



(19) 대한민국특허청(KR)
(12) 공개특허공보(A)

(11) 공개번호 10-2016-0047112
(43) 공개일자 2016년05월02일

(51) 국제특허분류(Int. Cl.)

G06F 17/30 (2006.01)

(21) 출원번호 10-2014-0143056

(22) 출원일자 2014년10월22일

심사청구일자 2014년10월22일

(71) 출원인

고려대학교 산학협력단

서울특별시 성북구 안암로 145, 고려대학교 (안암동5가)

(72) 발명자

박상성

서울특별시 동대문구 한천로58길 139, 103동 101호 (이문동, 이문동대우아파트)

장동식

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

(뒷면에 계속)

(74) 대리인

특허법인충현

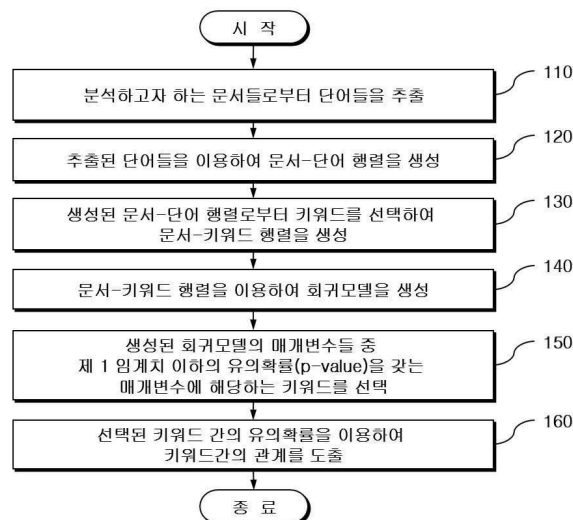
전체 청구항 수 : 총 6 항

(54) 발명의 명칭 통계적 분석에 기반한 기술계층다이어그램 모델을 이용한 특허분석방법

(57) 요약

본 발명은 특허 문서를 이용하여 기술을 분석하는 방법에 관한 것으로서, 분석하고자 하는 문서들로부터 단어들을 추출하는 단계, 상기 추출된 단어들을 이용하여 문서-단어 행렬을 생성하는 단계, 상기 생성된 문서-단어 행렬로부터 키워드를 선택하여 문서-키워드 행렬을 생성하는 단계, 상기 문서-키워드 행렬을 이용하여 회귀모델을 생성하는 단계, 상기 생성된 회귀모델의 매개변수들 중 제 1 임계치 이하의 유의확률(p-value)을 갖는 매개변수에 해당하는 키워드를 선택하는 단계, 및 상기 선택된 키워드 간의 유의확률을 이용하여 키워드 간의 관계를 도출하는 단계를 포함함으로써 통계적 분석을 이용하여 기술계층을 도출하고, 도출된 기술계층을 이용함으로써 객관적이고 정확한 특허 분석이 가능하다.

대표도 - 도1



(72) 발명자

이흥철

서울특별시 서초구 서초대로26길 19, 101동 1202호
(방배동, 브라운스톤방배)

김갑조

서울특별시 성북구 안암로 145 고려대학교 산업경
영공학과

이준혁

서울특별시 성북구 안암로 145 고려대학교 산업경
영공학과

이준석

경기도 김포시 유현로 19, 121동 1502호 (풍무동,
유현마을신동아아파트)

전성해

인천광역시 연수구 해돋이로120번길 16, 202동
1702호 (송도동, 송도풍림아이원2단지아파트)

강지호

경상남도 창원시 마산회원구 내서읍 호원로 359,
208동 2001호 (코오롱하늘채2차아파트)

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 ~ 2014.08.31이 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

명세서

청구범위

청구항 1

특허 문서를 이용하여 기술을 분석하는 방법에 있어서,
 분석하고자 하는 문서들로부터 단어들을 추출하는 단계;
 상기 추출된 단어들을 이용하여 문서-단어 행렬을 생성하는 단계;
 상기 생성된 문서-단어 행렬로부터 키워드를 선택하여 문서-키워드 행렬을 생성하는 단계;
 상기 문서-키워드 행렬을 이용하여 회귀모델을 생성하는 단계;
 상기 생성된 회귀모델의 매개변수들 중 제 1 임계치 이하의 유의확률(p-value)을 갖는 매개변수에 해당하는 키워드를 선택하는 단계; 및
 상기 선택된 키워드 간의 유의확률을 이용하여 키워드 간의 관계를 도출하는 단계를 포함하는 방법.

청구항 2

제 1 항에 있어서,
 상기 키워드 간의 관계를 도출하는 단계는,
 상기 선택된 키워드 중 제 2 임계치 이하의 유의확률을 갖는 매개변수에 해당하는 하나 이상의 키워드를 제 1 키워드로 선정하는 단계; 및
 상기 제 1 키워드와 제 1 키워드 이외의 키워드들 간의 회귀모델 분석을 통해, 제 3 임계치 이하의 유의확률을 갖는 매개변수에 해당하는 키워드를 상기 제 1 키워드와 관계성이 있는 키워드로 선정하는 단계를 포함하는 방법.

청구항 3

제 1 항에 있어서,
 상기 키워드 간의 관계를 도출하는 단계는,
 상기 키워드 간의 유의확률을 이용하여 계층을 나누는 것을 특징으로 하는 방법.

청구항 4

제 1 항에 있어서,
 상기 문서-키워드 행렬을 생성하는 단계는,
 상기 문서-단어 행렬로부터 제 4 임계치 이상의 발생 빈도 값을 갖는 단어들을 선정하는 단계; 및
 상기 선정된 단어들 중 기술과 관련이 없는 단어들을 제거하는 단계를 포함하는 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

기술 분야

- [0001] 본 발명은 특허를 분석하는 방법에 관한 것으로서, 보다 구체적으로, 통계적 분석을 이용하여 기술계층을 도출하고, 도출된 기술계층을 이용하여 특허를 분석하는 방법에 관한 것이다.

배경 기술

- [0002] 최근 기술예측(Technology Forecasting, TF)은 기술경영에서 중요하게 떠오르는 이슈이다. 대부분 MOT R&D정책은 다양한 기술예측의 결과에 의존한다. 전통적인 기술예측은 델파이 방법과 같은 정성적인 접근법을 이용하였다. 델파이 방법이 기술예측에서 대표적인 방법이었던, 결과가 주관적이라는 단점으로 인해 적합한 방법은 아니다.

- [0003] 이러한 문제점으로 인해 최근 대부분 기술예측 기법들은 특허분석과 같은 정량적인 접근법에 초점을 맞추어져 가고 있는 추세이다. 특허는 개발된 기술들에 대한 다양한 정보를 가지고 있으며, 기술예측에서 사용되는 객관적인 자료 형식 중 하나이다.

선행기술문헌

특허문헌

- [0004] (특허문헌 0001) 한국공개특허 "조사 대상 문서의 문서 특징 분석 장치(10-2006-0095565)"

발명의 내용

해결하려는 과제

- [0005] 본 발명이 해결하고자 하는 과제는 통계적 분석을 이용하여 기술계층을 도출하고, 도출된 기술계층을 이용하여 특허를 분석하는 방법을 제공하는 것이다.

과제의 해결 수단

- [0006] 본 발명은 상기 과제를 해결하기 위하여, 특허 문서를 이용하여 기술을 분석하는 방법에 있어서, 분석하고자 하는 문서들로부터 단어들을 추출하는 단계; 상기 추출된 단어들을 이용하여 문서-단어 행렬을 생성하는 단계; 상기 생성된 문서-단어 행렬로부터 키워드를 선택하여 문서-키워드 행렬을 생성하는 단계; 상기 문서-키워드 행렬을 이용하여 회귀모델을 생성하는 단계; 상기 생성된 회귀모델의 매개변수들 중 제 1 임계치 이하의 유의확률(p-value)을 갖는 매개변수에 해당하는 키워드를 선택하는 단계; 및 상기 선택된 키워드 간의 유의확률을 이용하여 키워드 간의 관계를 도출하는 단계를 포함하는 방법을 제공한다.

- [0007] 본 발명의 다른 실시예에 의하면, 상기 키워드 간의 관계를 도출하는 단계는, 상기 선택된 키워드 중 제 2 임계치 이하의 유의확률을 갖는 매개변수에 해당하는 하나 이상의 키워드를 제 1 키워드로 선정하는 단계; 및 상기 제 1 키워드와 제 1 키워드 이외의 키워드들 간의 회귀모델 분석을 통해, 제 3 임계치 이하의 유의확률을 갖는 매개변수에 해당하는 키워드를 상기 제 1 키워드와 관계성이 있는 키워드로 선정하는 단계를 포함하는 방법일 수 있다.

- [0008] 본 발명의 다른 실시예에 의하면, 상기 키워드 간의 관계를 도출하는 단계는, 상기 키워드 간의 유의확률을 이

용하여 계층을 나누는 것을 특징으로 하는 방법일 수 있다.

[0009] 본 발명의 다른 실시예에 의하면, 상기 문서-키워드 행렬을 생성하는 단계는, 상기 문서-단어 행렬로부터 제 4 임계치 이상의 발생 빈도 값을 갖는 단어들을 선정하는 단계; 및 상기 선정된 단어들 중 기술과 관련이 없는 단어들을 제거하는 단계를 포함하는 방법일 수 있다.

[0010] 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 추출하는 것을 특징으로 하는 방법일 수 있다.

[0011] 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

[0012] 본 발명에 따르면, 핵심키워드 도출을 위하여 기존의 전문가의견, 델파이기법 등 정성적으로 추출하던 방법을 벗어나, 통계적 분석을 이용하여 기술계층을 도출하고, 도출된 기술계층을 이용하여 특허를 분석할 수 있다. 이에 따라 객관적인 특허 분석이 가능하다.

도면의 간단한 설명

[0013] 도 1은 본 발명의 일 실시예에 따른 특허분석방법의 흐름도이다.

도 2 내지 3은 본 발명의 다른 실시예에 따른 특허분석방법의 흐름도이다.

도 4는 본 발명의 또 다른 실시예에 따른 특허분석방법의 흐름도이다.

도 5 내지 6은 본 발명의 실시예에 따른 특허분석방법에 따라 도출된 기술계층을 나타낸 것이다.

발명을 실시하기 위한 구체적인 내용

[0014] 본 발명에 관한 구체적인 내용의 설명에 앞서 이해의 편의를 위해 본 발명이 해결하고자 하는 과제의 해결 방안의 개요 혹은 기술적 사상의 핵심을 우선 제시한다.

[0015] 본 발명의 일 실시예에 따른 특허 문서를 이용하여 기술을 분석하는 방법에 있어서, 분석하고자 하는 문서들로부터 단어들을 추출하는 단계, 상기 추출된 단어들을 이용하여 문서-단어 행렬을 생성하는 단계, 상기 생성된 문서-단어 행렬로부터 키워드를 선택하여 문서-키워드 행렬을 생성하는 단계, 상기 문서-키워드 행렬을 이용하여 회귀모델을 생성하는 단계, 상기 생성된 회귀모델의 매개변수들 중 제 1 임계치 이하의 유의확률(p-value)을 갖는 매개변수에 해당하는 키워드를 선택하는 단계, 및 상기 선택된 키워드 간의 유의확률을 이용하여 키워드 간의 관계를 도출하는 단계를 포함한다.

[0016] 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.

[0017] 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 실시예에 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.

[0018] 도 1은 본 발명의 일 실시예에 따른 특허분석방법의 흐름도이다.

[0019] 본 발명의 일 실시예에 따른 특허 문서를 이용하여 기술을 분석하는 방법은 다음의 일련의 과정으로 통해 구현된다.

[0020] 110 단계는 분석하고자 하는 문서들로부터 단어들을 추출하는 단계이다.

- [0021] 보다 구체적으로, 분석하고자 하는 문서들로부터 텍스트 마이닝 방법 등을 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.
- [0022] 상기 문서들을 분석하고자 하는 기술 또는 기술분야에 대해 조사된 문서들일 수 있다.
- [0023] 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. 상기 특허문서 중, 특허의 요약(주록)을 이용하는 것이 바람직하다.
- [0024] 120 단계는 상기 추출된 단어들을 이용하여 문서-단어 행렬을 생성하는 단계이다.
- [0025] 보다 구체적으로, 분석한 문서와 추출된 단어들을 이용하여 문서-단어 행렬(Document-Term Matrix, DTM)을 생성한다. Corpus와 텍스트 마이닝 기법의 text repository를 사용하여 문서-단어 행렬을 구성한다. 문서-단어 행렬은 문서와 단어 간의 관계를 나타낸 행렬로, 문서에 각 단어가 나타나는 빈도수를 나타낸다. 행렬의 행과 열은 단어와 문서들로 구성된다. 각 원소는 각 문서에서 단어 발생 빈도 값이다.
- [0026] 130 단계는, 상기 생성된 문서-단어 행렬로부터 키워드를 선택하여 문서-키워드 행렬을 생성하는 단계이다.
- [0027] 보다 구체적으로, 상기 단어들 중 키워드를 선택하여 문서-키워드 행렬을 생성한다. 단어들 중 키워드를 선택함으로써 행의 수를 줄여 빠르고 정확한 분석을 가능하도록 한다.
- [0028] 문서-키워드 행렬을 생성함에 있어서, 도 3의 310 단계 내지 320 단계를 수행할 수 있다. 310 단계에서 문서-단어 행렬로부터 제 4 임계치 이상의 발생 빈도 값을 갖는 단어들을 선정할 수 있다. 상기 제 4 임계치는 미리 설정되어 있거나, 문서-단어 행렬의 행의 수에 따라 결정될 수 있다. 발생 빈도가 높을수록 키워드일 확률이 높은바, 발생 빈도 값에 따라 단어를 선정할 수 있다. 320 단계에서 선정된 단어들 중 기술과 관련이 없는 단어들을 제거하여 문서-키워드 행렬을 생성할 수 있다. 발생 빈도는 높으나, 기술과 관련이 없는 단어들을 제거함으로써 정확성을 높일 수 있다. 예를 들어, "is", "the"와 같이 의미 없는 단어들을 제거할 수 있다. 제거되는 단어들을 미리 설정되어 있거나 사용자의 설정 입력에 따라 특정 단어들이 제거될 수 있다.
- [0029] 140 단계는 상기 문서-키워드 행렬을 이용하여 회귀모델을 생성하는 단계이다.
- [0030] 보다 구체적으로, 문서-키워드 행렬을 이용하여 회귀모델을 생성한다. 회귀모델은 하나 또는 그 이상의 독립변수의 종속변수에 대한 영향의 추정을 할 수 있는 통계기법으로 키워드의 통계적 분석이 가능하다. 즉, 회귀모델을 생성하여 선택된 키워드들 중 통계적으로 유의한 키워드를 찾을 수 있다. 키워드 선택은 통계적으로 유의한 분석이 아닌바, 회귀모델을 이용함으로써 통계적으로 유의한 분석이 가능하다. 상기 문서-키워드 행렬로부터 회귀모델을 생성한다. 회귀모델은 다음과 같이 나타낼 수 있다.

수학식 1

$$Z = \beta_0 + \beta_1 X_1 + \dots + \beta_k X_k + \epsilon$$

- [0031]
- [0032] Z는 종속변수이고, X_s 는 독립변수이다. 종속변수는 영향을 받는 기술이고, 독립변수는 개발된 기술들을 의미한다. 즉, 종속변수는 분석을 하고자 하는 기술이고, 독립변수는 해당 기술에 관련하여 개발된 기술들로 문서분석을 통해 도출되는 키워드들에 해당한다. ϵ 는 오차항이다. 회귀 매개변수 β_k 는 두 기술들 사이의 인과관계의 강도를 나타낸다. Full regression model에서는 아래와 같이, 모든 독립변수들을 사용하고 가설을 세운다.

수학식 2

$$H_0: \beta_i = 0 \text{ vs. } H_1: \beta_i \neq 0, \quad i = 1, 2, \dots, k$$

Z 기술에 영향을 미치는 X 기술들을 찾기 위해, 가설검정을 실시할 수 있다. 귀무가설 H_0 는 i 번째 회귀 매개변수가 0이라는 것을 나타낸다. 이 의미는 i 번째 기술은 기술 Z에 영향을 미치지 않는다는 것이다. 또한 대립가설 H_1 은 i 번째 회귀변수가 0이 아니라는 것을 나타낸다. 즉, 기술 Z는 기술 X_i 에 의존하고 있다는 것이다. 기술 X_i 가 Z에 유의적인 영향을 미치는지 판단하기 위해, 통계적 검정을 통하여 H_0 를 거절할 수 있다. 가설검정을 위해 자유도가 $n-(k+1)$ 인 t분포를 사용한다. n과 k는 데이터 크기와 변수의 개수이다. 만약 H_0 가 참이라면, 아래와 같이 검정통계량을 계산할 수 있다.

수학식 3

$$t = \frac{\hat{\beta}}{S_{\hat{\beta}}}$$

$\hat{\beta}_i$ 는 β_i 의 추정치이고, $S_{\hat{\beta}_i}$ 는 $\hat{\beta}_i$ 의 표준오차이다. H_0 를 거절하기 위해서는 다음조건을 만족시켜야 한다.

수학식 4

$$|t| > t_{(n-(k+1)), \alpha/2}$$

α 는 유의수준이고, 또한 t 분포표를 이용하여 $t_{(n-(k+1)), \alpha/2}$ 의 값을 얻을 수 있다.

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

보다 구체적으로, 빠르고 정확한 분석을 위하여, 유의한 키워드를 선택하여 키워드의 수를 줄일 수 있다. 이를 위하여, 140 단계에서 생성된 회귀모델의 매개변수들 중 제 1 임계치 이하의 유의확률(p-value)을 갖는 매개변수에 해당하는 키워드를 선택할 수 있다. 유의확률은 매개변수가 유의한지 안 한지 확인하기 위해 요구되는 가장 작은 유의수준으로, 만약 매개변수의 회귀결과에서 유의확률이 0.05보다 작다면 이 매개변수는 유의한 것으로 판단할 수 있다. 검정통계량의 유의확률이 0.05보다 작을 때, X_i 의 기술은 유의하다고 판단한다. 즉, 축소된 회귀모델을 생성하기 위하여, 유의확률이 0.05보다 작은 매개변수에 해당하는 키워드들을 선택한다. 핵심 키워드를 선정함으로써 정확도와 속도 면에서 매우 효율적이다.

160 단계는 상기 선택된 키워드 간의 유의확률을 이용하여 키워드 간의 관계를 도출하는 단계이다.

보다 구체적으로, 상기 선택된 유의한 키워드들을 분석하고자 하는 기술과의 관계를 키워드 간의 유의확률을 이용하여 도출할 수 있다. 상기 키워드 간의 관계는 상기 키워드 간의 유의확률을 이용하여 계층을 나누어 나타낼 수 있다. 계층을 나누어, 기술 계층 다이어그램(A Hierarchical Diagram of Technology, HDT)으로 나타낼 수 있다. 상위 계층 기술과 하위 계층의 기술에 대한 관계는 다음과 같이 나타낼 수 있다.

수학식 5

$$X_1 = u_{01} + u_{11} W_1$$

- [0044] X_1 은 영향을 받은 기술이고, W_1 은 개발된 기술이다. u_{11} 은 X_1 과 W_1 의 인과관계 강도이다. 상기 모델은 기술 Z 를 예측하기 위해 모든 인과관계 강도를 찾는 것이다. W_m 기술은 $(ump*vpn*rn)$ 에 인과관계 강도에 따라 기술 Z 개발에 영향을 미친다. 그러므로 TF의 Z 를 제외하고 모든 기술의 모든 인과관계 강도를 산출할 수 있다.
- [0045] 상기 키워드 간의 관계를 도출하기 위해, 도 2의 210 단계 내지 220 단계를 수행할 수 있다.
- [0046] 210 단계는 상기 선택된 키워드 중 제 2 임계치 이하의 유의확률을 갖는 매개변수에 해당하는 하나 이상의 키워드를 제 1 키워드로 선정하는 단계이다.
- [0047] 제 1 임계치 이하의 유의확률을 갖는 매개변수에 해당하는 키워드들 중 제 2 임계치 이하의 유의확률을 갖는 매개변수에 해당하는 하나 이상의 키워드를 제 1 키워드로 선정한다. 제 1 키워드는 분석하고자 하는 기술과 가장 관련이 높은 기술들로, 제 2 임계치 이하의 유의확률을 갖는 매개변수에 해당하는 키워드를 상위계층에 해당하는 제 1 키워드로 선정한다. 제 2 임계치는 미리 설정되어 있거나, 사용자의 입력에 따라 설정될 수 있다. 또는 유의확률의 낮은 순서대로 미리 설정된 수만큼 제 1 키워드로 선정할 수도 있다.
- [0048] 220 단계는 상기 제 1 키워드와 제 1 키워드 이외의 키워드들 간의 회귀모델 분석을 통해, 제 3 임계치 이하의 유의확률을 갖는 매개변수에 해당하는 키워드를 상기 제 1 키워드와 관계성이 있는 키워드로 선정하는 단계이다.
- [0049] 보다 구체적으로, 제 1 키워드를 선정한 후 다음 계층의 키워드를 선정하는 단계이다. 계층의 깊이는 미리 설정되어 있거나, 키워드의 수, 또는 분석하고자 하는 정도에 따라 달라질 수 있다. 제 1 키워드와 관련된 다음 계층의 키워드를 선정하기 위하여, 제 1 키워드와 제 1 키워드 이외의 키워드들 간의 회귀모델 분석을 수행한다. 상기 제 1 키워드와 제 1 키워드 이외의 키워드들 간의 회귀모델 분석을 통해, 제 3 임계치 이하의 유의확률을 갖는 매개변수에 해당하는 키워드를 상기 제 1 키워드와 관계성이 있는 키워드로 선정한다. 본 단계는 140 단계 내지 150 단계에 대응한다.
- [0050] 도 4는 본 발명의 또 다른 실시예에 따른 특허분석방법의 흐름도이다.
- [0051] 정해진 TF 분야의 특허문서를 검색하고, 특허데이터의 요약(초록)을 추출하여 문서-단어 행렬(DTM)을 구성한다. 문서-단어 행렬(DTM)으로부터 문서-키워드 행렬로 재구성한다. 이를 위하여, 문서-단어 행렬(DTM)에서 상위에 랭크된 단어(ranked terms)들을 검색하고, 의미 없는 단어는 제거한 후, 남아있는 단어들을 이용하여 키워드를 선택하여 문서-키워드 행렬 구성한다. 이후, Hierarchical 회귀분석 실시한다. 즉, 모든 키워드들을 사용하여 full regression model 구성한 후, 선택된 키워드를 사용하여 reduced regression model 구성하고, 완성된 HDT 구성하여 구성된 HDT모델 해석하여 분석하고자 하는 기술을 분석한다.
- [0052] 도 5 내지 6은 본 발명의 실시예에 따른 특허분석방법에 따라 도출된 기술계층을 나타낸 것이다.
- [0053] 본 발명의 실시예에 따른 특허분석을 이용한 결과는 도 6과 같다. KIPRIS와 USPTO(KIPRIS 2013; USPTO 2013)으로부터 검색한 특허문서를 이용하여 결과를 도출하였다. TF를 이용하는 기술분야로 telematics를 선택했다. data set은 미국, 유럽, 중국에 출원된 telematics 특허들로 구성되었고 전체 특허문서 수는 총 474건을 이용하였다. 출원 수는 2000년대부터 현저하게 증가했다. DTM구성을 위해 검색된 특허문서의 초록을 사용하여 행렬의 행과 열을 각각 문서와 단어들로 나타냈다. 행렬의 요소들은 각 문서에서 단어 발생 빈도수를 나타낸다. DTM을 사용하여 telematics의 TF를 위해 Hierarchical regression을 실시했다. Path analysis를 위한 변수선택을 위해 telematics를 포함한 상위 20개 키워드를 선정했다. 키워드들로는 center, communication, control, device, diagnostic, information, message, mobile, network, receiving, request, service, signal, telematics, terminal, transmit, unit, user, vehicle를 선정했다. 이를 통해 아래와 같은 full regression을 구했다.

수학적 6

[0054]
$$telematics = \beta_0 + \beta_1 center + \dots + \beta_{19} vehicle + \epsilon$$

[0055] 회귀모델에서 종속변수는 telematics이고 telematics를 제외한 모든 키워드들은 독립변수들로 사용되었다. 표1은 full regression 분석결과를 나타낸다.

표 1

[0056]

Independent variable	Beta	p-value
center	0.126	0.001
communication	-0.004	0.857
connection	-0.009	0.642
control	0.037	0.072
device	0.286	0.001
diagnostic	-0.044	0.024
information	0.022	0.346
message	0.024	0.236
mobile	0.003	0.893
network	0.060	0.003
receiving	0.086	0.001
request	0.063	0.004
service	0.169	0.001
signal	0.019	0.440
terminal	0.219	0.001
transmit	-0.154	0.001
unit	0.321	0.001
user	0.067	0.001
vehicle	0.152	0.001

[0057] Beta는 표준화된 변수의 회귀 매개변수를 나타내고, 모든 독립변수들의 효과를 비교할 수 있다. 즉, beta는 각 독립변수에서 단위변화와 연관된 “telematics”의 변화이다. 예를 들면, “center”의 단위변화는 “telematics”의 발생에 0.126씩 증가한다. 게다가, 우리는 “unit”이 “telematics”에 가장 큰 영향을 미친다는 것을 발견했다. 종속변수에 대한 독립변수의 유의성은 p-value에 의해 결정된다. 95%의 신뢰구간에서 p-value가 0.05보다 작다면 유의한 변수라고 결정할 수 있다. 따라서, “center”, “device”, “diagnostic”, “network”, “receiving”, “request”, “service”, “terminal”, “transmit”, “unit”, “user”, “vehicle”를 키워드로 선택하여 HDT모델의 변수로 이용하였다.

[0058] 다음은 아래와 같이 “telematics”의 TF를 위한 HDT생성을 위해 reduced regression model을 구성하였다.

표 2

[0059]

Independent variables	Beta	p-value
center	0.124	0.001
device	0.285	0.001
diagnostic	-0.045	0.024
network	0.061	0.019
receiving	0.101	0.002
request	0.063	0.001
service	0.171	0.003
terminal	0.220	0.001
transmit	-0.134	0.001
unit	0.325	0.001
user	0.069	0.001
vehicle	0.165	0.001

[0060] 모든 독립변수들의 p-value가 0.05이하이므로 독립변수들은 모두 유의하다. device, unit, terminal의 beta가 다른 것들보다 큰 값을 가졌기 때문에 도 5와 같이, 제 1 키워드로 선정하여 첫 번째 HDT모델을 구성하였다.

[0061] 결과적으로, 변수 “device”, “unit”, “terminal”는 직접적으로 “telematics”에 영향을 미치는 것을 받

견하였다. 또한 다른 독립변수들은 “device”, “unit”, “terminal”를 경유하여 “telematics”에 간접적으로 영향을 미쳤다. 그러므로 “telematics” TF를 위한 HDT모형을 완성하기 위해 세 종속변수들(i.e. device, unit, terminal)의 세 개 회귀모형을 구성하였고, 결과는 아래와 같다.

표 5

Independent variable	Dependent variable					
	device		unit		terminal	
	Beta	p-value	Beta	p-value	Beta	p-value
center	-0.093	<u>0.001</u>	0.093	<u>0.001</u>	0.001	0.962
diagnostic	0.040	0.140	-0.093	<u>0.001</u>	0.015	0.579
network	0.123	<u>0.001</u>	0.019	0.464	0.068	<u>0.011</u>
receiving	0.060	<u>0.048</u>	0.070	<u>0.020</u>	0.131	<u>0.001</u>
request	-0.020	0.253	0.025	0.391	0.055	0.062
service	0.033	0.253	0.044	0.125	0.055	<u>0.001</u>
transmit	-0.096	<u>0.001</u>	0.096	<u>0.001</u>	0.100	<u>0.001</u>
user	0.048	<u>0.001</u>	0.182	<u>0.001</u>	0.124	0.388
vehicle	0.016	<u>0.001</u>	0.079	<u>0.005</u>	-0.053	0.066

여기에서 세 독립변수에 따라, p-value에 기반하여 통계적으로 유의성을 사용하여 각 통계적으로 유의한 독립변수를 선택하였다. p-value가 0.05(신뢰수준 95%) 이하인 것들을 고려하여, device에 대한 독립변수로 p-value가 0.05이하인 center, network, receiving, transmit, user, vehicle를 선택하였다. 같은 방법으로 아래와 같이 unit과 terminal에 대한 독립변수를 결정하였다.

표 6

Dependent variable	Selected independent variable
device	center, network, receiving, transmit, user, vehicle
unit	center, diagnostic, receiving, transmit, user, vehicle
terminal	network, receiving, service, transmit

표4의 결과로부터, center, network, receiving, transmit, user, vehicle의 기술이 device 기술 개발에 영향을 미쳤다는 것을 알 수 있고 이 개발이 telematics의 기술개발에 영향을 미쳤다는 것을 알 수 있다. unit과 terminal 기술의 경우도 device 과 마찬가지로. 표4의 결과를 이용하여, 도 6의 telematics TF를 관한 complete hierarchical model을 만들었다.

이를 통해, “telematics” TF의 기술적 연관성을 알 수 있다. 각 연결강도는 아래부터 상단까지의 강도를 의미한다. 예를 들어, “receiving” 기술은 “terminal” 기술에 0.133의 연결강도만큼 영향을 미친다. 게다가 이 가중치는 p-value가 0.05보다 작으므로 유의하다. “terminal” 기술은 “telematics” 기술에 0.422의 연결강도만큼 연속적으로 영향을 미친다. 만약 “telematics” 기술에 대한 “receiving” 기술의 직접적인 영향을 알고 싶다면, 두 가중치를 곱해야한다: “terminal”에서의 “receiving” (0.133)과 “telematics”에서의 “terminal” (0.422)의 가중치. 그러므로 “receiving” 기술은 간접적으로 “telematics” 기술에 $0.056 (0.133 \times 0.422)$ 의 연결강도만큼 영향을 미친다. Hierarchical diagram의 나머지부분들도 같은 방법으로 설명될 수 있다.

130 단계의 키워드를 선택함에 있어서, 다음과 같은 일련의 과정(410 단계 내지 을 통해 키워드를 선택할 수 있다.

410 단계는 상기 생성된 문서-단어 행렬로부터 주성분분석을 통해 문서-주성분분석 행렬을 생성하는 단계이다.

보다 구체적으로, 추출된 단어들 중 핵심 키워드를 선정하는 것은 매우 어려우며, 선정된 키워드에 따라 이후 결과에 지대한 영향을 미치기 때문에 선불리 단어의 개수를 줄일 수 없다. 객관적이고 핵심 키워드를 추출하기 위하여, 주성분 분석(Principal component analysis, PCA)을 이용하여 행렬의 희소성을 해결하면서 동시에 손실

될 수 있는 정보를 가능한 보존할 수 있다. 주성분분석은 다양한 변수들에 대해 분석하는 소위 다변량(multivariate) 분석인데, 많은 변수들로부터 몇 개의 주성분들을 추출하는 방법이다. 즉, 주성분분석은 차원 축소(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. 주성분분석을 통해 생성되는 문서-주성분분석 행렬은 주성분(PC)의 개수는 처음 변수인 단어(Term)의 개수만큼 생성된다.

[0070] 여기서, 효율적인 분석을 위하여, PC의 개수를 줄일 수 있다. 상기 주성분분석 결과, 임계치 이상의 고유 값을 갖는 주성분만을 이용하여 축소된 문서-주성분분석 행렬을 생성할 수 있다. 유의미한 주성분만을 이용하여 분석을 수행하기 위하여, 주성분의 고유 값이 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 임계치 이상인 경우의 주성분만을 이용하여 문서-주성분분석 행렬을 축소한다. 임계치는 1일 수 있다. 즉, 고유 값이 1이상인 주성분만을 이용하여 주성분의 수를 줄여 차원축소가 가능하다.

[0071] 420 단계는 문서-주성분분석 행렬을 이용하여 회귀모형을 생성하는 단계이다. 보다 구체적으로, 문서-주성분분석 행렬을 이용하여 회귀모형을 생성한다. 회귀모형은 하나 또는 그 이상의 독립변수의 종속변수에 대한 영향의 추정을 할 수 있는 통계기법으로 주성분의 통계적 분석이 가능하다. 즉, 회귀모형을 생성하여 선택된 주성분들 중 통계적으로 유의한 주성분을 찾을 수 있다. 주성분분석은 통계적으로 유의한 분석이 아닌바, 회귀모형을 이용함으로써 통계적으로 유의한 분석이 가능하다. 상기 문서-주성분분석 행렬로부터 회귀모형을 생성한다. 회귀모형은 다음과 같이 나타낼 수 있다.

수학식 7

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

[0072]

[0073] 430 단계는 상기 생성된 회귀모형의 매개변수들 중 임계치 이하의 유의확률(p-value)을 갖는 매개변수에 해당하는 주성분을 선택하는 단계이다. 보다 구체적으로, 420 단계에서 생성된 회귀모형에서 통계적으로 유의한 주성분을 선택하기 위하여, 회귀모형의 매개변수를 이용한다. 수학식 7에서 β_1 과 β_2 는 유의확률(p-value)을 갖는다. 임계치 이하인 유의확률을 갖는 매개변수에 해당하는 주성분을 유의한 주성분으로 선택할 수 있다. 상기 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 회귀모형을 통해 핵심 키워드를 선정함으로써 정확도와 속도 면에서 매우 효율적이다.

[0074] 440 단계는 상기 선택된 주성분에 속하는 하나 이상의 단어를 핵심 키워드로 선정하는 단계이다. 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 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 임계치 이상의 주성분 점수를 갖는 단어를 핵심 키워드로 선택할 수 있다. 주성분 점수는 -1 내지 1일 수 있고, 상기 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. 임계치와 단어의 수 둘 모두에 따라 선정될 수도 있다.

[0075] 주성분분석(PCA)과 회귀모형을 이용하여 통계적으로 유의한 매개변수 β 를 찾아내고 선정된 매개변수를 통해 중요한 키워드를 선정하여 정량적 기술동향 분석을 할 수 있다.

[0076] 본 발명의 실시예는 문서의 단어를 추출하여, 문서-단어 행렬을 생성하고, 문서-키워드 행렬을 생성하고, 회귀모형을 생성하여, 키워드 간의 관계를 도출하는 하나 이상의 처리부(프로세서) 및 처리부에서 산출되는 결과 및 제거되는 단어들을 저장하는 하나 이상의 저장부(데이터베이스)를 포함할 수 있다.

[0077] 본 발명의 실시예들은 다양한 컴퓨터 수단을 통하여 수행될 수 있는 프로그램 명령 형태로 구현되어 컴퓨터 판

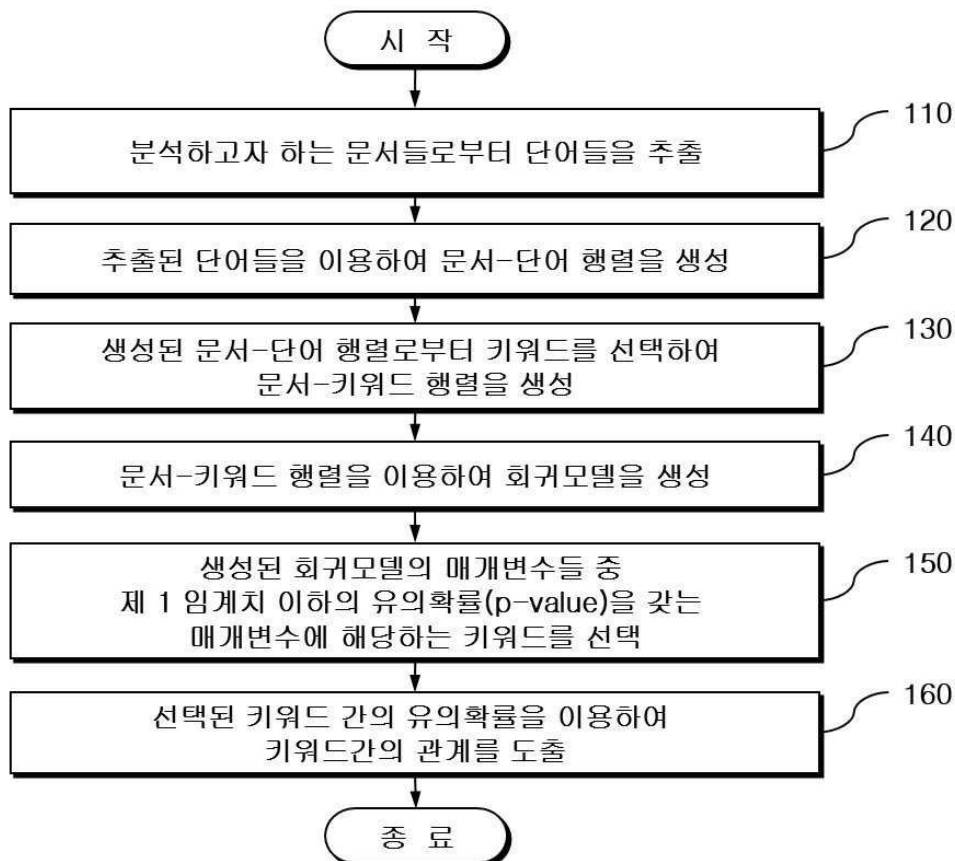
독 가능 매체에 기록될 수 있다. 상기 컴퓨터 판독 가능 매체는 프로그램 명령, 데이터 파일, 데이터 구조 등을 단독으로 또는 조합하여 포함할 수 있다. 상기 매체에 기록되는 프로그램 명령은 본 발명을 위하여 특별히 설계되고 구성된 것들이거나 컴퓨터 소프트웨어 당업자에게 공지되어 사용 가능한 것일 수도 있다. 컴퓨터 판독 가능 기록 매체의 예에는 하드 디스크, 플로피 디스크 및 자기 테이프와 같은 자기 매체(magnetic media), CD-ROM, DVD와 같은 광기록 매체(optical media), 플롭티컬 디스크(floptical 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.

[0078] 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.

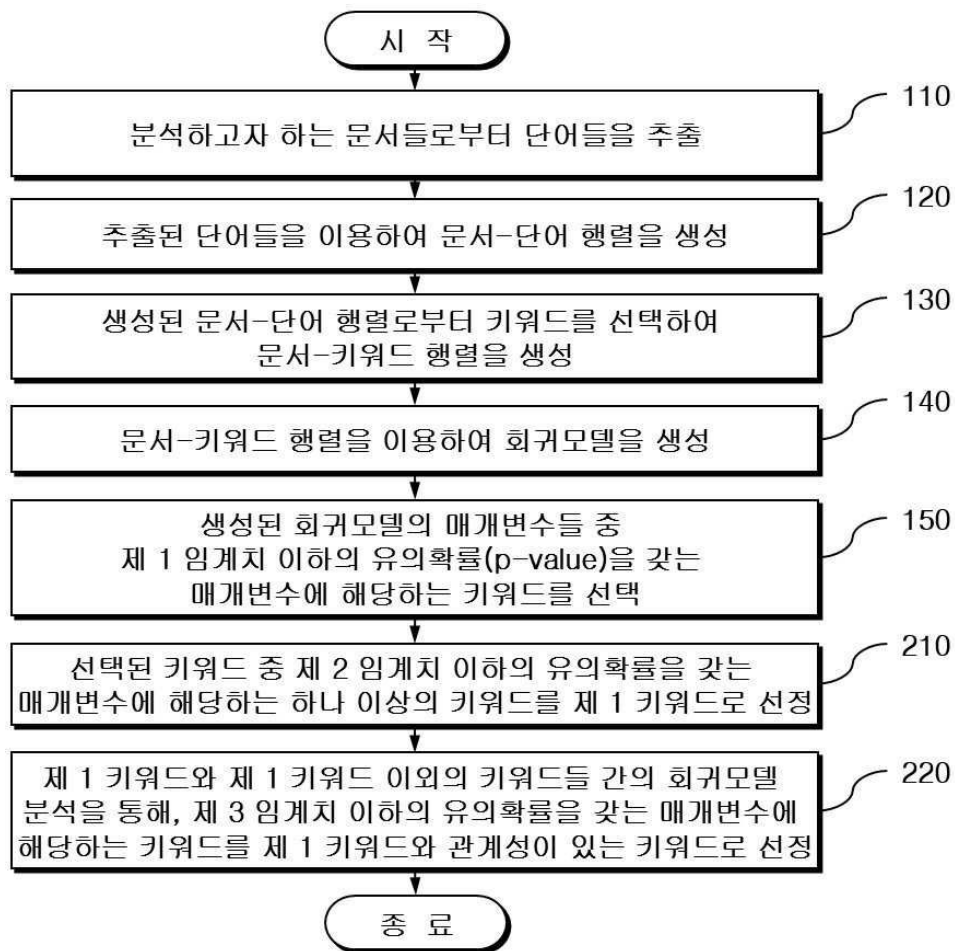
[0079] 따라서, 본 발명의 사상은 설명된 실시예에 국한되어 정해져서는 아니되며, 후술하는 특허청구범위뿐 아니라 이 특허청구범위와 균등하거나 등가적 변형이 있는 모든 것들은 본 발명 사상의 범주에 속한다고 할 것이다.

도면

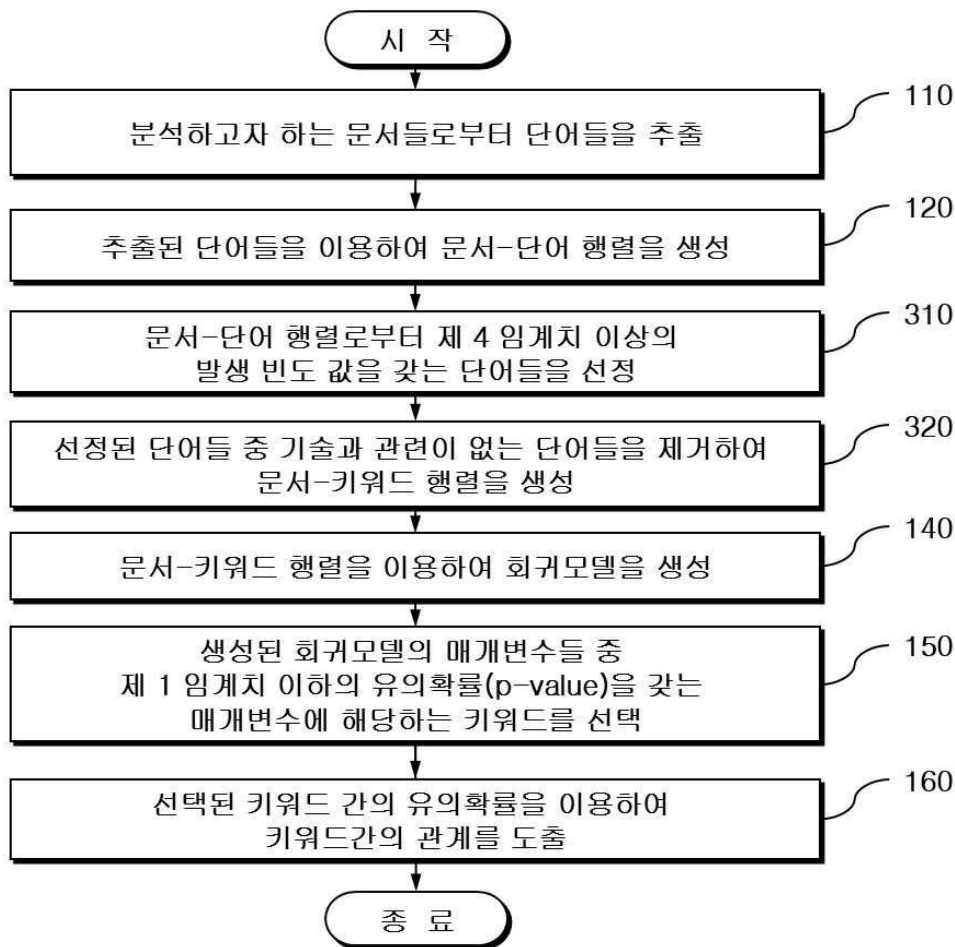
도면1



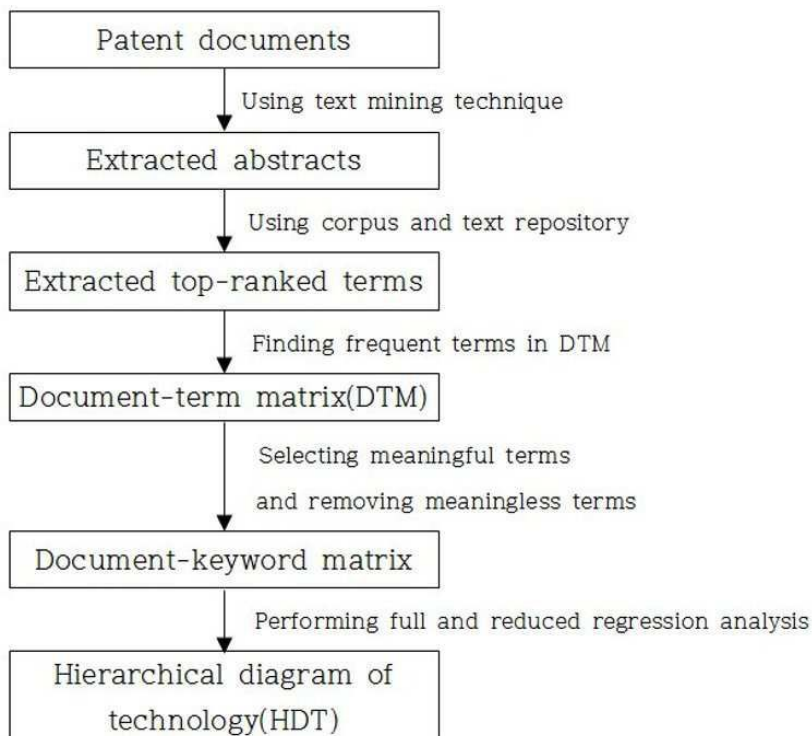
도면2



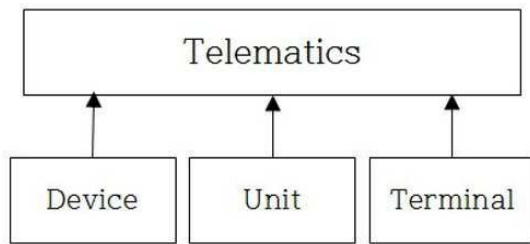
도면3



도면4



도면5



도면6

Telematics		
1 st	2 nd	3 rd
Terminal, (0.422)	Unit, (0.352)	Device, (0.282)
1 st : receiving (0.133)	1 st : user (0.195)	1 st : vehicle (0.147)
2 nd : transmit(0.130)	2 nd : transmit (0.106)	2 nd : user (0.138)
3 rd : service (0.113)	3 rd : center (0.103)	3 rd : network (0.097)
4 th : network (0.067)	4 th : diagnostic (-0.094)	4 th : transmit (0.097)
	5 th : vehicle (0.087)	5 th : center (-0.091)
	6 th : receiving (0.079)	6 th :receiving (0.063)