트를 기초로 키워드를 도출하고, 도출된 키워드를 이용하여 특허문서와 키워드와의 관계를 행렬로 표시한 키워드 행렬($X_{n \times k}$)을 생성한다(S100).

[0025] 다만 키워드 행렬을 생성하기 전에 먼저 특정 기술분야의 특허문서에 대한 검색을 수행한다. 이는 키워드 행렬을 생성하기 위한 미가공 데이터(raw data)를 추출하는 과정이다.

[0026] 도2는 특허문서를 검색하는 시스템도이다.

[0027] 도2에 도시된 바와 같이 사용자는 유저인터페이스를 통해 검색엔진으로 방대한 특허문서가 축적되어 있는 특허 DB를 검색한다. 특허 DB를 통해 검색된 특정 기술분야의 특허문서는 저장매체로 저장된다.

[0028] 특허문서는 인터넷 및 분산된 데이터베이스로부터 수집될 수도 있으며, 대표적으로 한국특허청(KIPO)이나 미국

특허청(USPTO) 등과 같이 공공적 성격을 가지고 있는 데이터베이스에 접속하여 하이퍼텍스트 전송프로토콜 형태나 텍스트 파일 형태의 특허문서들을 다운로드 받는다. 검색엔진을 통해 검색된 특허문서는 저장매체에 저장되고, 부상기술 예측을 위한 미가공 데이터(raw data)가 된다.

도3은 특허검색 수행결과 연도별 특허출원수를 도시한 그래프이다.

도3에 도시된 그래프는 본 발명의 일실시예에 따른 부상기술 예측방법을 수행하기 위해서 무선랜 정보기술 표준인 IEEE 802.11g 와 관련된 특허문서를 검색하여 연도별로 특허출원 건수를 도시한 그래프로써, 연도별 특허출원동향을 파악할 수 있다. 그러나 본 발명의 일실시예에 따른 부상기술 예측방법은 검색을 통한 미가공 데이터의 단순 연도별 출원동향 뿐만 아니라 미가공 데이터 개별적으로 기재되어 있는 텍스트 내용을 확인하게 된다.

이하 검색을 통해 추출된 미가공 데이터인 특허문서를 통해 키워드 행렬을 도출하는 방법에 대해서 살펴본다.

키워드 행렬을 도출하기 위해서 먼저 텍스트 마이닝을 수행한다. 특허 문서상에 기재되어 있는 텍스트 중에서 키워드를 추출하는 텍스트 마이닝은 전자출원형식의 파일 또는 PDF와 같이 파일 형식으로 이루어져 있는 비정형 데이터에 대하여 자연언어처리 기술과 문서 처리기술을 적용하여 유용한 정보를 추출, 가공하는 것이다. 검색된 전체 특허문서에 대한 텍스트 마이닝이 완료되면 전체 특허문서를 1에서 n까지 임의의 순서대로 나열하고 텍스트 마이닝 된 텍스트들을 1에서 k까지 임의의 순서대로 나열한다. 각 나열된 n개의 전체 특허문서를 아래 수학식 1과 같이 행렬로 표현하고, 각 나열된 k개의 텍스트를 아래 수학식 2와 같이 행렬로 표현한다.

수학식 1

$$Y_n = \begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ \vdots \\ y_n \end{pmatrix}$$

수학식 2

$$T_k = \begin{pmatrix} t_1 \\ t_2 \\ t_3 \\ \vdots \\ t_k \end{pmatrix}$$

수학식 1과 수학식 2 사이에는 아래 수학식 3과 같은 관계가 성립하고 다시 아래 수학식 4가 본 발명의 일실시예에 따른 부상기술을 예측하기 위한 키워드 행렬이 된다.

수학식 3

$$Y_n = (X_{n \times k})(T_k)$$

수학식 4

$$X_{n \times k} = \begin{pmatrix} x_{11} & x_{12} & x_{13} & \cdots & x_{1k} \\ x_{21} & x_{22} & \cdots & \cdots & \vdots \\ \vdots & \vdots & \cdots & \cdots & \vdots \\ x_{n1} & \cdots & \cdots & \cdots & x_{nk} \end{pmatrix}$$

[0037]

[0038]

수학식 4로 표현되는 키워드 행렬의 각 원소는 n개의 특허문서 각각이 k개의 키워드 각각을 포함하고 있는 개수를 나타낸다. 즉 x_{22} 는 2번째 특허문서가 2번째 키워드를 포함하고 있는 개수이다.

[0039]

키워드 행렬을 생성하고 난 이후에 특이값 분해(SVD)과정과 주성분 분석(PCA)을 통해 주성분 행렬을 도출한다(S200). 수학식 4와 같이 키워드 행렬은 일반적으로 열의 수가 행의 수보다 많기 때문에($k > n$) 주성분 분석(PCA)을 바로 수행하기가 어렵다. 이러한 문제를 해결하기 위해 특이값 분해(SVD)과정을 사용하여 행렬 데이터를 활용한 분석의 문제점을 해결한 이후에 주성분 분석(PCA)을 수행한다.

[0040]

주성분 분석(PCA)은 다양한 변수들이 관측될 때 이 변수들이 가지고 있는 전체 정보를 최대한 보유할 수 있는 적은 수의 새로운 변수들을 생성하여 전체 자료의 구조를 파악하기 쉽게 하는 방법이다.

[0041]

주성분 분석(PCA)을 통해 수학식 4로 표현된 키워드 행렬은 아래 수학식 5와 같은 주성분 행렬로 변환된다.

수학식 5

$$P_{n \times f} = \begin{pmatrix} p_{11} & p_{12} & p_{13} & \cdots & p_{1f} \\ p_{21} & p_{22} & \cdots & \cdots & \vdots \\ \vdots & \vdots & \cdots & \cdots & \vdots \\ p_{n1} & \cdots & \cdots & \cdots & p_{nf} \end{pmatrix}$$

[0042]

[0043]

$p_{n \times f}$ 는 n번째 특허문서의 f번째 주요 구성요소의 점수를 나타낸다.

[0044]

이처럼 본 발명의 일실시예에 따른 부상기술 예측방법은 위 수학식 5와 같은 주성분 행렬을 도출한 이후에 주성분 행렬을 이용하여 군집 개수를 결정한다(S300). 군집개수를 결정하기 위해서 BIC(Bayesian Information Criterion), AIC(Akaike Information Criterion) 또는 실루엣 측도(Silhouette width)를 활용한다.

[0045]

도4는 실루엣 측도를 활용한 군집 개수 결정 구성도이다.

[0046]

도5는 실루엣 측도를 활용한 군집 개수 결정의 결과 그래프이다.

[0047]

본 발명의 일실시예에 따른 부상기술 예측방법은 군집 개수를 결정하기 위해 실루엣 측도를 활용한다.

[0048]

도4에 도시된 바와 같이 주성분 행렬에 대해 실루엣 측도 연산을 수행하면 각 그룹별로 결과값이 도출된다. 앞선 예인 무선랜 정보기술 표준인 IEEE 802.11g 와 관련된 특허문서 검색을 통해 도출된 키워드 행렬을 주성분 행렬로 변환하고, 실루엣 측도를 활용한 군집 개수를 결정한 예를 통해 살펴보면, 도4 및 도5에서 확인 가능한 가장 높은 결과값을 나타내는 그룹수는 "3"으로 나타났다. 즉 다수의 특허 문서들을 분류하기 위한 군집개수는 "3"으로 결정된다.

[0049]

전체 특허 문서를 3개의 군집으로 분류하기 위해서 군집화 알고리즘을 이용하여 다수의 특허문서들 각각을 다수의 군집들 중 하나로 군집화한다(S400).

[0050]

도6은 k-medoid를 활용한 군집화 구성도이다.

[0051] 도6에 도시된 바와 같이 주성분 행렬로 군집화 알고리즘인 k-medoid를 수행한다. 여기서 k는 앞서 실루엣 측도를 통해 결정된 군집개수인 3이다. 이하 k-medoid를 수행하면 각 특허문서 별로 결과값이 도출된다. 결과값들을 서로 비교하여 유사한 결과값을 갖는 특허문서를 하나의 군집으로 묶는다. 도6에 도시된 표와 같이 X_1 의 결과값은 a_1 이고, X_3 의 결과값은 a_2 , X_4 의 결과값은 a_3 이다.

수학식 6

[0052]
$$a_1 \approx a_2 \approx a_3$$

[0053] 수학식 6과 같은 가정하에 특허문서 ($X_1, X_3, X_4, \dots$)를 a라는 군집으로 묶고, b_1, b_2 와 같은 유사한 결과값을 갖는 특허문서 ($X_2, X_5, \dots$)를 b라는 군집으로 묶는 등 총 a, b, c 세 개의 군집으로 각 특허문서를 분류한다. 본 발명의 일실시예에 따른 부상기술 예측방법은 앞선 예인 무선랜 정보기술 표준인 IEEE 802.11g와 관련된 기술에 대해서 군집 개수를 결정하는 실루엣 측도를 수행하여 "3"이라는 군집 개수를 예를 들어 설명했을 뿐 본 발명은 다양한 군집 개수로 특허문서의 분류가 가능하다.

[0054] 본 발명의 일실시예에 따른 부상기술 예측방법에 의해서 군집화된 특허문서는 다수의 군집들 각각의 평균 패밀리 특허규모, 평균 피인용수 및 출원후 평균 경과 기간을 기초로 다수의 군집들 중 부상기술에 대응하는 군집을 결정하는 단계를 수행한다(S500).

[0055] 부상기술에 대응하는 군집을 결정하는 단계는 다수의 군집들 각각의 기술 부상도를 계산하는 단계를 포함하고 기술 부상도는 아래 수학식 7과 같다.

수학식 7

[0056]
$$\text{기술부상도} = \frac{F_j}{C_j \times L_j}$$

[0057] F_j 는 j군집 내 특허들의 평균 패밀리 특허규모를 나타내고, C_j 는 j군집 내 특허들의 평균 피인용수를 나타내고, L_j 는 j군집 내 특허들의 평균 출원 후 경과기간을 나타낸다.

[0058] 앞선 예인 무선랜 정보기술 표준인 IEEE 802.11g와 관련된 기술을 통해 3개의 군집 각각에 대한 기술부상도를 계산한 결과가 아래 표1에 나타나 있다.

표 1

구분	군집1	군집2	군집3
군집내 top10 키워드	signal, network, plurality, communication, comprising, data, packet, wireless, method, channel	configured, device, access, apparatus, module, communication, transmit, wireless, packets, receiving	control, envelope, frequency, output, plurality, power, signal, constant, phase, comprising
점유율(%)	65.87	9.13	25
평균 패밀리 특허규모(F_j)	17.146	9.36	37.673
평균 피인용수 (C_j)	7.26	2.894	20.13
평균 출원후 경과기간(L_j)	7.58	6.79	6.99

기술부상도	0.312	0.476	0.268
-------	-------	-------	-------

- [0060] 각 군집별로 기술 부상도를 계산한 결과 군집2에 해당하는 특허가 포함된 기술 군집이 가장 기술 부상도가 높다는 결과가 도출되었고, 군집2에 해당하는 기술이 부상기술로 예상되었다.
- [0061] 이처럼 본 발명의 일실시예에 따른 부상기술 예측방법은 검색을 통한 특허문서를 활용하여 키워드를 추출하여 키워드 행렬을 생성하고, 다시 이를 가공하여 주성분 행렬을 생성하고, 주성분 행렬을 각종 알고리즘에 적용함으로써 특정 기술 내 부상기술 내지는 공백기술분야가 어느 분야인지 결정할 수 있는 신뢰성 있는 기준을 제시한다.
- [0062] 본 발명의 실시예들은 컴퓨터로 읽을 수 있는 기록매체에 컴퓨터가 읽을 수 있는 프로그램으로서 구현하는 것이 가능하다. 컴퓨터가 읽을 수 있는 기록매체는 컴퓨터 시스템에 의해 읽혀질 수 있는 데이터가 저장되는 모든 종류의 기록 장치를 포함한다. 컴퓨터가 읽을 수 있는 기록매체의 예로는 ROM, RAM, CD-ROM, 자기 테이프, 플로피 디스크, 광데이터 저장장치 등이 있다.
- [0063] 또한, 컴퓨터가 읽을 수 있는 기록매체는 네트워크로 연결된 컴퓨터 시스템에 분산되어, 분산방식으로 컴퓨터가 읽을 수 있는 프로그램이 저장되고 실행될 수 있다.
- [0064] 그리고, 본 발명을 구현하기 위한 기능적인(functional) 프로그램, 코드 및 코드 세그먼트들은 본 발명이 속하는 기술분야의 프로그래머들에 의해 용이하게 추론될 수 있다.
- [0065] 본 발명의 일실시예에 따른 부상기술 예측장치는 다수의 특허문서들에서 텍스트 마이닝하여 키워드를 도출하는 키워드 도출부, 키워드 도출부에서 도출된 키워드와 특허문서들을 매칭시켜 키워드 행렬을 생성하고, 키워드 행렬에 특이값 분해를 수행하며, 주성분 분석(PCA)을 통해 주성분 행렬을 생성하는 행렬 생성부, 다수의 특허문서들을 분류하기 위한 군집 개수를 결정하고, 다수의 특허문서들 각각을 다수의 군집들 중 하나로 군집화하는 클러스터링부 및 다수의 군집들 중 부상기술에 대응하는 군집을 결정하는 군집결정부를 포함한다.
- [0066] 본 발명의 일실시예에 따른 부상기술 예측장치에 대해서 도면을 통해 상세히 설명한다. 앞서 설명과 중복되는 구성에 대한 설명은 생략한다.
- [0067] 도7은 본 발명의 일실시예에 따른 부상기술 예측장치의 구성도이다.
- [0068] 키워드 도출부(10)는 특정 기술분야에서 검색을 통해 획득된 다수의 특허문서들에서 텍스트를 추출하여 키워드를 도출한다. 텍스트 추출, 즉 텍스트 마이닝은 전자출원형식의 파일 또는 PDF와 같이 파일 형식으로 이루어져 있는 비정형 데이터에 대하여 자연언어처리 기술과 문서 처리기술을 적용하여 유용한 정보를 추출, 가공하는 것이다.
- [0069] 행렬 생성부(20)는 도출된 키워드를 입력받아 특허문서와 매칭시켜 키워드 행렬을 생성한다. 키워드 행렬은 앞서 설명한 수학적식 4와 같다. 또한 행렬 생성부(20)는 키워드 행렬에 특이값 분해(SVD)와 주성분 분석(PCA)을 통해 주성분 행렬을 생성한다. 주성분 행렬은 앞서 설명한 수학적식 5와 같다.
- [0070] 클러스터링부(30)는 군집 개수를 결정한다. 군집 개수를 결정하기 위해 BIC(Bayesian Information Criterion), AIC(Akaike Information Criterion) 또는 Silhouette측도를 활용한다. 또한 클러스터링부(30)는 특허 문서들 각각에 대해 군집화를 수행한다. 군집화를 위해 군집화 알고리즘인 k-평균 알고리즘(k-mean algorithm) 또는 k-메도이드 알고리즘(k-medoid algorithm)을 활용한다.
- [0071] 군집결정부(40)는 행렬 생성부(20)에서 생성된 주성분 행렬을 활용하여 클러스터링부(30)에서 특허문서를 분류하는 군집화가 완료된 이후 각 군집에 포함되어 있는 특허문헌의 평균 패밀리 특허규모(F_j), 평균 피인용수(C_j) 및 출원후 평균 경과기간(L_j)을 기초로 다수의 군집들 중 부상기술에 대응하는 군집을 결정하기 위한 기술 부상도를 연산한다. 기술 부상도를 연산하기 위한 식은 앞서 설명한 수학적식 7과 같다.
- [0072] 결국 특허 문서를 이용하여 부상 기술을 예측하는 본 발명의 일실시예에 따른 부상기술 예측장치를 통해 부상기술에 대한 신뢰성 있는 예측이 가능하고, 비교적 간단한 프로세스를 통해 부상기술에 대한 예측정보 획득이 가능하다.
- [0073] 본 발명은 도면에 도시된 실시 예를 참고로 설명되었으나 이는 예시적인 것에 불과하며, 본 기술 분야의 통상의 지식을 가진 자라면 이로부터 다양한 변형 및 균등한 타 실시 예가 가능하다는 점을 이해할 것이다. 따라서, 본

발명의 진정한 기술적 보호 범위는 첨부된 등록청구범위의 기술적 사상에 의해 정해져야 할 것이다.

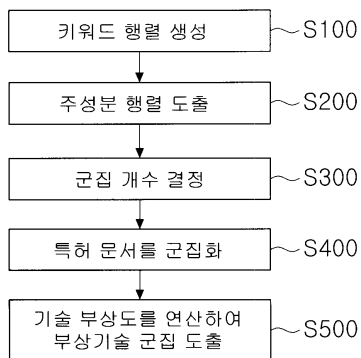
부호의 설명

[0074]

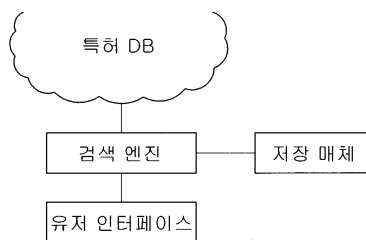
- | | |
|-------------|------------|
| 1 부상기술 예측장치 | 10 키워드 도출부 |
| 20 행렬 생성부 | 30 클러스터링부 |
| 40 군집결정부 | |

도면

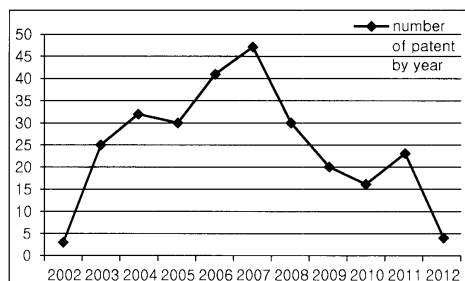
도면1



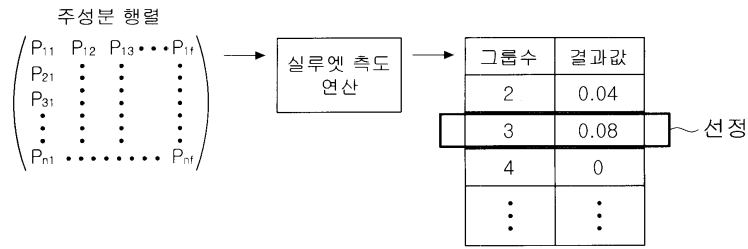
도면2



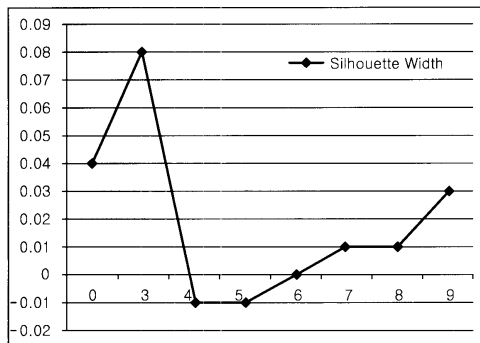
도면3



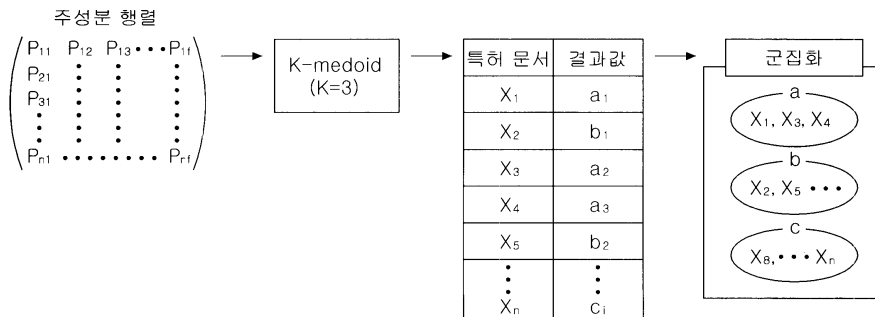
도면4



도면5



도면6



도면7

