(19) 대한민국특허청(KR)
(12) 등록특허공보(B1)

(45) 공고일자 2016년09월22일
(11) 등록번호 10-1658794
(24) 등록일자 2016년09월13일

(51) 국제특허분류(Int. Cl.)
G06F 17/30 (2006.01) *G06Q 50/18* (2012.01)
(52) CPC특허분류
G06F 17/30598 (2013.01)
G06F 17/30707 (2013.01)
(21) 출원번호 10-2015-7034318(분할)
(22) 출원일자(국제) 2013년02월28일
심사청구일자 2015년12월01일
(85) 번역문제출일자 2015년12월01일
(65) 공개번호 10-2015-0142070
(43) 공개일자 2015년12월21일
(62) 원출원 특허 10-2014-7026134
원출원일자(국제) 2013년02월28일
심사청구일자 2014년09월18일
(86) 국제출원번호 PCT/JP2013/055330
(87) 국제공개번호 WO 2013/129548
국제공개일자 2013년09월06일
(30) 우선권주장
JP-P-2012-044382 2012년02월29일 일본(JP)
(56) 선행기술조사문헌
KR1020080041388 A

(73) 특허권자
가부시키가이샤 유빅
일본국 도쿄토 미나토쿠 코난 2-12-23 메이산 타카하마 빌딩 7층
(72) 발명자
모리모토 마사히로
일본 도쿄 미나토구 코난 2-12-23 메이산 타카하마 빌딩 7층 가부시키가이샤 유빅
시라이 요시카츠
일본 도쿄 미나토구 코난 2-12-23 메이산 타카하마 빌딩 7층 가부시키가이샤 유빅
(뒷면에 계속)
(74) 대리인
허성원, 이동욱, 서동헌

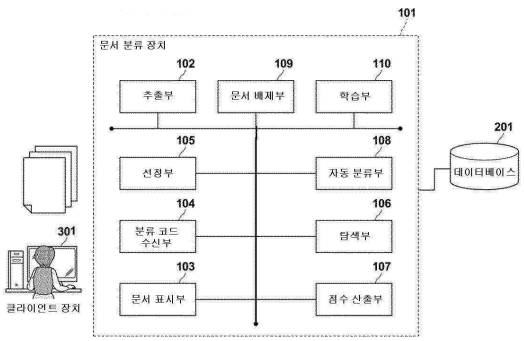
전체 청구항 수 : 총 7 항 심사관 : 고재용

(54) 발명의 명칭 **문서 분류 시스템, 문서 분류 방법 및 문서 분류 프로그램**

(57) 요 약

소송에서 증거로 제출하기 위하여 수집된 디지털화된 문서 정보를 분석하고, 소송에 이용이 용이하도록 분류한다. 문서 정보로부터 복수의 문서를 샘플링하는 것에 의하여, 상기 복수의 문서를 사용자에 의한 분류 대상으로 추출하는 추출부와, 상기 추출된 복수의 문서에 대하여, 각 문서를 분류하기 위한 것으로, 상기 사용자가 부여한 분류 코드를 수신하는 분류 코드 수신부와, 상기 분류 코드가 부여된 문서로부터 공통으로 출현하는 키워드를 선정하는 선정부와, 상기 분류 코드가 부여된 문서로부터 상기 선정된 키워드와 상기 키워드의 가중치를 대응시켜 기록하는 데이터베이스와, 상기 분류 코드가 부여되지 않는 미분류문서에 포함되는 키워드와 상기 데이터베이스에서 상기 키워드에 대응시켜 부여한 가중치에 따라, 상기 미분류문서와 상기 분류 코드의 관련성을 평가한 점수를 산출하는 산출부를 포함하는 문서 분류 시스템이 제공된다.

대 표 도 - 도1



(52) CPC특허분류
G06Q 50/18 (2013.01)
(72) 발명자
타케다 히데키
일본 도쿄 미나토구 코난 2-12-23 메이산 타카하마 빌딩 7층 가부시키가이샤 유빅
하스코 카즈미
일본 도쿄 미나토구 하마마츠쵸 1-1-2

명 세 서

청구범위

청구항 1

문서 정보로부터 복수의 문서를 샘플링하는 것에 의하여, 상기 복수의 문서를 사용자에 의한 분류 대상으로 추출하는 추출부와,

상기 추출된 복수의 문서에 대하여, 각 문서를 분류하기 위한 것으로, 상기 사용자가 부여한 분류 코드를 수신하는 분류 코드 수신부와,

상기 분류 코드가 부여된 문서로부터 공통으로 출현하는 키워드를 선정하는 선정부와,

상기 분류 코드가 부여된 문서로부터 상기 선정된 키워드와 상기 키워드의 가중치를 대응시켜 기록하는 데이터베이스와,

상기 분류 코드가 부여되지 않는 미분류문서에 포함되는 키워드와 상기 데이터베이스에서 상기 키워드에 대응되는 가중치에 따라, 상기 미분류문서와 상기 분류 코드의 결부된 강도를 평가한 점수를 산출하는 산출부를 포함하는 문서 분류 시스템.

청구항 2

제1항에 있어서,

상기 선정부에 의해 선정된 키워드에 대한 가중치를 학습하는 학습부를 더 포함하는 것을 특징으로 하는 문서 분류 시스템.

청구항 3

제1항에 있어서,

상기 산출부가 산출한 점수에 기초하여, 상기 미분류문서에 상기 분류 코드를 부여하는 자동 분류부를 더 포함하는 것을 특징으로 하는 문서 분류 시스템.

청구항 4

제1항 내지 제3항 중 어느 한 항에 있어서,

상기 미분류문서로부터, 상기 데이터베이스에 기록된 키워드를 탐색하는 탐색부를 더 포함하는 것을 특징으로 하는 문서 분류 시스템.

청구항 5

문서 정보로부터 복수의 문서를 샘플링하는 것에 의하여, 상기 복수의 문서를 사용자에 의한 분류 대상으로 추출하는 추출부와,

상기 추출된 복수의 문서에 대하여, 각 문서를 분류하기 위한 것으로, 상기 사용자가 부여한 분류 코드를 수신하는 분류 코드 수신부와,

상기 분류 코드가 부여된 문서로부터 선정된 키워드와 상기 키워드의 가중치를 대응시켜 기록하는 데이터베이스와,

상기 분류 코드가 부여되지 않는 미분류문서로부터, 상기 데이터베이스에 기록된 키워드를 탐색하는 탐색부와,

상기 미분류문서로부터 탐색된 키워드와 상기 데이터베이스에서의 상기 키워드에 대응되는 가중치에 따라, 상기 미분류문서와 상기 분류 코드의 결부된 강도를 평가한 점수를 산출하는 산출부를 포함하고,

상기 데이터베이스는, 상기 분류 코드와 상관관계가 있는 관련용어와 상기 관련용어의 가중치를 대응시켜 기록하고,

상기 탐색부는 상기 미분류문서로부터 상기 관련용어를 탐색하고,

상기 산출부는, 상기 미분류문서와 상기 분류 코드의 결부된 강도를, 상기 관련용어와 상기 관련용어에 대응되는 가중치에 따라 상기 점수를 산출하는 문서 분류 시스템.

청구항 6

문서 정보로부터 복수의 문서를 샘플링하는 것에 의하여, 상기 복수의 문서를 사용자에 의한 분류 대상으로 추출하는 추출 단계와,

상기 추출된 복수의 문서에 대하여, 각 문서를 분류하기 위한 것으로, 상기 사용자가 부여한 분류 코드를 수신하는 분류 코드 수신 단계와,

상기 분류 코드가 부여된 문서로부터 공통으로 출현하는 키워드를 선정하는 선정 단계와,

상기 분류 코드가 부여된 문서로부터 상기 선정된 키워드와 상기 키워드의 가중치를 대응시켜 기록하는 데이터베이스를 참조하여, 상기 분류 코드가 부여되지 않는 미분류문서에 포함되는 키워드와 상기 데이터베이스에서의 상기 키워드에 대응되는 가중치에 따라, 상기 미분류문서와 상기 분류 코드의 결부된 강도를 평가한 점수를 산출하는 산출 단계를 포함하는 컴퓨터가 실행하는 문서 분류 방법.

청구항 7

컴퓨터에,

문서 정보로부터 복수의 문서를 샘플링하는 것에 의하여, 상기 복수의 문서를 사용자에 의한 분류 대상으로 추출하는 추출 기능과,

상기 추출된 복수의 문서에 대하여, 각 문서를 분류하기 위한 것으로, 상기 사용자가 부여한 분류 코드를 수신하는 분류 코드 수신 기능과,

상기 분류 코드가 부여된 문서로부터 공통으로 출현하는 키워드를 선정하는 선정 기능과,

상기 분류 코드가 부여된 문서로부터 상기 선정된 키워드와 상기 키워드의 가중치를 대응시켜 기록하는 데이터베이스를 참조하여, 상기 분류 코드가 부여되지 않는 미분류문서에 포함되는 키워드와 상기 데이터베이스에서의 상기 키워드에 대응되는 가중치에 따라, 상기 미분류문서와 상기 분류 코드의 결부된 강도를 평가한 점수를 산출하는 산출 기능을 실현시키기 위한 문서 분류 프로그램을 기록한 컴퓨터로 읽을 수 있는 기록매체.

발명의 설명

기 술 분 야

[0001] 본 발명은, 문서 분류 시스템, 문서 분류 방법 및 문서 분류 프로그램에 관한 것으로서, 특히 소송에 관한 문서 정보의 문서 분류 시스템, 문서 분류 방법 및 문서 분류 프로그램에 관한 것이다.

배 경 기 술

[0002] 종래, 부정한 액세스 또는 기밀 정보 누설 등 컴퓨터 관련 범죄나 법적 분쟁이 생겼을 때, 원인 규명이나 수사에 필요한 기기 또는 데이터, 전자적 기록을 수집·분석하고, 그 법적인 증거성을 밝히려는 수단과 기술이 제안되어 있다.

[0003] 특히 미국 민사 소송에서는, 전자 증거 개시(eDiscovery) 등이 요구되고 있으며, 해당 소송의 원고 및 피고의 어느 쪽이나, 관련 디지털 정보를 모두 증거로 제출하는 책임을 진다. 따라서 컴퓨터나 서버에 기록된 디지털 정보를 증거로 제출하여야 한다.

[0004] 한편, 정보기술(IT)의 급속한 발달 및 보급과 더불어, 오늘날의 비즈니스 세계는 대부분의 정보가 컴퓨터로 작성되어 있기 때문에, 동일한 기업 내에서도 다수의 디지털 정보가 범람하고 있다.

[0005] 따라서 법정에 증거 자료 제출을 위한 준비 작업을 하는 과정에서, 해당 소송에 반드시 관련되지 않은 기밀이나 디지털 정보까지도 증거 자료로 포함되어 버리는 실수가 일어나기 쉽다. 또한, 해당 소송과 관련되지 않은 기밀이나 문서 정보를 제출해 버리는 문제가 있었다.

[0006] 최근 포렌식 시스템(Forensic system)의 문서 정보에 관한 기술이, 특허문헌 1 내지 특허문헌 3에 제안되어 있다. 특허문헌 1에는, 이용자 정보에 포함된 적어도 한 명 이상의 이용자로부터 특정 이용자를 지정하고, 지정된 특정 이용자에 관한 액세스 이력 정보에 기초하여, 특정 이용자가 액세스한 디지털 문서 정보만을 추출하고, 추출된 디지털 문서 정보의 문서 파일 각각이, 소송에 관련된 것인지 여부를 나타내는 부대 정보를 설정하고, 부대 정보에 기초하여 소송 관련 문서 파일을 출력하는 포렌식 시스템에 대해 개시되어 있다.

[0007] 또한, 특허문헌 2에는, 기록된 디지털 정보를 표시하고, 복수의 문서 파일마다 이용자 정보에 포함된 이용자 중 어느 이용자에게 관련된 것인지를 나타내는 이용자 특정 정보를 설정하고, 상기 설정된 이용자 특정 정보를 저장부에 기록하도록 설정하고, 적어도 한 명 이상의 이용자를 지정하고, 지정된 이용자에 대응하는 이용자 특정 정보가 설정된 문서 파일을 검색하고, 표시부를 통해 검색된 문서 파일이 소송에 관련된 것인지 여부를 나타내는 부대 정보를 설정하고, 부대 정보에 기초하여 소송 관련 문서 파일을 출력하는 포렌식 시스템에 대해 개시되어 있다.

[0008] 또한, 특허문헌 3에는, 디지털 문서 정보에 포함된 하나 이상의 문서 파일의 지정을 수신하고, 지정된 문서 파일을 어떤 언어로 번역할 것인지의 지정을 수신하고, 수신한 지정 문서 파일을 수신한 지정 언어로 번역하고, 기록부에 기록된 디지털 문서 정보에서 지정된 문서 파일과 동일한 내용을 나타내는 공통 문서 파일을 추출하고, 추출된 공통 문서 파일이 번역된 문서 파일의 번역 내용을 원용함으로써 번역되었다는 것을 나타내는 번역 관련 정보를 생성하고, 번역 관련 정보에 따라 소송 관련 문서 파일을 출력하는 포렌식 시스템에 대해 개시되어 있다.

선행기술문헌

특허문헌

[0009] (특허문헌 0001) 특개2011-209930호 공보

(특허문헌 0002) 특개2011-209931호 공보

(특허문헌 0003) 특개2012-32859호 공보

발명의 내용

해결하려는 과제

[0010] 그러나 가령 특허문헌 1 내지 특허문헌 3과 같은 포렌식 시스템에서는, 복수의 컴퓨터 및 서버를 이용한 이용자의 방대한 문서 정보를 수집하게 된다.

[0011] 이러한 디지털화된 방대한 문서 정보를 소송의 증거 자료로 타당한지 여부를 분류하는 작업은, 리뷰어(reviewer)라는 사용자가 눈으로 확인하고, 해당 문서 정보를 하나하나 분류해 나갈 필요가 있고, 상당한 노력이 소요되는 문제가 있었다.

[0012] 따라서, 본 발명은, 상기 사정을 감안하여, 디지털화된 문서 정보를 수집한 후, 해당 문서 정보에 대하여, 분류 코드를 자동으로 부여함으로써 소송에 이용할 문서 정보의 분류 작업의 부담을 경감할 수 있는 문서 분류 시스템, 문서 분류 방법 및 문서 분류 프로그램을 제공하는 것을 목적으로 하는 것이다.

과제의 해결 수단

[0013] 본 발명의 문서 분류 시스템은, 복수의 컴퓨터 또는 서버에 기록된 디지털 정보를 획득하고, 상기 획득된 디지털 정보에 포함된 문서 정보를 분석하고, 소송에 이용이 용이하도록 분류하는 문서 분류 시스템에 관한 것으로, 문서 정보로부터 소정 수의 문서를 포함하는 데이터 세트인 문서 그룹을 추출하는 추출부와, 추출된 문서 그룹을 화면 상에 표시하는 문서 표시부와, 표시된 문서 그룹에 대해 사용자가 소송과의 관련성에 기초하여 부여한 분류 코드를 수신하는 분류 코드 수신부와, 분류 코드에 기초하여 추출된 문서 그룹을 분류 코드별로 분류하고, 상기 분류된 문서 그룹에서 공통으로 출현하는 키워드를 분석하여 선정하는 선정부와, 선정된 키워드를 기록하는 데이터베이스와, 데이터베이스에 기록된 키워드를 문서 정보에서 탐색하는 탐색부와, 탐색부의 탐색 결과와 선정부의 해석 결과를 이용하여 분류 코드와 문서와의 관련성을 나타내는 점수를 산출하는 점수 산출부와, 점수

결과에 기초하여 자동으로 분류 코드를 부여하는 자동 분류부를 포함한다.

[0014] 「문서」라 함은 하나 이상의 키워드를 포함하는 데이터를 말한다. 예를 들어 이메일, 프레젠테이션 자료, 표계산 자료, 협의 자료, 계약서, 조직도, 사업 계획서 등이 있다.

[0015] 「키워드」는 한 언어에서 일정한 의미를 가진 문자열의 정리를 말한다. 예를 들어, 「문서를 분류한다」는 문장에서 키워드를 선정하면, 「문서」, 「분류」로 할 수 있다.

[0016] 「분류 코드」는 문서를 분류할 때 사용하는 식별자를 말한다. 예를 들어 소송에서 문서 정보를 증거로 사용할 때에 증거의 종류에 따라 부여할 수 있다.

[0017] 「점수」는 어떤 문서에서 특정 분류 코드와 결부된 강도를 정량적으로 평가한 것을 말한다. 예를 들어 점수 산출부는, 문서 그룹 중에 출현하는 키워드와, 각 키워드가 갖는 가중치에 따라 점수를 산출할 수도 있다. 한 예로서 해당 가중치는, 키워드가 갖는, 각 분류 코드의 전달 정보량을 바탕으로 결정할 수도 있다.

[0018] 또한, 본 발명의 문서 분류 시스템에서, 추출부는 문서 정보에서 문서 그룹을 추출할 때에, 무작위로 샘플링을 할 수 있다.

[0019] 본 발명의 문서 분류 시스템에서, 탐색부는 키워드를 분류 코드가 부여되지 않은 문서로 구성된 문서 정보에서 탐색하는 기능을 구비하고, 점수 산출부는 탐색부의 탐색 결과와 선정부의 해석 결과를 이용하여 분류 코드와 문서의 관련성을 나타내는 점수를 산출하고, 자동 분류부는 분류 코드 수신부에서 분류 코드의 부여를 수신하지 않은 문서를 추출하고 해당 문서에 대해 자동으로 분류 코드를 부여하는 기능을 구비할 수도 있다.

[0020] 또한, 본 발명의 문서 분류 시스템에서, 데이터베이스는 분류 코드와 관련성이 있는 관련 용어를 추출하고 기록하는 기능을 구비하고, 탐색부는 관련 용어를 문서 정보에서 탐색하는 기능을 구비하고, 점수 산출부는 탐색부가 관련 용어를 탐색한 결과를 바탕으로 점수를 산출하는 기능을 구비하고, 또한 자동 분류부는 관련 용어를 이용하여 산출한 점수에 따라 자동으로 분류 코드를 부여하는 기능을 구비할 수도 있다.

[0021] 또한, 본 발명의 문서 분류 시스템은, 문서 그룹에 포함된 문서 중, 선정부가 선정한 키워드, 관련 용어 및 분류 코드와 상관 관계가 있는 키워드를 포함하지 않는 문서를 선정하고, 자동 분류부의 분류 대상에서 선정된 문서를 배제하는 문서 배제부를 구비할 수도 있다.

[0022] 본 발명의 문서 분류 시스템은, 또한, 선정부의 분석 결과 및 점수 산출부가 산출한 점수에 기초하여 선정부가 선정한 데이터베이스에 기록된 분류 코드와의 상관 관계를 갖는 키워드 및 관련 용어를 증감시키는 학습부를 구비할 수도 있다.

[0023] 본 발명의 문서 분류 방법은, 복수의 컴퓨터 또는 서버에 기록된 디지털 정보를 획득하고, 해당 획득된 디지털 정보에 포함된 문서 정보를 분석하고, 소송에 이용이 용이하도록 분류하는 문서 분류 방법에 관한 본 발명에서, 문서 정보로부터 소정 수의 문서를 포함하는 데이터 세트인 문서 그룹을 추출하고, 추출된 문서 그룹을 화면 상에 표시하고, 표시된 문서 그룹에 대해 사용자가 소송과의 관련성에 기초하여 부여한 분류 코드를 수신하고, 분류 코드에 기초하여 추출된 문서 그룹을 분류 코드별로 분류하고, 해당 분류된 문서 그룹에서 공통으로 출현하는 키워드를 해석하여 선정하고, 선정된 키워드를 기록하고, 기록된 키워드를 문서 정보에서 탐색하고, 탐색 결과 및 해석 결과를 이용하여 분류 코드와 문서와의 관련성을 나타내는 점수를 산출하여, 점수 결과에 기초하여 자동으로 분류 코드를 부여하는 기능을 실현하는 것이다.

[0024] 본 발명의 문서 분류 프로그램은, 복수의 컴퓨터 또는 서버에 기록된 디지털 정보를 획득하고, 해당 획득된 디지털 정보에 포함된 문서 정보를 분석하고, 소송에 이용이 용이하도록 분류하는 문서 분류 프로그램에 관한 본 발명에서, 컴퓨터에, 문서 정보로부터 소정 수의 문서를 포함하는 데이터 세트인 문서 그룹을 추출하는 기능과, 추출된 문서 그룹을 화면 상에 표시하는 기능과, 표시된 문서 그룹에 대해 사용자가 소송과의 관련성에 기초하여 부여한 분류 코드를 수신하여 분류 코드에 기초하여 추출된 문서 그룹을 분류 코드별로 분류하고, 상기 분류된 문서 그룹에서 공통으로 출현하는 키워드를 해석하여 선정하는 기능과, 선정한 키워드를 기록하는 기능과, 기록된 키워드를 문서 정보에서 탐색하는 기능과, 탐색 결과와 해석 결과를 이용하여 분류 코드와 문서와의 관련성을 나타내는 점수를 산출하는 기능과, 점수 결과에 따라 자동으로 분류 코드를 부여하는 기능을 실현하는 것이다.

발명의 효과

[0025] 본 발명에 따른 문서 분류 시스템, 문서 분류 방법 및 문서 분류 프로그램은, 문서 정보로부터 소정 수의 문서

를 포함하는 데이터 세트인 문서 그룹을 추출하고, 추출된 문서 그룹을 화면 상에 표시하고, 표시된 문서 그룹에 대해 사용자가 소송과의 관련성에 기초하여 부여한 분류 코드를 수신하여 해당 분류 코드를 기반으로 추출된 문서 그룹을 분류 코드별로 분류하고, 상기 분류된 문서 그룹에서 공통으로 출현하는 키워드를 해석하여 선정하고, 선정된 키워드를 기록하고, 기록된 키워드를 문서 정보에서 탐색하고, 탐색 결과와 해석 결과를 이용하여 분류 코드와 문서와의 관련성을 나타내는 점수를 산출하고, 점수 결과에 기초하여 자동으로 분류 코드를 부여함으로써, 리뷰어의 분류 작업 노력의 경감을 도모할 수 있다.

[0026] 또한, 본 발명의 문서 분류 시스템에서, 탐색부는 키워드를 분류 코드가 부여되지 않은 문서로 구성된 문서 정보에서 탐색하는 기능을 구비하고, 점수 산출부는 탐색부의 탐색 결과와 선정부의 해석 결과를 이용하여 분류 코드와 문서와의 관련성을 나타내는 점수를 산출하고, 자동 분류부는 분류 코드 수신부에서 분류 코드의 부여를 수신하지 않은 문서를 추출하고 해당 문서에 대하여 자동으로 분류 코드를 부여하는 기능을 구비할 때에, 분류 코드 수신부에서 분류 코드의 부여를 수신하지 않은 문서 정보에 대하여, 리뷰어가 분류한 규칙성을 바탕으로 자동으로 분류 코드를 부여할 수 있다.

[0027] 또한, 본 발명은 선정부의 분석 결과와 점수 산출부에서 산출한 점수에 기초하여 선정부가 선정한 데이터베이스에 기록된 분류 코드와의 상관 관계를 갖는 키워드 및 관련 용어를 증감시키는 학습부를 구비한 때에는, 분류 회수를 거듭할 때마다 분류 정밀도를 향상시킬 수 있다.

[0028] 또한, 본 발명은 데이터베이스가 분류 코드와 관련성이 있는 관련 용어를 추출 및 기록하고, 탐색부가 관련 용어를 문서 정보에서 탐색하고, 점수 산출부는 탐색부가 관련 용어를 탐색한 결과를 바탕으로 점수를 산출하고, 자동 분류부가 관련 용어를 이용하여 산출한 점수에 따라 자동으로 분류 코드를 부여하는 것으로, 문서 그룹에 포함된 문서 중 선정부가 선정한 키워드, 관련 용어 및 분류 코드와 상관 관계가 있는 키워드를 포함하지 않는 문서를 선정하고, 자동 분류부의 분류 대상에서 선정된 문서를 배제할 때에는, 문서 분류를 보다 효율적으로 수행할 수 있다. 이것은 수집된 디지털 정보의 소송에서의 이용을 용이하게 한다.

도면의 간단한 설명

[0029] 도 1은 본 발명의 제1실시예에 따른 문서 분류 시스템의 구성도이고,

도 2는 본 발명의 실시예에 따른 선정부에서의 해석 결과를 도시한 그래프이고,

도 3은 본 발명의 실시예에 따른 각 단계의 처리 흐름을 도시한 흐름도이고,

도 4는 본 발명의 실시예에 따른 데이터베이스의 처리 흐름을 도시한 흐름도이고,

도 5는 본 발명의 실시예에 따른 탐색부의 처리 흐름을 도시한 흐름도이고,

도 6은 본 발명의 실시예에 따른 점수 산출부의 처리 흐름을 도시한 흐름도이고,

도 7은 본 발명의 실시예에 따른 자동 분류부의 처리 흐름을 도시한 흐름도이고,

도 8은 본 발명의 실시예에 따른 추출부의 처리 흐름을 도시한 흐름도이고,

도 9는 본 발명의 실시예에 따른 문서 표시부의 처리 흐름을 도시한 흐름도이고,

도 10은 본 발명의 실시예에 따른 분류 코드 수신부의 처리 흐름을 도시한 흐름도이고,

도 11은 본 발명의 실시예에 따른 선정부의 처리 흐름을 도시한 흐름도이고,

도 12는 본 발명의 실시예에 따른 문서 배제부의 처리 흐름을 도시한 흐름도이고,

도 13은 본 발명의 실시예에 따른 학습부의 처리 흐름을 도시한 흐름도이고,

도 14는 본 발명의 실시예에 따른 문서 표시 화면이다.

발명을 실시하기 위한 구체적인 내용

[0030] [제1실시예]

[0031] 아래에서, 본 발명의 실시예를 첨부한 도면에 의해 설명한다. 도 1에 제1실시예에 따른 문서 분류 시스템의 구성도를 도시한다.

[0032] 제1실시예는, 특허 침해 소송에서의 문서 제출 명령에 대응할 때에, 피의 제품인 제품 A에 관한 문서를 분류 처

리하는 경우의 실시예이다.

[0033] 본 발명에 따른 문서 분류 시스템은, 문서 정보로부터 소정 수의 문서를 포함하는 데이터 세트인 문서 그룹을 추출하는 추출부(102)와, 추출된 문서 그룹을 화면 상에 표시하는 문서 표시부(103)와, 표시된 문서 그룹에 대해 리뷰어라는 사용자가 소송과의 관련성에 기초하여 부여한 분류 코드를 수신하는 분류 코드 수신부(104)와, 분류 코드에 기초하여 추출된 문서 그룹을 분류 코드별로 분류하고, 해당 분류된 문서 그룹에서 공통으로 출현하는 키워드를 해석하여 선정하는 선정부(105)와, 선정된 키워드를 기록하는 데이터베이스(201)와, 데이터베이스(201)에 기록된 키워드를 문서 정보에서 탐색하는 탐색부(106)와, 탐색부(106)의 탐색 결과와 선정부(105)의 해석 결과를 이용하여 분류 코드와 문서와의 관련성을 나타내는 점수를 산출하는 점수 산출부(107)와, 점수 결과에 기초하여 자동으로 분류 코드를 부여하는 자동 분류부(108)와, 자동 분류부(108)의 분류 대상에서 선정된 문서를 배제하는 문서 배제부(109)와, 선정부(105)의 분석 결과 및 점수 산출에서 산출한 점수에 기초하여 선정부(105)가 선정한 키워드, 데이터베이스(201)에 기록된 분류 코드와의 상관 관계를 갖는 키워드 및 관련 용어를 증감시키는 학습부(110)를 포함하고 있다.

[0034] 제1실시예에서, 해당 문서 분류 시스템은, 추출부(102)와 문서 표시부(103)와 분류 코드 수신부(104)와 선정부(105)와 탐색부(106)와 점수 산출부(107)와 자동 분류부(108)와 문서 배제부(109)와 학습부(110)를 포함하는 문서 분류 장치(101), 데이터베이스(201) 및 리뷰어가 이용하는 클라이언트 장치(301)로 구성된다. 클라이언트 장치(301)는 하나의 문서 분류 시스템 내에 복수 개를 포함할 수도 있다.

[0035] 문서 분류 장치(101) 및 클라이언트 장치(301)는, 컴퓨터 또는 서버이며, 각종 입력에 따라 CPU가 ROM에 기록된 프로그램을 실행함으로써 각종 기능부로써 동작한다.

[0036] 분류 코드는, 문서를 분류할 때 사용하는 식별자를 말한다. 소송에서 문서 정보를 증거로 이용할 때에, 증거의 종류에 따라 부여할 수도 있다. 제1실시예에서, 분류 코드로서 이번 소송에서 증거 능력이 없는 문서를 나타내는 「무관」, 증거로 제출할 필요가 있음을 나타내는 「유관」, 및 제품 A와 특히 관련이 있는 문서임을 나타내는 「중요」의 3 가지 코드를 포함하고 있으며, 이 중 「중요」 코드가 부여되는 문서를 분류하는 것이다.

[0037] 여기서 말하는 문서는, 소송에서 증거로 제출하는 디지털 정보로, 하나 이상의 단어가 포함된 데이터를 말한다. 예를 들어, 이메일, 프레젠테이션 자료, 표 계산 자료, 협의 자료, 계약서, 조직도, 사업 계획서 등이 있다. 또한, 스캔 데이터를 문서로 취급할 수도 있다. 이 경우 스캔 데이터를 텍스트 데이터로 변환할 수 있도록 문서 분류 시스템 내에 OCR(Optical Character Reader) 장치를 포함할 수 있다. OCR 장치에 의해 텍스트 데이터로 변경하여 스캔 데이터에서 키워드 및 관련 용어를 해석하고 탐색할 수 있다.

[0038] 예를 들어, 제1실시예에서는, 제품 A에 관한 협의에 대한 내용이 기재된 의사록이나 이메일 등에 「유관」 코드가 부여되고, 제품 A의 개발 계획서나 설계서 등에 「중요」 코드가 부여되고, 제품 A와 무관한 정례회 등의 자료에 「무관」 코드가 부여된다.

[0039] 또한, 키워드는 어떤 언어에서 일정한 의미를 가진 문자열의 정리를 말한다. 예를 들어, 「문서를 분류한다」는 문장에서 키워드를 선정하면 「문서」 「분류」로 될 수 있다. 제1실시예에서는, 「침해」나 「소송」, 「특허공보○○호」라는 키워드가 중점적으로 선정된다.

[0040] 데이터베이스(201)는 전자 매체에 데이터를 기록하는 기록 장치이며, 문서 분류 장치(101)의 내부에 있을 수도 있으며, 예를 들어 저장 장치로 외부에 설치할 수도 있다.

[0041] 문서 분류 장치(101), 데이터베이스(201) 및 클라이언트 장치(301)는, 유선 또는 무선 네트워크를 통해 접속되어 있다. 클라우드 컴퓨팅의 형태로 이용할 수도 있다.

[0042] 데이터베이스(201)는 각 분류 코드에 대한 키워드를 기록하고 있는 것이다. 또한, 과거의 분류 처리의 결과에서 제품 A와 관련성이 높은 문서 중에 포함되면 즉시 「중요」 코드를 부여한다고 판단할 수 있는 키워드를 사전(事前)에 등록할 수 있다. 예를 들어 제품 A의 주요 기능 이름이나, 「소송」, 「경고」, 「특허공보」와 같은 키워드이다. 또한, 마찬가지로 과거의 분류 처리의 결과에서 제품 A와 관련성이 높기 때문에 「중요」 코드가 부여된 문서 그룹과 관련성이 높은 일반 용어를 추출하여, 관련 용어로 등록할 수도 있다.

[0043] 일단 데이터베이스(201)에 등록된 키워드 및 관련 용어는, 학습부(110)에 의한 학습 결과에 따라 증감되는 것 이외에, 수동으로도 추가 등록 및 삭제가 가능하다.

[0044] 추출부(102)는 문서 정보에서 문서 그룹을 추출할 때, 무작위로 샘플링을 할 수 있다. 제1실시예에서는 전체 문서 정보 중 20%의 문서를 무작위로 추출하여 리뷰어에 의한 분류 대상으로 한다. 추출부(102)가 전체 문서 정보

에서 추출할 문서의 비율은 자유롭게 설정할 수도 있다. 또한, 추출원이 될 대상을 전체 문서 정보의 일부로 할 수도 있다.

[0045] 문서 표시부(103)는 클라이언트 장치(301)에 대하여, 도 14에 도시된 바와 같은 문서 표시 화면(I1)을 제시한다. 문서 표시 화면(I1)은, 도 14과 같이 중앙에 분류 대상이 되는 문서를 표시하고, 좌측에 분류 코드를 표시하도록, 화면 구조에서 한 화면 내에 분류 대상 문서와 부여할 분류 코드를 표시할 수도 있다. 문서를 표시하는 부분과 분류 코드를 표시하는 부분이 각각 다른 화면이 되는 화면 구조로 할 수도 있다.

[0046] 제1실시예에서, 문서 표시 화면(I1) 내의 분류 코드 1은 「무관」 코드, 분류 코드 2는 「유관」 코드 및 분류 코드 3은 「중요」 코드를 의미한다. 또한, 「유관」 코드를 부여받은 문서 중, 소분류 1은 제품 A의 가격과 관계가 있는 문서에 대해 부여되고, 소분류 2는 제품 A의 개발 일정과 관계가 있는 문서에 대해 부여되는 것이다. 소분류는 하나의 분류 코드에 복수로 구비할 수 있으며, 구비하지 못할 수도 있다.

[0047] 분류 코드 수신부(104)는 문서 표시부(103)가 표시한 문서 정보 중, 리뷰어가 눈으로 확인하고, 하나하나 분류 코드를 결정한 문서에 대해, 해당 결정에 따라 분류 코드를 부여하고, 해당 문서를 분류할 수 있다. 문서의 분류는 부여된 분류 코드에 의해 수행할 수 있다.

[0048] 선정부(105)는 분류 코드 수신부(104)가 분류한 문서 정보를 해석하고, 「무관」, 「유관」 및 「중요」의 각각의 분류 코드를 부여받은 문서 정보에서, 공통으로 빈출하는 키워드를 해당 분류 코드의 키워드로 선정한다.

[0049] 도 2는 선정부(105)에서 「중요」 코드가 부여된 문서를 분석한 결과를 나타낸 그래프이다.

[0050] 도 2에서, 세로축(R_hot)은, 리뷰어에 의해 「중요」 코드가 부여된 모든 문서 중, 「중요」 코드에 결부된 키워드로 선정된 키워드를 포함하며 또한「중요」 코드가 부여된 문서의 비율을 나타내고 있다. 가로축은 리뷰어가 분류 처리를 실시한 모든 문서 중, 선정부(105)에 의해 선정된 키워드를 포함하는 문서의 비율을 나타내고 있다.

[0051] 제1실시예에서, 선정부(105)는, 직선 R_hot = R_all 보다 상부에 나타날 수 있게 되는 키워드를 그 분류 코드의 키워드로 선정할 수 있다.

[0052] 탐색부(106)는, 대상이 되는 문서 중에서 특정 키워드를 탐색하는 기능을 갖는다. 탐색부(106)는 선정부(105)에서 선정된 키워드 또는 데이터베이스(201)에서 추출된 관련 용어를 포함하는 문서를, 탐색할 때에, 분류 코드 수신부(104)에 의해 분류 코드의 부여를 수신하지 않은 문서로 구성되는 문서 그룹을 대상으로서 탐색하는 것이다.

[0053] 점수 산출부(107)는, 문서 그룹 중에 출현하는 키워드와 각 키워드가 갖는 가중치에 따라 다음 식에서 점수를 산출할 수 있다. 점수는 어떤 문서에서 특정한 분류 코드와의 결부 강도를 정량적으로 평가하는 것을 말한다.

[0054] $$Scr = \sum_{i=0}^{N} i*(m_i * wgt_i^2) / \sum_{i=0}^{N} i * wgt_i^2 \quad \text{- - - (1)}$$

[0055] m_i : i번째 키워드 또는 관련 용어의 출현 빈도

[0056] wgt_i^2 : i번째 키워드 또는 관련 용어의 가중치

[0057] 자동 분류부(108)는 산출된 점수에 기초하여 문서 정보에 자동으로 분류 코드를 부여할 때에, 분류 코드 수신부(104)에서 분류 코드의 부여를 수신하지 않은 문서를 추출하고, 해당 문서에 대하여 자동으로 분류 코드를 부여하는 기능을 가질 수도 있다.

[0058] 문서 배제부(109)는 분류 대상이 되는 문서 정보 중, 데이터베이스(201)에서 사전(事前)에 등록된 키워드 및 관련 용어, 및 선정부(105)에서 선정된 키워드의 어느 것도 포함하지 않는 문서를 탐색하고, 해당 문서를 분류 대상에서 사전(事前)에 배제할 수 있다.

[0059] 학습부(110)는 분류 처리의 결과를 바탕으로 각 키워드의 가중치를 학습하고, 해당 학습 결과를 토대로 데이터베이스(201)에 등록되어있는 키워드 및 관련 용어의 증감을 수행한다.

[0060] 각 키워드가 갖는 가중치는, 해당 키워드가 갖는 각 분류 코드에 의한 전달 정보량을 바탕으로 결정할 수도 있다. 해당 가중치는 다음 식에서, 분류 처리를 거듭할 때마다 학습하여 정확도를 향상시킬 수 있다.

[0061] $$wgt_{i,L}=\sqrt{wgt_{L-i}^{2}+\gamma_{L}\,wgt_{L-i}^{2}-\vartheta}=\sqrt{wgt_{i,L}^{2}+\sum_{i=1}^{L}\gamma_{L}\,wgt_{i,L}^{2}-\vartheta} \quad \text{- - - (2)}$$

[0062] $wgt_{i,0}$: 학습 전의 i 번째 선정 키워드의 가중치 (초기값)

[0063] $wgt_{i,L}$: L회 학습 후의 i번째 선정 키워드의 가중치

[0064] γ_{L} : L회 학습에서의 학습 파라미터

[0065] ϑ : 학습 효과의 임계값

[0066] 또한, 학습부에서는 신경망을 이용하여 분류 결과를 가중치에 반영시키는 학습 방법을 취하는 것도 가능하다.

[0067] 클라이언트 장치(301)는 리뷰어가 조작하며, 문서 정보를 확인하여 부여하는 분류 코드를 결정하는 데 사용하는 장치이다.

[0068] 제1실시예에서는, 도 3에 도시한 흐름도에 따라 5개의 단계로 분류 처리를 한다.

[0069] 제1단계에서는, 과거의 분류 처리의 결과를 이용하여 키워드와 관련 용어의 사전(事前) 등록을 수행한다. 이때 등록된 키워드는 제품 A의 침해 행위로 되는 기능의 명칭이나 기술의 명칭 등, 문서 내에 포함되면 즉시 「중요」 코드가 부여되는 키워드이다.

[0070] 제2단계에서는, 제1단계에서 등록된 키워드를 포함하는 문서를 전체 문서 정보에서 탐색하고, 해당 문서를 발견하면 「중요」 코드를 부여한다.

[0071] 제3단계에서는, 제1단계에서 등록된 관련 용어를 전체 문서 정보에서 검색하고, 해당 관련 용어를 포함하는 문서의 점수를 산출하고, 분류를 수행한다.

[0072] 제4단계에서는, 리뷰어에 의한 분류 코드의 결정을 실시한 후에, 리뷰어가 분류한 규칙성을 바탕으로 자동으로 분류 코드의 부여를 수행한다.

[0073] 제5단계에서는, 제1단계 내지 제4단계의 결과를 이용하여 학습을 수행한다.

[0074] <제1단계>

[0075] 제1단계에서 데이터베이스(201)의 처리 흐름을, 도 4를 이용하여 상세하게 설명한다. 데이터베이스(201)에서 몇 번째 단계의 처리를 할 것인지를 판단하고, 제1단계의 처리를 선택한다(스텝 1: 제1단계). 이 단계에서는 먼저 데이터베이스(201)에서 키워드를 사전(事前) 등록한다(스텝 2). 이때 등록되는 것은, 과거의 분류 처리의 결과에서 제품 A와 관련성이 높고 문서 내에 포함되면 즉시 「중요」 코드를 부여한다고 판단할 수 있는 키워드이다. 또한, 마찬가지로 과거의 분류 처리의 결과에서 제품 A와 관련성이 높기 때문에 「중요」 코드가 부여된 문서 그룹과 관련성이 높은 일반 용어를 추출하여(스텝 3), 관련 용어로 등록한다(스텝 4).

[0076] <제2단계>

[0077] 제2단계에서 데이터베이스(201), 탐색부(106) 및 자동 분류부(108)의 처리 흐름을 도 4, 도 5 및 도 7을 이용하여 상세하게 설명한다.

[0078] 데이터베이스(201)에서 몇 번째 단계의 처리를 할 것인지를 판단하고, 제2단계의 처리를 선택한다(스텝 1: 제2단계). 데이터베이스(201)에서 다시 사전(事前)에 등록해 둘 필요가 있는 키워드가 있는 경우(스텝 5: 예), 추가 등록을 실시한다(스텝 6). 추가로 등록할 키워드가 없는 경우(스텝 5: 아니오) 및 스텝 6의 처리 완료 후, 탐색부(106)에서 몇 번째 단계의 처리를 할 것인지를 판단하고, 제2단계의 처리를 선택한다(스텝 11: 제2단계). 이 단계에서 탐색부(106)는 먼저 데이터베이스(201)에 제1단계 및 제2단계에서 사전(事前) 등록된 키워드가 있

는지를 판정한다(스텝 12). 사전(事前)에 등록된 키워드가 존재하지 않는 경우(스텝 12: 아니오), 제2단계의 처리는 종료한다.

[0079] 사전(事前)에 등록된 키워드가 존재하는 경우(스텝 12: 예), 분류 대상이 되는 문서 정보 속에 해당 키워드를 포함하는 문서가 없는지 분류 대상이 되는 전체 문서 정보에 대해 탐색을 실시한다(스텝 13). 탐색된 키워드가 포함된 문서가 존재하지 않는 경우(스텝 14: 아니오), 제2단계의 처리를 종료한다. 한편, 탐색된 키워드가 포함된 문서를 발견한 경우(스텝 14: 예), 자동 분류부(108)에 통지한다(스텝 15).

[0080] 자동 분류부(108)에서는, 탐색부(106)로부터 해당 통지를 받은 경우(스텝 29: 제2단계, 스텝 30: 예), 해당 통지의 대상이 된 문서에 대하여 「중요」 코드를 부여하고, 처리를 종료한다. 탐색부(106)에서 해당 통지를 받지 않은 경우(스텝 29: 제2단계, 스텝 30: 아니오), 아무 처리도 하지 않는다.

[0081] <제3단계>

[0082] 제3단계에서 데이터베이스(201), 탐색부(106), 점수 산출부(107) 및 자동 분류부(108)의 처리 흐름을, 도 4, 도 5, 도 6 및 도 7을 이용하여 상세하게 설명한다.

[0083] 데이터베이스(201)에서 몇 번째 단계의 처리를 할 것인지를 판단하고, 제3단계의 처리를 선택한다(스텝 1: 제3단계). 데이터베이스(201)에서 다시 사전(事前)에 등록해 둘 필요가 있는 관련 용어가 있는 경우(스텝 7: 예) 추가 등록을 실시한다(스텝 8). 관련 용어의 추가 등록이 필요없는 경우(스텝 7: 아니오), 제3단계의 처리를 종료한다.

[0084] 스텝 8의 처리 완료 후, 탐색부(106)에서 몇 번째 단계의 처리를 할 것인지를 판단하고, 제3단계의 처리를 선택한다(스텝 11: 제3단계). 이 단계에서 탐색부(106)는, 데이터베이스(201) 내에 제1단계 및 제2단계에서 등록된 관련 용어가 있는지를 판정한다(스텝 16). 사전(事前)에 등록된 키워드가 존재하지 않는 경우(스텝 16: 아니오), 제3단계의 처리는 종료한다.

[0085] 관련 용어가 존재하는 경우(스텝 16: 예), 분류 대상이 되는 문서 정보 속에 해당 관련 용어를 포함하는 문서가 없는지 분류 대상이 되는 전체 문서 정보에 대해 탐색을 실시한다(스텝 17). 탐색한 키워드가 포함된 문서가 존재하지 않는 경우(스텝 18: 아니오), 제3단계의 처리를 종료한다. 한편, 탐색한 관련 용어를 포함하는 문서를 발견한 경우(스텝 18: 예), 점수 산출부(107)에 통지한다(스텝 19).

[0086] 점수 산출부(107)는, 탐색부(106)로부터 해당 통지를 받은 경우(스텝 24: 제3단계, 스텝 23: 예), 위의 식 (1)을 사용하여 문서 속에서 발견한 관련 용어의 종류와 해당 관련 용어가 갖는 가중치에서 각 문서의 점수를 산출하여, 자동 분류부(108)에 통지한다(스텝 28). 탐색부(106)에서 관련 용어를 발견한 통지를 받지 않은 경우(스텝 24: 제3단계, 스텝 23: 아니오), 제3단계의 처리를 종료한다.

[0087] 자동 분류부(108)는, 점수 산출부(107)로부터 점수의 통지를 받은 경우(스텝 29: 제3단계, 스텝 32: 예), 점수가 임계값을 초과했는지의 판정을 문서별로 수행하고, 점수가 임계값을 초과한 문서에는 「중요」 코드를 부여하고, 점수가 임계값을 초과한 문서가 없는 경우는 부여하지 않고 처리를 종료한다(스텝 33).

[0088] <제4단계>

[0089] 제3단계에서 데이터베이스(201), 탐색부(106), 점수 산출부(107), 자동 분류부(108), 추출부(102), 문서 표시부(103), 분류 코드 수신부(104) 및 선정부(105)의 처리 흐름을 각각 도 4, 도 5, 도 6, 도 7, 도 8, 도 9, 도 10 및 도 11을 이용하여 상세하게 설명한다.

[0090] 제4단계에서는, 먼저 추출부(102)에서, 분류 대상이 되는 문서 정보에서 무작위로 문서를 샘플링하여, 리뷰어가 수동으로 분류 코드를 부여하는 대상이 되는 문서 그룹을 추출한다(스텝 34). 문서 표시부(103)에서 추출된 문서 그룹을 문서 표시 화면(I1) 상에 표시한다(스텝 35).

[0091] 리뷰어는 문서 표시화면(I1)에 표시된 문서 그룹에 대해, 각 문서의 내용을 읽고 나서, 제품 A와 해당 문서의 내용 사이에 관련성이 있는지 여부를 판단하고, 「중요」 코드를 부여할지 여부를 결정한다. 리뷰어가 「중요」 코드를 부여하는 문서란, 예를 들어, 제품 A의 선행 기술을 조사한 결과 보고서라든지, 제품 A의 제조는 특허침해라고 타인으로부터 경고받은 경고장 등이다.

[0092] 리뷰어에 의해 부여된 분류 코드는 분류 코드 수신부(104)에 의해 수신되고(스텝 36), 부여된 분류 기호에 따라 문서가 분류된다(스텝 37).

[0093] 선정부(105)는 스텝 37에서 분류된 각 문서에 대해 키워드 해석을 수행하고(스텝 38), 「중요」 코드를 부여받은 문서에 공통으로 출현 횟수가 많은 키워드를 선정한다(스텝 39).

[0094] 그런 다음 데이터베이스(201)는, 스텝 39에서 선정부(105)가 선정한 키워드가 제품 A와 관계가 있음을 나타내는 「중요」 코드에 관한 키워드로 미등록인 경우(스텝 1: 제3단계, 스텝 9: 예), 해당 키워드의 등록을 실시한다. 해당 키워드가 이미 등록된 경우, 아무 처리도 수행하지 않는다(스텝 1: 제3 단계, 스텝 9: 아니오).

[0095] 탐색부(106)는, 「중요」 코드에 관한 키워드가 데이터베이스(201)에 등록되어 있지 않은 경우(스텝 20: 아니요), 제4단계의 처리를 종료한다. 해당 키워드가 등록되어있는 경우(스텝 20: 예), 추출부(102)에서 추출된 리뷰어에 의해 분류된 문서를 탐색 대상에서 빼고, 나머지 각 문서를 대상으로, 해당 키워드 탐색을 실행한다(스텝 21). 해당 탐색에서 문서 중에서 키워드를 발견한 경우(스텝 22: 예), 점수 산출부(107)에 통지한다(스텝 23).

[0096] 점수 산출부(107)는, 키워드 발견의 통지를 받은 경우(스텝 27: 예) 위의 식 (1)을 이용하여 각 문서에 대한 점수를 산출하고, 자동 분류부에 통지한다.

[0097] 자동 분류부(108)는, 점수 산출부(107)로부터 통지를 받으면(스텝 32: 예), 문서별로 점수가 임계값을 초과했는지의 판정을 실시하고, 임계값을 초과한 문서에는 「중요」 코드를 부여하고, 초과하지 않은 문서는 부여하지 않고 처리를 종료한다 (스텝 33).

[0098] <제5단계>

[0099] 제5단계에서 문서 배제부(109) 및 학습부(110)의 처리 흐름을 각각 도 12 및 도 13을 이용하여 설명한다.

[0100] 문서 배제부(109)에서, 분류 대상이 되는 문서 정보 중, 제1 내지 제4 단계의 처리가 미실시인 문서 그룹에 대하여, 제1, 제2 단계에서 사전(事前)에 등록된 키워드, 제1, 제3 단계에서 등록된 관련 용어 및 제4 단계에서 등록된 키워드를 포함하는 문서가 있는지 여부를 탐색하고, 어느 것도 발견되지 않은 문서가 있는 경우(스텝 40: 예), 해당 문서를 분류 대상에서 사전(事前)에 배제한다(스텝 41).

[0101] 학습부(110)는, 제1 내지 제4의 처리 결과를 바탕으로, 각 키워드의 가중치를 식 (2)에 의해 학습한다. 해당 학습 결과를 데이터베이스(201)에 반영한다(스텝 42).

[0102] [기타 실시예]

[0103] 본 발명의 다른 실시예를 설명한다.

[0104] 제1실시예에서는, 특히 특허 침해 소송 사건에서의 실시예를 설명하였으나, 본 발명에서 문서 분류 시스템은, 카르텔이나 독점 금지법 등, 전자 증거 개시(eDiscovery) 제도를 채택하고 있으며, 문서 제출 의무가 있는 모든 소송에서 이용할 수 있다.

[0105] 또한, 제1실시예에서, 리뷰어가 분류한 규칙성을 바탕으로 자동으로 분류 코드를 부여하는 제4단계의 처리를, 제1단계 내지 제3단계의 처리 후에 실시하였지만, 제1단계 내지 제3단계의 처리를 하지 않고, 제4단계의 처리만 단독으로 수행할 수도 있다.

[0106] 또한, 최초에 추출부(102)에 의해 문서 정보에서 일부 문서 그룹을 추출하고, 해당 추출한 문서 그룹에 대해 먼저 제4단계의 처리를 최초로 실시한다. 그 후, 제4단계에서 등록한 키워드를 바탕으로 제1단계 내지 제3단계의 처리를 수행하는 실시예일 수도 있다.

[0107] 탐색부(106)에서, 제1실시예의 제4단계에서는, 분류 코드 수신부(104)에서, 분류 코드가 수신되지 않은 문서에 대하여 선정부(105)가 선정한 키워드 검색을 실시하였지만, 전체 문서 정보를 대상으로 해당 키워드 검색을 수행할 수 있다.

[0108] 자동 분류부(108)에서 제1실시예의 제4단계에서는, 분류 코드 수신부(104)에서 분류 코드가 수신되지 않은 문서만을 분류 코드의 자동 부여의 대상으로 하였지만, 전체 문서 정보를 해당 자동 부여의 대상으로 할 수 있다.

[0109] 본 발명에 따른 문서 분류 시스템, 문서 분류 방법 및 문서 분류 프로그램은, 문서 정보로부터 소정 수의 문서를 포함하는 데이터 세트인 문서 그룹을 추출하고, 추출된 문서 그룹을 화면 상에 표시하고, 표시된 문서 그룹에 대해 리뷰어가 소송과의 관련성에 기초하여 부여한 분류 코드를 수신하고, 해당 분류 코드에 기초하여 추출된 문서 그룹을 분류 코드별로 분류하고, 상기 분류된 문서 그룹에서 공통으로 출현하는 키워드를 해석하여 선정하고, 선정된 키워드를 기록하고, 기록된 키워드를 문서 정보에서 탐색하고, 탐색 결과와 해석 결과를 이용하

여 분류 코드와 문서와의 관련성을 나타내는 점수를 산출하고, 점수 결과에 따라 자동으로 분류 코드를 부여함으로써, 리뷰어의 분류 작업 노력의 경감을 도모할 수 있다.

[0110] 또한, 본 발명의 문서 분류 시스템에서, 탐색부는 키워드를 분류 코드가 부여되지 않은 문서로 구성된 문서 정보에서 탐색하는 기능을 갖고, 점수 산출부는 탐색부의 탐색 결과와 선정부의 해석 결과를 이용하여 분류 코드와 문서와의 관련성을 나타내는 점수를 산출하고, 자동 분류부는 분류 코드 수신부에서 분류 코드의 부여를 수신하지 않은 문서를 추출하고 해당 문서에 대해 자동으로 분류 코드를 부여하는 기능을 구비할 때, 분류 코드 수신부에서 분류 코드의 부여를 수신하지 않은 문서 정보에 대해, 리뷰어가 분류한 규칙성을 바탕으로 자동으로 분류 코드를 부여할 수 있다.

[0111] 또한, 본 발명은, 선정부의 분석 결과와, 점수 산출부가 산출한 점수에 기초하여 선정부가 선정한 데이터베이스에 기록된 분류 코드와의 상관 관계를 갖는 키워드 및 관련 용어를 증감시키는 학습부를 구비할 때, 분류 회수를 거듭할 때마다 분류 정밀도를 향상시킬 수 있다.

[0112] 또한, 본 발명은, 데이터베이스가 분류 코드와 관련성이 있는 용어를 추출 및 기록하고, 탐색부가 관련 용어를 문서 정보에서 탐색하고, 점수 산출부는 탐색부가 관련 용어를 탐색한 결과를 바탕으로 점수를 산출하고, 자동 분류부가 관련 용어를 이용하여 산출한 점수에 기초하여 자동으로 분류 코드를 부여하는 것으로, 문서 그룹에 포함된 문서 중 선정부가 선정한 키워드, 관련 용어 및 분류 코드와 상관 관계가 있는 키워드를 포함하지 않는 문서를 선정하고, 자동 분류부의 분류 대상에서 선정된 문서를 배제할 때, 문서 분류를 보다 효율적으로 수행할 수 있다. 이것은 수집된 디지털 정보의 소송에서의 이용을 용이하게 한다.

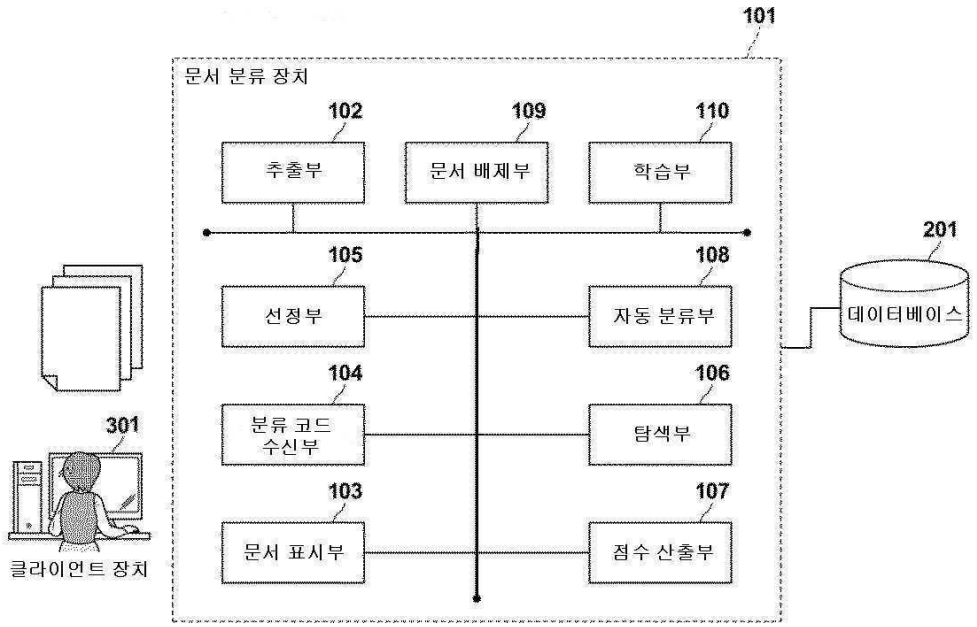
부호의 설명

[0113]

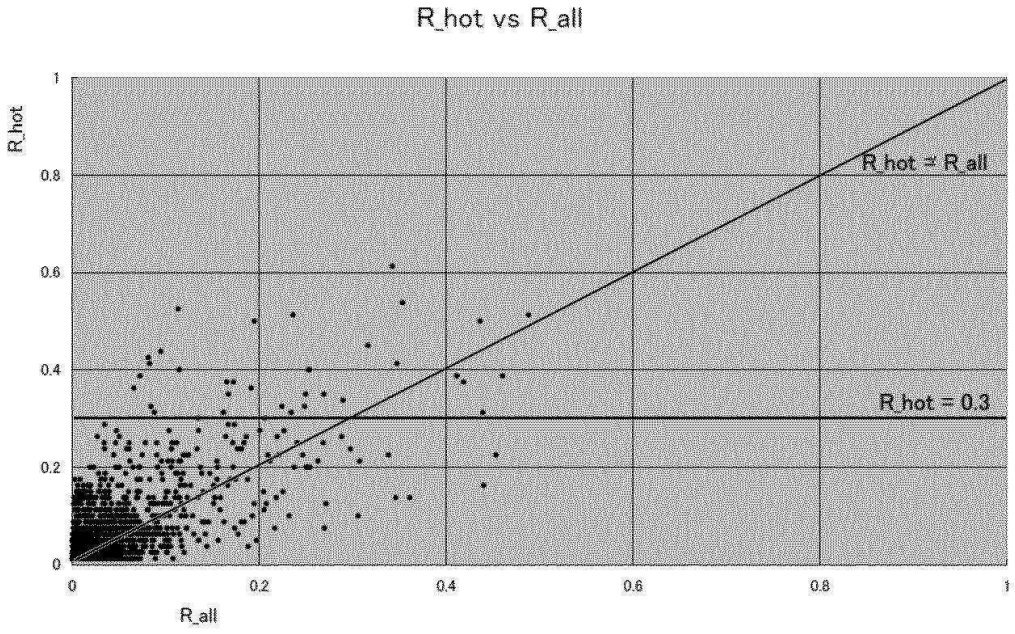
101 문서 분류 장치	102 추출부
103 문서 표시부	104 분류 코드 수신부
105 선정부	106 탐색부
107 점수 산출부	108 자동 분류부
109 문서 배제부	110 학습부
201 데이터베이스	301 클라이언트 장치
I1 문서 표시 화면	

도면

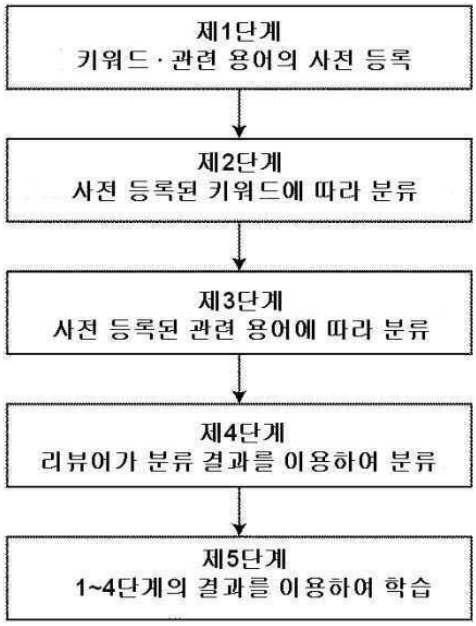
도면1


101
문서 분류 장치
102
추출부
109
문서 배제부
110
학습부
105
선정부
108
자동 분류부
104
분류 코드 수신부
106
탐색부
103
문서 표시부
107
점수 산출부
201
데이터베이스
301
클라이언트 장치

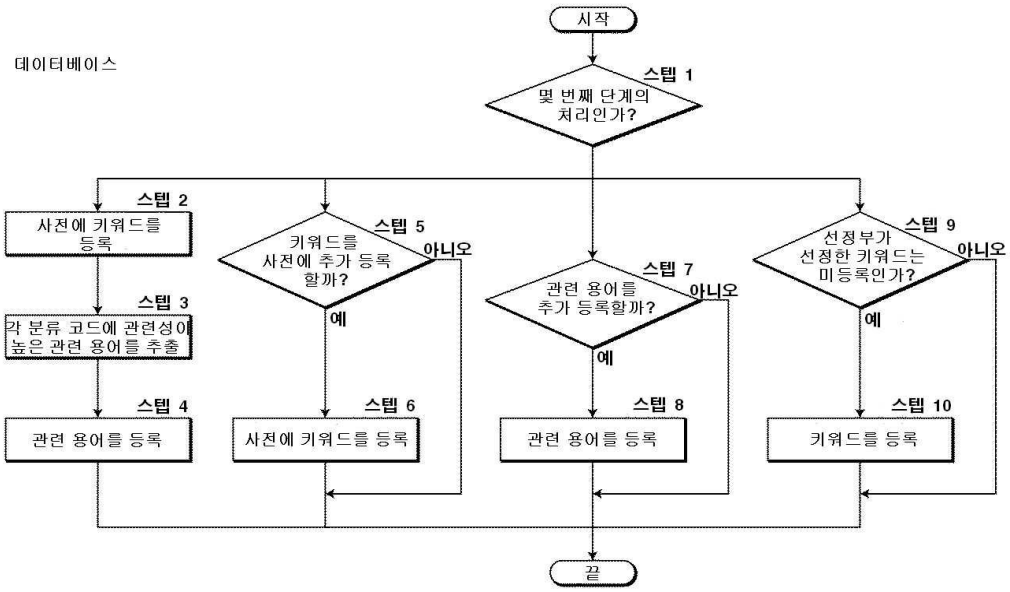
도면2


R_hot vs R_all
R_hot
R_all
R_hot = R_all
R_hot = 0.3
0
0.2
0.4
0.6
0.8
1

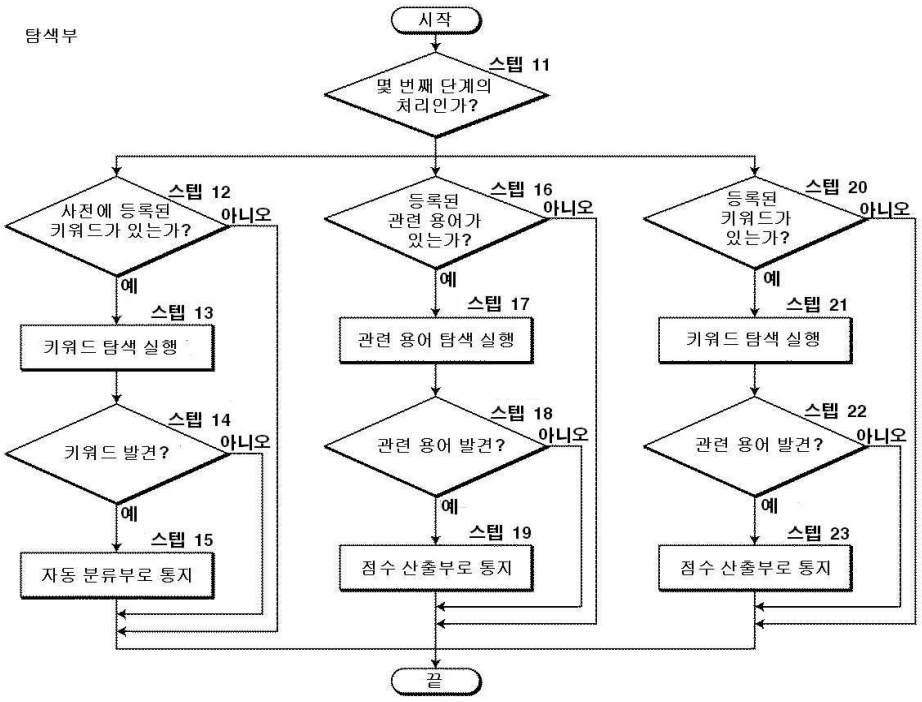
도면3


제1단계
키워드·관련 용어의 사전 등록
제2단계
사전 등록된 키워드에 따라 분류
제3단계
사전 등록된 관련 용어에 따라 분류
제4단계
리뷰어가 분류 결과를 이용하여 분류
제5단계
1~4단계의 결과를 이용하여 학습

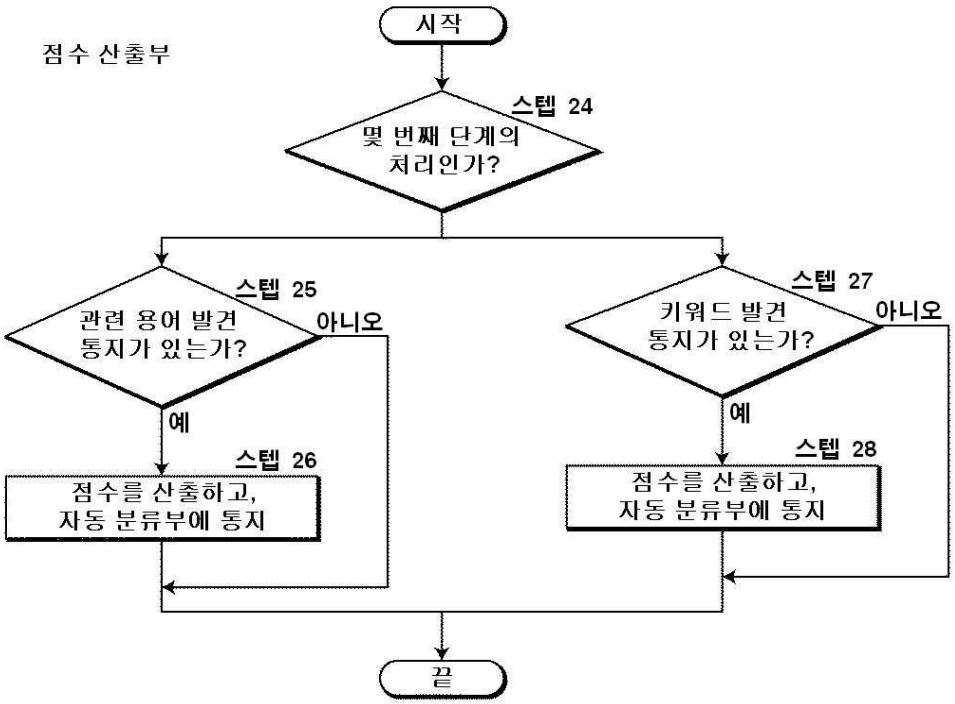
도면4


데이터베이스
시작
스텝 1
몇 번째 단계의 처리인가?
스텝 2
사전에 키워드를 등록
스텝 3
각 분류 코드에 관련성이 높은 관련 용어를 추출
스텝 4
관련 용어를 등록
스텝 5
키워드를 사전에 추가 등록할까?
아니오
예
스텝 6
사전에 키워드를 등록
스텝 7
관련 용어를 추가 등록할까?
아니오
예
스텝 8
관련 용어를 등록
스텝 9
선정부가 선정한 키워드는 미등록인가?
아니오
예
스텝 10
키워드를 등록
끝

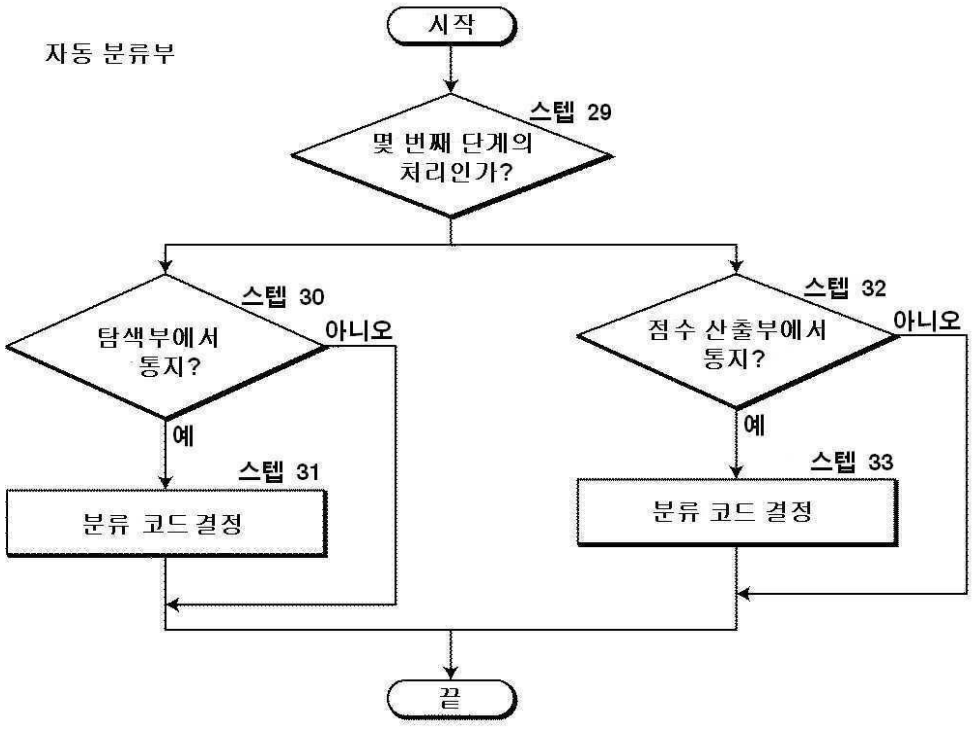
도면5


탐색부
시작
스텝 11
몇 번째 단계의 처리인가?
스텝 12
사전에 등록된 키워드가 있는가?
아니오
예
스텝 13
키워드 탐색 실행
스텝 14
키워드 발견?
아니오
예
스텝 15
자동 분류부로 통지
스텝 16
등록된 관련 용어가 있는가?
아니오
예
스텝 17
관련 용어 탐색 실행
스텝 18
관련 용어 발견?
아니오
예
스텝 19
점수 산출부로 통지
스텝 20
등록된 키워드가 있는가?
아니오
예
스텝 21
키워드 탐색 실행
스텝 22
관련 용어 발견?
아니오
예
스텝 23
점수 산출부로 통지
끝

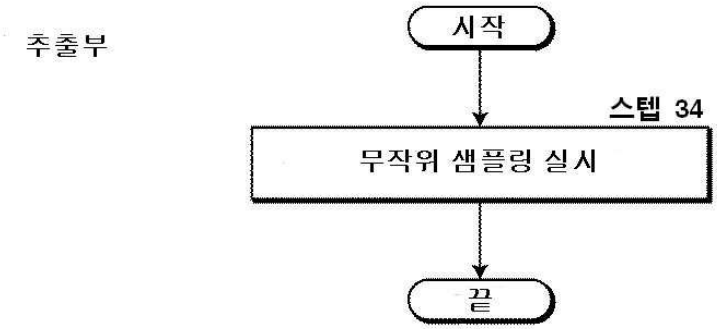
도면6


점수 산출부
시작
스텝 24
몇 번째 단계의 처리인가?
스텝 25
관련 용어 발견 통지가 있는가?
아니오
예
스텝 26
점수를 산출하고, 자동 분류부에 통지
스텝 27
키워드 발견 통지가 있는가?
아니오
예
스텝 28
점수를 산출하고, 자동 분류부에 통지
끝

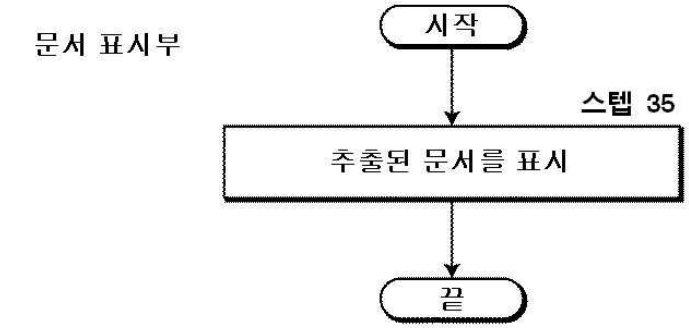
도면7


자동 분류부
시작
스텝 29
몇 번째 단계의 처리인가?
스텝 30
탐색부에서 통지?
아니오
예
스텝 31
분류 코드 결정
스텝 32
점수 산출부에서 통지?
아니오
예
스텝 33
분류 코드 결정
끝

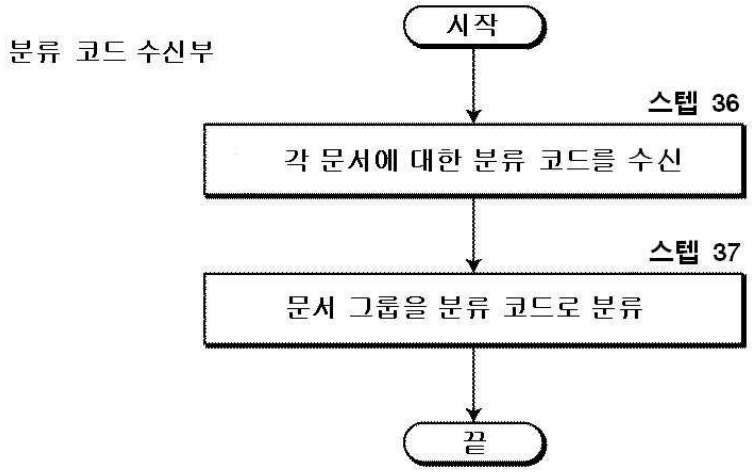
도면8


추출부
시작
스텝 34
무작위 샘플링 실시
끝

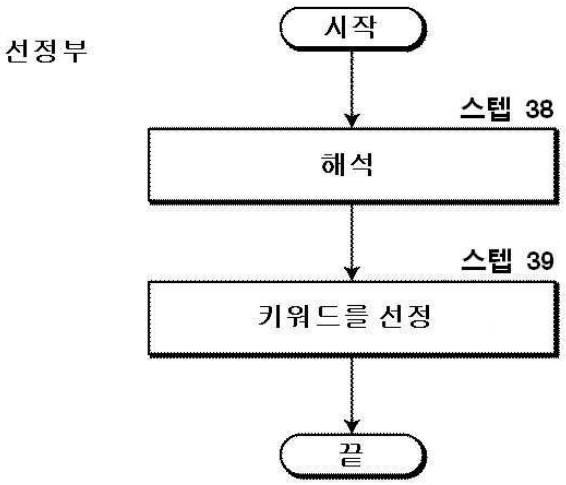
도면9


문서 표시부
시작
스텝 35
추출된 문서를 표시
끝

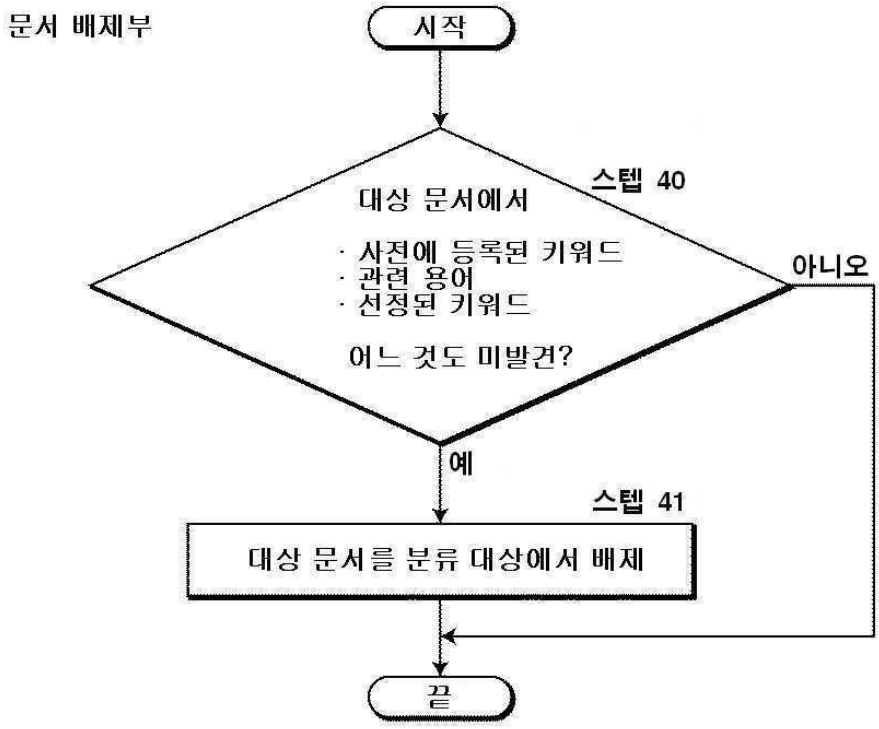
도면10


분류 코드 수신부
시작
스텝 36
각 문서에 대한 분류 코드를 수신
스텝 37
문서 그룹을 분류 코드로 분류
끝

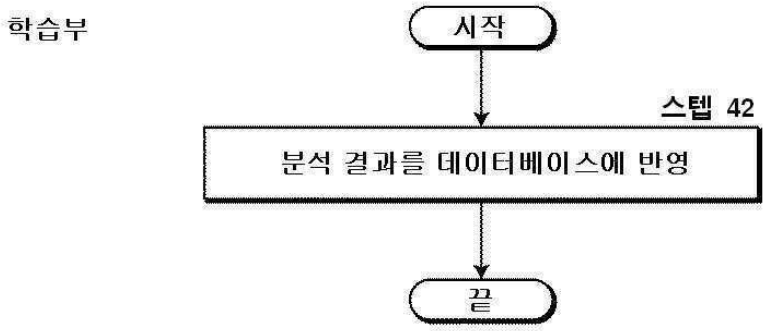
도면11


선정부
시작
스텝 38
해석
스텝 39
키워드를 선정
끝

도면12


문서 배제부
시작
스텝 40
대상 문서에서
· 사전에 등록된 키워드
· 관련 용어
· 선정된 키워드
어느 것도 미발견?
아니오
예
스텝 41
대상 문서를 분류 대상에서 배제
끝

도면13


학습부
시작
스텝 42
분석 결과를 데이터베이스에 반영
끝

도면14