(19) 대한민국특허청(KR)
(12) 등록특허공보(B1)

(45) 공고일자 2016년03월08일
(11) 등록번호 10-1600870
(24) 등록일자 2016년03월02일

(51) 국제특허분류(Int. Cl.)
G06F 17/30 (2006.01) *G06F 17/21* (2006.01)
(21) 출원번호 10-2014-0114029
(22) 출원일자 2014년08월29일
심사청구일자 2014년08월29일
(56) 선행기술조사문헌
KR100893129 B1
"특허분석을 이용한 지능형 시스템의 기술 예측", 전성해, 한국지능시스템학회 논문지 2011, Vol.21, No.1, pp.100-105, 2011.02.28*
"특허분석을 위한 빅 데이터학습", 전성해, 한국지능시스템학회 논문지 제23권 제5호, pp.406-411, 2013.10.30*
*는 심사관에 의하여 인용된 문헌

(73) 특허권자
고려대학교 산학협력단
서울특별시 성북구 안암로 145, 고려대학교 (안암동5가)
(72) 발명자
박상성
서울특별시 동대문구 한천로58길 107 104동 2105호(이문동, 현대아파트)
장동식
서울특별시 성북구 안암로9길 30 2동 1102호(안암동3가, 삼익아파트)
(뒷면에 계속)
(74) 대리인
특허법인충현

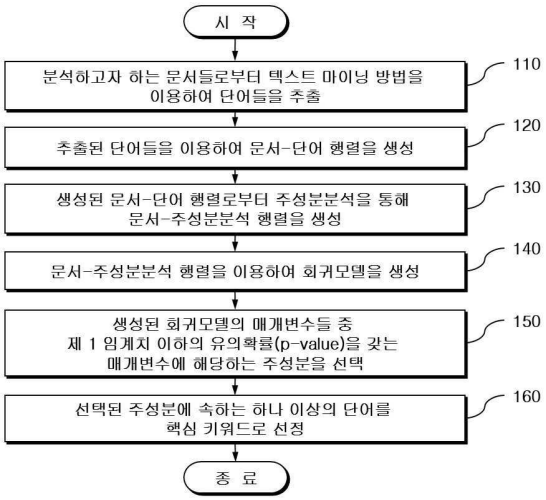
전체 청구항 수 : 총 5 항 심사관 : 김병균

(54) 발명의 명칭 **통계적 방법을 이용한 핵심 키워드 추출방법**

(57) 요 약

본 발명은 핵심 키워드를 추출하는 방법에 관한 것으로서, 분석하고자 하는 문서들로부터 텍스트 마이닝 방법을 이용하여 단어들을 추출하는 단계, 추출된 단어들을 이용하여 문서-단어 행렬을 생성하는 단계, 생성된 문서-단어 행렬로부터 주성분분석을 통해 문서-주성분분석 행렬을 생성하는 단계, 문서-주성분분석 행렬을 이용하여 회귀모델을 생성하는 단계, 생성된 회귀모델의 매개변수들 중 제 1 임계치 이하의 유의확률(p-value)을 갖는 매개변수에 해당하는 주성분을 선택하는 단계, 및 선택된 주성분에 속하는 하나 이상의 단어를 핵심 키워드로 선정하는 단계를 포함함으로써 핵심 키워드를 통계적 방법에 근거하여 객관적으로 추출할 수 있다.

대 표 도 - 도1



(72) 발명자

이준석

경기도 김포시 풍무로 2 121동1502호(풍무동, 유현마을신동아아파트)

김갑조

서울특별시 성북구 안암로 145 고려대학교 산업경영공학과 (안암동5가)

이준혁

서울특별시 성북구 안암로 145 고려대학교 산업경영공학과 (안암동5가)

김종찬

서울특별시 성북구 안암로 145 고려대학교 산업경영공학과 (안암동5가)

이 발명을 지원한 국가연구개발사업

과제고유번호 2013053572
부처명 교육부
연구관리전문기관 한국연구재단
연구사업명 일반연구자지원(교육부)
연구과제명 시맨틱 기술 기반 IP 마이닝 시스템 개발
기 여 율 1/2
주관기관 고려대학교 산학협력단
연구기간 2013.09.01 ~ 2014.08.31

이 발명을 지원한 국가연구개발사업

과제고유번호 22A20130012646
부처명 교육부
연구관리전문기관 한국연구재단
연구사업명 제조·물류분야에서의 빅데이터 운용사업팀
연구과제명 BK21PLUS
기 여 율 1/2
주관기관 고려대학교 산학협력단
연구기간 2014.03.01 ~ 2015.02.28

명 세 서

청구범위

청구항 1

핵심 키워드를 추출하는 방법에 있어서,

분석하고자 하는 문서들로부터 텍스트 마이닝 방법을 이용하여 단어들을 추출하는 단계;

상기 추출된 단어들을 이용하여 문서-단어 행렬을 생성하는 단계;

상기 생성된 문서-단어 행렬로부터 주성분분석을 통해 문서-주성분분석 행렬을 생성하는 단계;

상기 주성분분석 결과, 제 2 임계치 이상의 고유 값을 갖는 주성분만을 이용하여 축소된 문서-주성분분석 행렬을 생성하는 단계;

상기 축소된 문서-주성분분석 행렬을 이용하여 회귀모델을 생성하는 단계;

상기 생성된 회귀모델의 매개변수들 중 제 1 임계치 이하의 유의확률(p-value)을 갖는 매개변수에 해당하는 주성분을 선택하는 단계; 및

상기 선택된 주성분에 속하는 하나 이상의 단어를 핵심 키워드로 선정하는 단계를 포함하는 방법.

청구항 2

삭제

청구항 3

제 1 항에 있어서,

상기 선택된 주성분에 속하는 하나 이상의 단어를 핵심 키워드로 선정하는 단계는,

상기 주성분에 속하는 단어들 중 제 3 임계치 이상의 주성분 점수를 갖는 단어 또는 주성분 점수가 높은 순서에 따라 소정의 단어를 핵심 키워드로 선정하는 것을 특징으로 하는 방법.

청구항 4

제 1 항에 있어서,

상기 주성분분석은 상관행렬을 이용하여 수행되는 것을 특징으로 하는 방법.

청구항 5

제 1 항에 있어서,

상기 분석하고자 하는 문서들은 특허문서이고,

상기 단어를 추출하는 단계는,

특허문서의 발명의 명칭, 요약, 특허청구범위, 또는 발명의 상세한 설명 중 하나 이상의 부분에서 단어들을 추출하는 것을 특징으로 하는 방법.

청구항 6

제 1 항에 있어서,

상기 단어들을 추출하는 단계는,

문장분석(parsing) 및 말뭉치분석(corpus)를 이용하여 수행되는 것을 특징으로 하는 방법.

발명의 설명

기 술 분 야

[0001] 본 발명은 핵심 키워드를 추출하는 방법에 관한 것으로서, 보다 구체적으로, 회귀모델을 이용하여 데이터들의 특징을 대표하는 핵심 키워드를 통계적 방법에 근거하여 객관적으로 추출하는 핵심 키워드를 추출하는 방법에 관한 것이다.

배 경 기 술

[0002] 근래에 지식재산에 대한 중요성은 점점 더 부각되고 있으며 여러 기업들과 국가는 지식재산을 보호하고 활용하기 위해 많은 노력을 하고 있다. 또한 이러한 지식재산을 활용할 수 있는 방법을 연구하여 기업의 전략 또는 국가의 전략을 수립하는데 많이 반영하고 있는 추세이다. 하지만 기존의 방법들은 전문가들의 의견을 바탕으로 하는 정성적인 방법이 많이 사용되었다. 그 결과, 전문가들의 사회적 위치, 수준 등으로 인해 예측의 객관성을 확보하기가 어려웠고 전문가마다 의견이 다를 수 있어 기업 혹은 국가의 미래전략을 수립하는데 있어 많은 어려움을 야기하였다. 이러한 정성적인 분석은 많은 시간이 걸린다는 단점이 존재하였다. 정성적 분석에서 발생되는 여러 단점을 보완하고 사용자에게 보다 객관적인 정보를 제공하기 위해 정량적 방법이 많이 연구되고 있다. 정량적 분석방법은 여러 자료를 바탕으로 사용자에게 객관적인 정보를 제공할 수 있으며 수집된 자료를 분석도구를 통하여 분석하므로 정성적 방법으로 분석할 때보다 많은 시간을 절약할 수 있다. 또한 객관적인 정보를 전달하는 장점으로 인해 전문가가 아닌 비전문가 역시 손쉽게 원하는 정보를 파악할 수 있다.

선행기술문헌

특허문헌

[0003] (특허문헌 0001) 한국공개특허 "조사 대상 문서의 문서 특징 분석 장치(10-2006-0095565)"

발명의 내용

해결하려는 과제

[0004] 본 발명이 해결하고자 하는 과제는 핵심 키워드를 통계적 방법에 근거하여 객관적으로 추출하는 핵심 키워드를 추출하는 방법을 제공하는 것이다.

과제의 해결 수단

[0005] 본 발명은 상기 과제를 해결하기 위하여, 핵심 키워드를 추출하는 방법에 있어서, 분석하고자 하는 문서들로부터 텍스트 마이닝 방법을 이용하여 단어들을 추출하는 단계; 상기 추출된 단어들을 이용하여 문서-단어 행렬을 생성하는 단계; 상기 생성된 문서-단어 행렬로부터 주성분분석을 통해 문서-주성분분석 행렬을 생성하는 단계; 상기 문서-주성분분석 행렬을 이용하여 회귀모델을 생성하는 단계; 상기 생성된 회귀모델의 매개변수들 중 제 1 임계치 이하의 유의확률(p-value)을 갖는 매개변수에 해당하는 주성분을 선택하는 단계; 및 상기 선택된 주성분에 속하는 하나 이상의 단어를 핵심 키워드로 선정하는 단계를 포함하는 방법을 제공한다.

[0006] 본 발명의 다른 실시예에 의하면, 상기 주성분분석 결과, 제 2 임계치 이상의 고유 값을 갖는 주성분만을 이용하여 축소된 문서-주성분분석 행렬을 생성하는 단계를 더 포함하는 방법일 수 있다.

[0007] 본 발명의 다른 실시예에 의하면, 상기 선택된 주성분에 속하는 하나 이상의 단어를 핵심 키워드로 선정하는 단계는, 상기 주성분에 속하는 단어들 중 제 3 임계치 이상의 주성분 점수를 갖는 단어 또는 주성분 점수가 높은 순서에 따라 소정의 단어를 핵심 키워드로 선정하는 것을 특징으로 하는 방법일 수 있다.

[0008] 본 발명의 다른 실시예에 의하면, 상기 주성분분석은 상관행렬을 이용하여 수행되는 것을 특징으로 하는 방법일 수 있다.

[0009] 본 발명의 다른 실시예에 의하면, 상기 분석하고자 하는 문서들은 특허문서이고, 상기 단어를 추출하는 단계는, 특허문서의 발명의 명칭, 요약, 특허청구범위, 또는 발명의 상세한 설명 중 하나 이상의 부분에서 단어들을 추

출하는 것을 특징으로 하는 방법일 수 있다.

[0010] 본 발명의 다른 실시예에 의하면, 상기 단어들을 추출하는 단계는, 문장분석(parsing) 및 말뭉치분석(corpus)를 이용하여 수행되는 것을 특징으로 하는 방법일 수 있다.

발명의 효과

[0011] 본 발명에 따르면, 핵심키워드 도출을 위하여 기존의 전문가의견, 델파이기법 등 정성적으로 추출하던 방법을 벗어나, 주성분분석과 회귀모델을 통해 정량적으로 핵심키워드를 추출할 수 있다. 이에 따라 객관적인 특허 분석이 가능하다.

도면의 간단한 설명

[0012] 도 1은 본 발명의 일 실시예에 따른 핵심 키워드를 추출하는 방법의 흐름도이다.

도 2는 본 발명의 다른 실시예에 따른 핵심 키워드를 추출하는 방법의 흐름도이다.

도 3은 본 발명의 또 다른 실시예에 따른 핵심 키워드를 추출하는 방법의 흐름도이다.

발명을 실시하기 위한 구체적인 내용

[0013] 본 발명에 관한 구체적인 내용의 설명에 앞서 이해의 편의를 위해 본 발명이 해결하고자 하는 과제의 해결 방안의 개요 혹은 기술적 사상의 핵심을 우선 제시한다.

[0014] 본 발명의 일 실시예에 따른 핵심 키워드를 추출하는 방법은, 분석하고자 하는 문서들로부터 텍스트 마이닝 방법을 이용하여 단어들을 추출하는 단계, 상기 추출된 단어들을 이용하여 문서-단어 행렬을 생성하는 단계, 상기 생성된 문서-단어 행렬로부터 주성분분석을 통해 문서-주성분분석 행렬을 생성하는 단계, 상기 문서-주성분분석 행렬을 이용하여 회귀모델을 생성하는 단계, 상기 생성된 회귀모델의 매개변수들 중 제 1 임계치 이하의 유의확률(p-value)을 갖는 매개변수에 해당하는 주성분을 선택하는 단계, 및 상기 선택된 주성분에 속하는 하나 이상의 단어를 핵심 키워드로 선정하는 단계를 포함한다.

[0015] 이하 첨부된 도면을 참조하여 본 발명이 속하는 기술 분야에서 통상의 지식을 가진 자가 본 발명을 용이하게 실시할 수 있는 실시 예를 상세히 설명한다. 그러나 이들 실시예는 본 발명을 보다 구체적으로 설명하기 위한 것으로, 본 발명의 범위가 이에 의하여 제한되지 않는다는 것은 당업계의 통상의 지식을 가진 자에게 자명할 것이다.

[0016] 본 발명이 해결하고자 하는 과제의 해결 방안을 명확하게 하기 위한 발명의 구성을 본 발명의 바람직한 실시예에 근거하여 첨부 도면을 참조하여 상세히 설명하되, 당해 도면에 대한 설명시 필요한 경우 다른 도면의 구성요소를 인용할 수 있음을 미리 밝혀둔다. 아울러 본 발명의 바람직한 실시 예에 대한 동작 원리를 상세하게 설명함에 있어 본 발명과 관련된 공지 기능 혹은 구성에 대한 구체적인 설명 그리고 그 이외의 제반 사항이 본 발명의 요지를 불필요하게 흐릴 수 있다고 판단되는 경우, 그 상세한 설명을 생략한다.

[0017] 도 1은 본 발명의 일 실시예에 따른 핵심 키워드를 추출하는 방법의 흐름도이다.

[0018] 110 단계는 분석하고자 하는 문서들로부터 텍스트 마이닝 방법을 이용하여 단어들을 추출하는 단계이다.

[0019] 보다 구체적으로, 분석하고자 하는 문서들로부터 텍스트 마이닝 방법을 이용하여 단어들을 추출한다. 텍스트 마이닝(Text Mining) 방법은 자연어로 구성된 비정형 텍스트 데이터에서 패턴 또는 관계를 추출하여 가치와 의미 있는 정보를 찾아내는 마이닝 기법으로, 사람들이 말하는 언어를 이해할 수 있는 자연어처리 기술에 기반한 기술이다. 상기 단어들은 문장분석(parsing) 및 말뭉치분석(corpus)를 이용하여 추출할 수 있다.

[0020] 상기 분석하고자 하는 문서들은 텍스트 형식의 문서들로, 특히 특허문서일 수 있다. 특허문서에 대한 단어를 추출함에 있어서, 유의한 단어들을 추출하는 정확성 및 효율성을 위하여, 특허문서의 발명의 명칭, 요약, 특허청구범위, 또는 발명의 상세한 설명 중 하나 이상의 부분에서 단어들을 추출할 수 있다. 특허문서의 발명의 명칭, 요약, 특허청구범위에는 해당 특허문서에서 다루고자 하는 발명의 핵심적인 부분이 포함되고 해당 발명의 상세한 내용은 발명의 상세한 설명에 포함되는바, 특허문서의 발명의 명칭, 요약, 특허청구범위, 또는 발명의 상세한 설명 중 하나 이상의 부분에서 단어들을 추출할 수 있다.

[0021] 120 단계는 상기 추출된 단어들을 이용하여 문서-단어 행렬을 생성하는 단계이다.

[0022] 보다 구체적으로, 텍스트 마이닝 방법을 통해 추출된 단어를 이용하여 분석한 문서와 추출된 단어들을 이용하여 문서-단어 행렬을 생성한다. 문서-단어 행렬은 문서와 단어간의 관계를 나타낸 행렬로, 문서에 각 단어가 나타나는 빈도수를 나타낸다. 상기 문서-단어 행렬은 다음과 같이 나타낼 수 있다.

표 1

[0023]

	Term 1	Term 2	Term 3	…	Term N
Document 1	0	0	1	…	0
Document 2	1	0	0	…	0
Document 3	0	0	0	…	1
Document 4	1	1	0	…	0
Document 5	1	0	0	…	1

[0024] 상기와 같이, 각 단어가 각 문서에 나타나는 빈도수로 Document 1에 Term 1이 출현된 횟수는 0이다.

[0025] 130 단계는 상기 생성된 문서-단어 행렬로부터 주성분분석을 통해 문서-주성분분석 행렬을 생성하는 단계이다.

[0026] 보다 구체적으로, 120 단계에서 추출된 단어들 중 핵심 키워드를 선정하는 것은 매우 어려우며, 선정된 키워드에 따라 이후 결과에 지대한 영향을 미치기 때문에 섣불리 단어의 개수를 줄일 수 없다. 객관적이고 핵심 키워드를 추출하기 위하여, 주성분 분석(Principal component analysis, PCA)을 이용하여 행렬의 희소성을 해결하면서 동시에 손실될 수 있는 정보를 가능한 보존할 수 있다. 130 단계에서는 문서-단어 행렬로부터 주성분분석을 통해 문서-주성분분석 행렬을 생성한다.

[0027] 주성분분석은 다양한 변수들에 대해 분석하는 소위 다변량(multivariate) 분석인데, 많은 변수들로부터 몇 개의 주성분들을 추출하는 방법이다. 즉, 주성분분석은 차원축소(dimension reduction)를 위한 것으로, 여기서 주성분이라는 것은 많은 변수들을 설명하는 주된(principal) 성분이라는 의미이다. 상기 주성분분석은 상관행렬을 이용하여 수행될 수 있다. 또는 공분산행렬을 이용하여 수행될 수도 있다.

[0028] 문서-단어 행렬로부터 주성분분석을 통해 생성되는 문서-주성분분석 행렬은 다음과 같이 나타낼 수 있다.

표 2

[0029]

	PC1	PC2	PC3	…	PC *n*
Document 1	0.0024	0.01358	-0.00213	…	0.17852
Document 2	0.001	-0.002	-0.01018	…	0.02364
Document 3	0.00005	0.36778	0.03258	…	0.00025
Document 4	-0.0278	0.03	0.0089	…	-0.05987
Document 5	0.5249	-0.0125	-0.1257	…	0.02113

[0030] 주성분분석을 통해 생성되는 문서-주성분분석 행렬은 주성분(PC)의 개수는 처음 변수인 단어(Term)의 개수만큼 생성된다.

[0031] 여기서, 효율적인 분석을 위하여, 도 2의 210 단계를 거쳐 PC의 개수를 줄일 수 있다.

[0032] 210 단계는 상기 주성분분석 결과, 제 2 임계치 이상의 고유 값을 갖는 주성분만을 이용하여 축소된 문서-주성분분석 행렬을 생성하는 단계이다.

[0033] 보다 구체적으로, 유의미한 주성분만을 이용하여 분석을 수행하기 위하여, 주성분의 고유 값이 제 2 임계치 이상의 고유 값을 갖는 주성분만을 이용하여 행렬의 열의 수를 줄일 수 있다. 주성분분석을 통해 각 주성분에 대한 고유 값(eigen value)와 각 주성분에서의 키워드별 주성분 점수를 산출할 수 있다. 상기 산출된 각 주성분의 고유 값이 제 2 임계치 이상인 경우의 주성분만을 이용하여 문서-주성분분석 행렬을 축소한다. 상기 제 2 임계치는 1일 수 있다. 즉, 고유 값이 1이상인 주성분만을 이용하여 주성분의 수를 줄여 차원축소가 가능하다. 표 2에 주어진 것과 같이, 주성분의 수가 n 개 존재하였을 때, 고유 값이 제 2 임계치 이상인 주성분이 PC1과 PC2인

경우, n 개에서 2 개로 차원이 축소된 문서-주성분분석 행렬은 다음과 같이 나타낼 수 있다.

표 3

[0034]

	PC1	PC2
Document 1	0.0024	0.01358
Document 2	0.001	-0.002
Document 3	0.00005	0.36778
Document 4	-0.0278	0.03
Document 5	0.5249	-0.0125

[0035] 140 단계는 상기 문서-주성분분석 행렬을 이용하여 회귀모델을 생성하는 단계이다.

[0036] 보다 구체적으로, 문서-주성분분석 행렬을 이용하여 회귀모델을 생성한다. 회귀모델은 하나 또는 그 이상의 독립변수의 종속변수에 대한 영향의 추정을 할 수 있는 통계기법으로 주성분의 통계적 분석이 가능하다. 즉, 회귀모델을 생성하여 선택된 주성분들 중 통계적으로 유의한 주성분을 찾을 수 있다. 주성분분석은 통계적으로 유의한 분석이 아닌바, 회귀모델을 이용함으로써 통계적으로 유의한 분석이 가능하다. 상기 문서-주성분분석 행렬로부터 회귀모델을 생성한다. 회귀모델은 다음과 같이 나타낼 수 있다.

수학식 1

[0037]
$$y = \beta_0 + \beta_1 \chi_1 + \beta_2 \chi_2$$

[0038] 150 단계는 상기 생성된 회귀모델의 매개변수들 중 제 1 임계치 이하의 유의확률(p-value)을 갖는 매개변수에 해당하는 주성분을 선택하는 단계이다.

[0039] 보다 구체적으로, 130 단계에서 생성된 회귀모델에서 통계적으로 유의한 주성분을 선택하기 위하여, 회귀모델의 매개변수를 이용한다. 수학식 1에서 β_1과 β_2는 유의확률(p-value)을 갖는다. 제 1 임계치 이하인 유의확률을 갖는 매개변수에 해당하는 주성분을 유의한 주성분으로 선택할 수 있다. 상기 제 1 임계치는 0.05일 수 있다. 유의확률이 0.05 이하인 매개변수에 해당하는 주성분을 유의한 주성분으로 선택할 수 있다. 만약 여러 개의 단어가 존재한다면 모든 단어들의 중요도를 임의로 정하기 어려우며 각각의 단어가 실제로 영향을 미치는지 여부를 확인하기 어렵다. 따라서, 주성분분석과 회귀모델을 통해 핵심 키워드를 선정함으로써 정확도와 속도 면에서 매우 효율적이다.

[0040] 160 단계는 상기 선택된 주성분에 속하는 하나 이상의 단어를 핵심 키워드로 선정하는 단계이다.

[0041] 보다 구체적으로, 주성분에 포함된 단어들 중 하나 이상의 단어를 핵심 키워드로 선정한다. 주성분에 포함된 단어들 중 핵심 키워드를 선정한다. 주성분에 포함된 단어들이 복수인 경우, 상기 주성분에 속하는 단어들 중 제 3 임계치 이상의 주성분 점수를 갖는 단어 또는 주성분 점수가 높은 순서에 따라 소정의 단어를 핵심 키워드로 선택할 수 있다. 주성분은 복수의 단어를 포함할 수 있고, 각 단어들의 주성분 부하(선형 계수) 즉, 주성분 점수를 이용하여 최종적으로 핵심 키워드를 선정할 수 있다. 상기 주성분 점수가 제 3 임계치 이상의 주성분 점수를 갖는 단어를 핵심 키워드로 선택할 수 있다. 주성분 점수는 -1 내지 1일 수 있고, 상기 제 3 임계치는 0일 수 있다. 또는 주성분 점수가 높은 단어들 순서대로 핵심 키워드를 선정할 수 있다. 미리 설정된 수만큼 선정하거나, 또는 단어 및 문서의 양에 따라 선정되는 단어의 수가 달라질 수 있다. 제 3 임계치와 단어의 수 둘 모두에 따라 선정될 수도 있다. 수학식 1에서 β_1의 유의확률이 0.05 이하인 경우, PC1이 유의한 주성분이고, 상기 주성분에 속한 단어들은 다음과 같을 수 있다.

표 4

[0042]

	PC1
market	0.125
patent	0.86
application	0.52
date	0.001
data	0.68

[0043] 제 3 임계치가 0일 경우, 상기 5 개의 단어가 핵심 키워드로 선정되고, 제 3 임계치가 0.5일 경우, patent, application, data가 핵심 키워드로 선정될 수 있다. 주성분 점수가 가장 높은 단어를 핵심 키워드로 선정하는 경우, patent을 핵심 키워드로 선정될 수 있다.

[0044] 주성분분석(PCA)과 회귀모델을 이용하여 통계적으로 유의한 매개변수β를 찾아내고 선정된 매개변수를 통해 중요한 키워드를 선정하여 정량적 기술동향 분석을 할 수 있다.

[0045] 도 3은 본 발명의 또 다른 실시예에 따른 핵심 키워드를 추출하는 방법의 흐름도이다.

[0046] 특허문서 분석을 위한 핵심 키워드를 추출하기 위하여, 먼저, 특허 문서를 조사한다. 이후, 문서로부터 텍스트 마이닝을 수행한다. parsing과 corpus를 이용하여 단어를 추출하여 문서-단어 행렬(DPTM)을 생성한다. 이후 주성분분석을 이용하여 문서-주성분분석 행렬(DPCM)을 생성한다. 주성분의 수를 줄이기 위하여, 주성분의 고유 값이 1 이상이 주성분을 선택한다. 이를 통해 축소된 문서-주성분분석 행렬을 생성한다. 축소된 문서-주성분분석 행렬을 이용하여 회귀모델을 생성한다. 회귀모델의 매개변수의 유의확률이 유의한지 판단하고, 주성분 점수가 높은 단어들을 핵심 키워드로 선정할 수 있다.

[0047] 본 발명의 실시예는 문서의 단어를 추출하여, 문서-단어 행렬을 생성하고, 주성분분석을 수행하여 문서-주성분분석 행렬을 생성하고, 회귀모델을 생성하여, 최종 핵심 키워드를 선정하는 하나 이상의 처리부(프로세서) 및 처리부에서 산출되는 결과를 저장하는 하나 이상의 저장부(데이터베이스)를 포함할 수 있다.

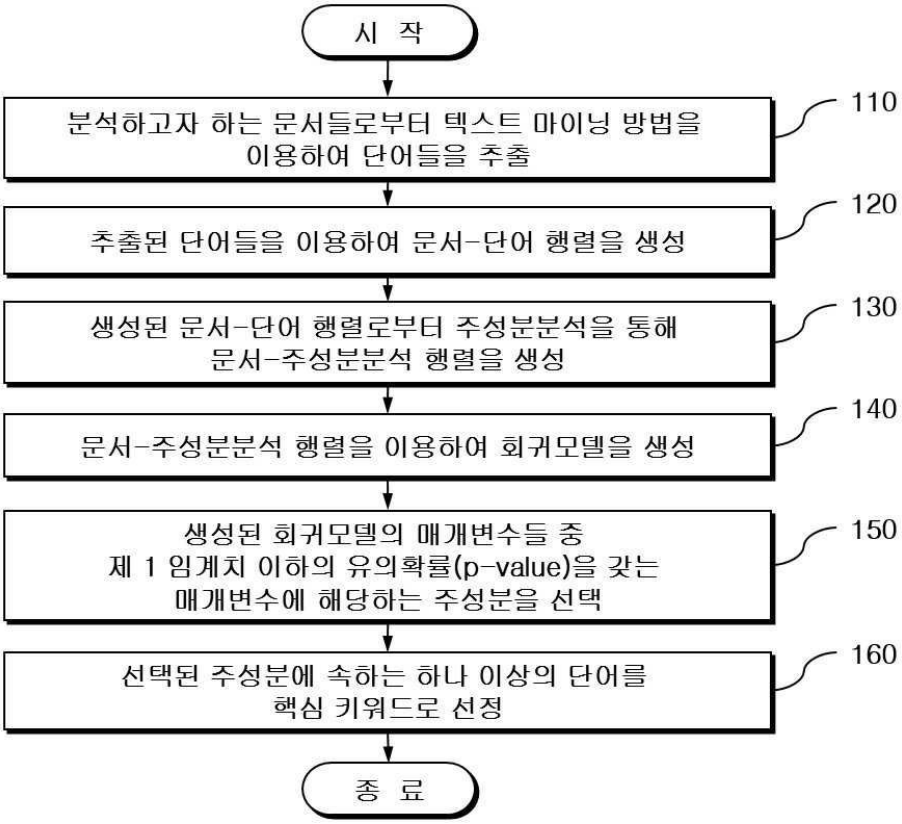
[0048] 본 발명의 실시예들은 다양한 컴퓨터 수단을 통하여 수행될 수 있는 프로그램 명령 형태로 구현되어 컴퓨터 판독 가능 매체에 기록될 수 있다. 상기 컴퓨터 판독 가능 매체는 프로그램 명령, 데이터 파일, 데이터 구조 등을 단독으로 또는 조합하여 포함할 수 있다. 상기 매체에 기록되는 프로그램 명령은 본 발명을 위하여 특별히 설계되고 구성된 것들이거나 컴퓨터 소프트웨어 당업자에게 공지되어 사용 가능한 것일 수도 있다. 컴퓨터 판독 가능 기록 매체의 예에는 하드 디스크, 플로피 디스크 및 자기 테이프와 같은 자기 매체(magnetic media), CD-ROM, DVD와 같은 광기록 매체(optical media), 플롭티컬 디스크(floptical disk)와 같은 자기-광 매체(magneto-optical media), 및 롬(ROM), 램(RAM), 플래시 메모리 등과 같은 프로그램 명령을 저장하고 수행하도록 특별히 구성된 하드웨어 장치가 포함된다. 프로그램 명령의 예에는 컴파일러에 의해 만들어지는 것과 같은 기계어 코드뿐만 아니라 인터프리터 등을 사용해서 컴퓨터에 의해서 실행될 수 있는 고급 언어 코드를 포함한다. 상기된 하드웨어 장치는 본 발명의 동작을 수행하기 위해 하나 이상의 소프트웨어 모듈로서 작동하도록 구성될 수 있으며, 그 역도 마찬가지이다.

[0049] 이상과 같이 본 발명에서는 구체적인 구성 요소 등과 같은 특정 사항들과 한정된 실시예 및 도면에 의해 설명되었으나 이는 본 발명의 보다 전반적인 이해를 돕기 위해서 제공된 것일 뿐, 본 발명은 상기의 실시예에 한정되는 것은 아니며, 본 발명이 속하는 분야에서 통상적인 지식을 가진 자라면 이러한 기재로부터 다양한 수정 및 변형이 가능하다.

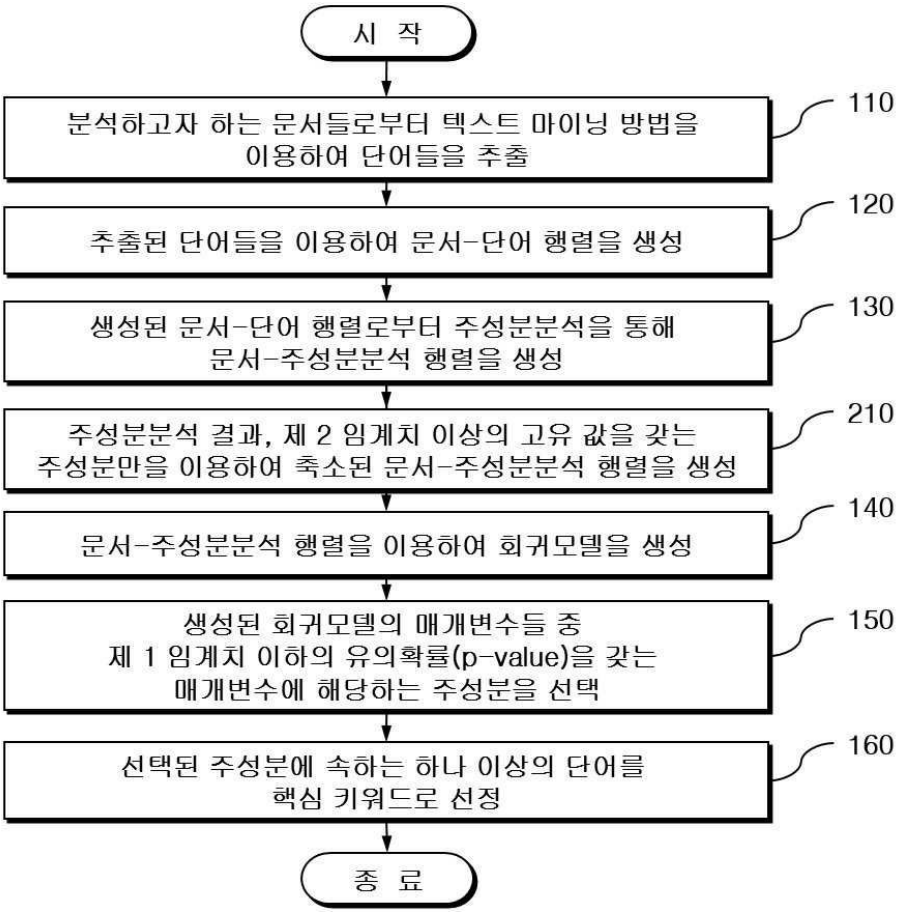
[0050] 따라서, 본 발명의 사상은 설명된 실시예에 국한되어 정해져서는 아니되며, 후술하는 특허청구범위뿐 아니라 이 특허청구범위와 균등하거나 등가적 변형이 있는 모든 것들은 본 발명 사상의 범주에 속한다고 할 것이다.

도면

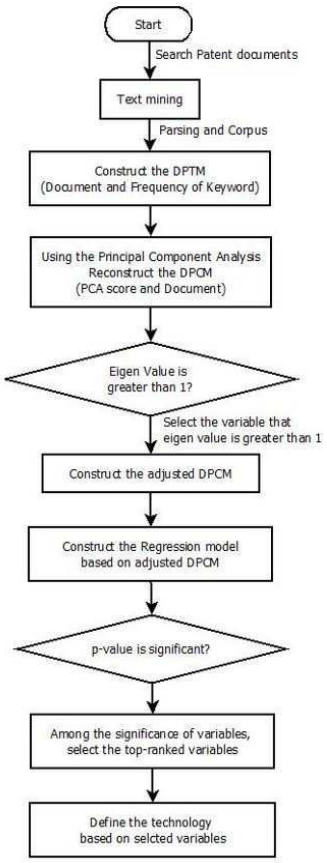
도면1


시 작
분석하고자 하는 문서들로부터 텍스트 마이닝 방법을 이용하여 단어들을 추출
110
추출된 단어들을 이용하여 문서-단어 행렬을 생성
120
생성된 문서-단어 행렬로부터 주성분분석을 통해 문서-주성분분석 행렬을 생성
130
문서-주성분분석 행렬을 이용하여 회귀모델을 생성
140
생성된 회귀모델의 매개변수들 중 제 1 임계치 이하의 유의확률(p-value)을 갖는 매개변수에 해당하는 주성분을 선택
150
선택된 주성분에 속하는 하나 이상의 단어를 핵심 키워드로 선정
160
종 료

도면2


시 작
분석하고자 하는 문서들로부터 텍스트 마이닝 방법을 이용하여 단어들을 추출
110
추출된 단어들을 이용하여 문서-단어 행렬을 생성
120
생성된 문서-단어 행렬로부터 주성분분석을 통해 문서-주성분분석 행렬을 생성
130
주성분분석 결과, 제 2 임계치 이상의 고유 값을 갖는 주성분만을 이용하여 축소된 문서-주성분분석 행렬을 생성
210
문서-주성분분석 행렬을 이용하여 회귀모델을 생성
140
생성된 회귀모델의 매개변수들 중 제 1 임계치 이하의 유의확률(p-value)을 갖는 매개변수에 해당하는 주성분을 선택
150
선택된 주성분에 속하는 하나 이상의 단어를 핵심 키워드로 선정
160
종 료

도면3


Start
Search Patent documents
Text mining
Parsing and Corpus
Construct the DPTM
(Document and Frequency of Keyword)
Using the Principal Component Analysis
Reconstruct the DPCM
(PCA score and Document)
Eigen Value is
greater than 1?
Select the variable that
eigen value is greater than 1
Construct the adjusted DPCM
Construct the Regression model
based on adjusted DPCM
p-value is significant?
Among the significance of variables,
select the top-ranked variables
Define the technology
based on selcted variables