



(19) 대한민국특허청(KR)
(12) 등록특허공보(B1)

(45) 공고일자 2019년08월07일
(11) 등록번호 10-2008387
(24) 등록일자 2019년08월01일

(51) 국제특허분류(Int. Cl.)
G06F 16/00 (2019.01)

(52) CPC특허분류
G06F 16/93 (2019.01)
G06F 2216/11 (2013.01)

(21) 출원번호 10-2018-0050064

(22) 출원일자 2018년04월30일

심사청구일자 2018년04월30일

(56) 선행기술조사문헌

KR1020020038966 A*

(뒷면에 계속)

(73) 특허권자

인하대학교 산학협력단

인천광역시 미추홀구 인하로 100(용현동, 인하대학교)

(72) 발명자

이우기

경기도 군포시 산본로432번길 25, 1222동 1302호
(산본동, 한양목련아파트)

송중수

인천광역시 남구 재념이길 143, 문화네트빌 203동
303호

(뒷면에 계속)

(74) 대리인

특허법인태백

전체 청구항 수 : 총 10 항

심사관 : 권현수

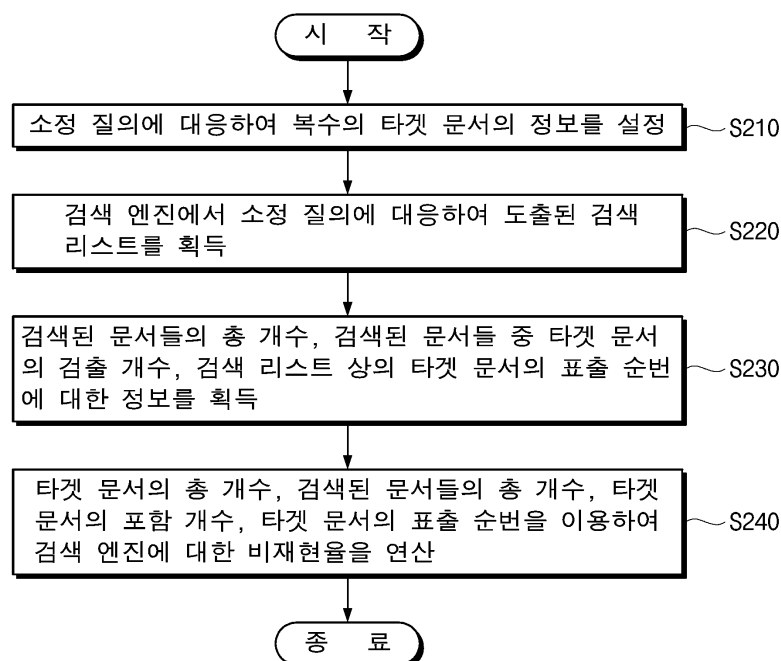
(54) 발명의 명칭 비재현율 기반의 특허 검색 엔진 평가 시스템 및 그 방법

(57) 요약

본 발명은 비재현율 기반의 특허 검색 엔진 평가 시스템 및 그 방법에 관한 것이다. 본 발명에 따르면, 비재현율 기반의 특허 검색 엔진 평가 시스템의 특허 검색 엔진 평가 방법에 있어서, 소정 질의에 대응하는 복수의 타겟 문서의 정보를 설정하는 단계와, 평가 대상이 되는 검색 엔진에서 소정 질의에 대응하여 도출된 검색 리스트를

(뒷면에 계속)

대표도 - 도2



획득하는 단계와, 상기 검색 리스트 내에서 검색된 문서들의 총 개수, 상기 검색된 문서들 중 상기 타겟 문서의 검출 개수, 상기 검색 리스트 상의 상기 타겟 문서의 표출 순번에 대한 정보를 획득하는 단계, 및 상기 설정된 타겟 문서의 총 개수, 상기 검색된 문서들의 총 개수, 상기 타겟 문서의 검출 개수, 상기 타겟 문서의 표출 순번을 이용하여 상기 검색 엔진에 대한 비재현율을 연산하는 단계를 포함하는 비재현율 기반의 특허 검색 엔진 평가 방법을 제공한다.

본 발명에 따른 비재현율 기반의 특허 검색 엔진 평가 시스템 및 그 방법에 따르면, 검색 엔진의 검색 정확도 및 사용자의 검색 시간을 고려하여 검색 엔진의 검색 능력의 척도를 현실적으로 평가할 수 있는 이점이 있다.

(72) 발명자

권현도

경상북도 안동시 강남1길 98-19, 305동 506호 (정하동)

이철기

강원도 속초시 청봉로1길 21, 라동 202호

임지영

인천광역시 서구 청라라임로 40, 134동 1202호 (연희동, 청라골드클래스 커넬웨이)

(56) 선행기술조사문헌

KR1020080096887 A

KR1020120011662 A

JP2013191119 A

KR100295032 B1

*는 심사관에 의하여 인용된 문헌

이 발명을 지원한 국가연구개발사업

과제고유번호 1415153677

부처명 산업통상자원부

연구관리전문기관 한국산업기술평가관리원

연구사업명 산업핵심기술개발사업

연구과제명 지식기반 플랜트 O&M을 위한 지식베이스 프레임워크 통합시스템 개념설계 개발

기 여 율 1/1

주관기관 한국과학기술원

연구기간 2017.10.01 ~ 2018.05.31

공지예외적용 : 있음

명세서

청구범위

청구항 1

비재현율 기반의 특허 검색 엔진 평가 시스템의 특허 검색 엔진 평가 방법에 있어서,
 상기 특허 검색 엔진 평가 시스템은 소정 질의에 대응하는 복수의 타겟 문서의 정보를 설정하는 단계;
 평가 대상이 되는 검색 엔진에서 소정 질의에 대응하여 도출된 검색 리스트를 획득하는 단계;
 상기 검색 리스트 내에서 검색된 문서들의 총 개수, 상기 검색된 문서들 중 상기 타겟 문서의 검출 개수, 상기 검색 리스트 상의 상기 타겟 문서의 표출 순번에 대한 정보를 획득하는 단계; 및
 상기 설정된 타겟 문서의 총 개수, 상기 검색된 문서들의 총 개수, 상기 타겟 문서의 검출 개수, 상기 타겟 문서의 표출 순번을 이용하여 상기 검색 엔진에 대한 비재현율을 연산하는 단계를 포함하며,
 상기 비재현율은,
 아래의 수학적식을 통해 연산되는 비재현율 기반의 특허 검색 엔진 평가 방법.

$$\text{비재현율} = 1 - \frac{\sum_{i=1}^n \beta_i}{\sum_{i=1}^{|r|} \beta_{r_i} + (n - |r|) \cdot \beta_N}$$

여기서, $\beta_x = x^2$ 의 함수 규칙을 따르며, N은 상기 검색된 문서들의 총 개수, n은 상기 설정된 타겟 문서의 총 개수, $|r|$ 는 상기 타겟 문서의 검출 개수, i는 상기 타겟 문서의 인덱스, r_i 는 i번째 타겟 문서의 표출 순번($1 \leq r_i \leq N$)을 나타낸다.

청구항 2

청구항 1에 있어서,
 상기 정보를 획득하는 단계는,
 상기 검색된 문서들과 상기 설정된 복수의 타겟 문서의 정보를 비교하여 상기 검색된 문서들 중 상기 타겟 문서의 검출 개수를 획득하는 비재현율 기반의 특허 검색 엔진 평가 방법.

청구항 3

청구항 1에 있어서,
 상기 비재현율은,
 상기 검색 리스트 내에 포함된 상기 타겟 문서의 표출 순번이 크거나, 상기 설정된 타겟 문서의 총 개수 대비 상기 타겟 문서의 검출 개수가 작을수록 상승하는 비재현율 기반의 특허 검색 엔진 평가 방법.

청구항 4

청구항 3에 있어서,
 상기 비재현율의 연산 시에, 상기 타겟 문서의 표출 순번에 대해 제곱 가중치를 부여하는 비재현율 기반의 특허 검색 엔진 평가 방법.

청구항 5

청구항 3에 있어서,

상기 비재현율은 0과 1 사이의 값을 가지는 비재현율 기반의 특허 검색 엔진 평가 방법.

청구항 6

삭제

청구항 7

소정 질의에 대응하여 복수의 타겟 문서의 정보를 설정하는 정보 설정부;

평가 대상이 되는 검색 엔진에서 소정 질의에 대응하여 도출된 검색 리스트를 획득하는 검색 결과 획득부;

상기 검색 리스트 내에서 검색된 문서들의 총 개수, 상기 검색된 문서들 중 상기 타겟 문서의 검출 개수, 상기 검색 리스트 상의 상기 타겟 문서의 표출 순번에 대한 정보를 획득하는 평가 인자 획득부; 및

상기 설정된 타겟 문서의 총 개수, 상기 검색된 문서들의 총 개수, 상기 타겟 문서의 검출 개수, 상기 타겟 문서의 표출 순번을 이용하여 상기 검색 엔진에 대한 비재현율을 연산하는 비재현율 연산부를 포함하며,

상기 비재현율은,

아래의 수학적식을 통해 연산되는 비재현율 기반의 특허 검색 엔진 평가 시스템:

$$\text{비재현율} = 1 - \frac{\sum_{i=1}^n \beta_i}{\sum_{i=1}^{|r|} \beta_{r_i} + (n - |r|) \cdot \beta_N}$$

여기서, $\beta_x = x^2$ 의 함수 규칙을 따르며, N은 상기 검색된 문서들의 총 개수, n은 상기 설정된 타겟 문서의 총 개수, $|r|$ 는 상기 타겟 문서의 검출 개수, i는 상기 타겟 문서의 인덱스, r_i 는 i번째 타겟 문서의 표출 순번($1 \leq r_i \leq N$)을 나타낸다.

청구항 8

청구항 7에 있어서,

상기 평가 인자 획득부는,

상기 검색된 문서들과 상기 설정된 복수의 타겟 문서의 정보를 비교하여 상기 검색된 문서들 중 상기 타겟 문서의 검출 개수를 획득하는 비재현율 기반의 특허 검색 엔진 평가 시스템.

청구항 9

청구항 7에 있어서,

상기 비재현율은,

상기 검색 리스트 내에 포함된 상기 타겟 문서의 표출 순번이 크거나, 상기 설정된 타겟 문서의 총 개수 대비 상기 타겟 문서의 검출 개수가 작을수록 상승하는 비재현율 기반의 특허 검색 엔진 평가 시스템.

청구항 10

청구항 9에 있어서,

상기 비재현율의 연산 시에, 상기 타겟 문서의 표출 순번에 대해 제공 가중치를 부여하는 비재현율 기반의 특허 검색 엔진 평가 시스템.

청구항 11

청구항 9에 있어서,

상기 비재현율은 0과 1 사이의 값을 가지는 비재현율 기반의 특허 검색 엔진 평가 시스템.

청구항 12

삭제

발명의 설명

기술 분야

[0001] 본 발명은 비재현을 기반의 특허 검색 엔진 평가 시스템 및 그 방법으로서, 검색 시간 및 정확도를 고려하여 검색 엔진의 검색 능력을 평가할 수 있는 비재현을 기반의 특허 검색 엔진 평가 시스템 및 그 방법에 관한 것이다.

배경 기술

[0002] 모든 정보가 디지털 포맷으로 생성/유통/검색되는 환경에서 기하급수적으로 늘어나는 데이터를 효과적으로 처리 및 관리할 수 있는 방안은 중요한 과제이다. 이를 일정 부분 컴퓨터의 도움을 받아 해결하는 대표적인 방법은 코딩 예측(predictive coding) 혹은 TAR(technology-assisted review) 방법이라 일컬으며, 모두 미국 법원과 정부에서도 공식적으로 인정하는 문서 리뷰 기술들이다.

[0003] 이들은 법률 분야의 전자적 증거 수집(eDiscovery), 증거 중심 의학(evidence-based medicine)에서의 체계적 검토(systematic review), 지적재산권 분야에서의 특허 검색(prior-art search or Patent Map) 등에 사용되고 있다.

[0004] 변호사는 법적 소송 하에 있는 대리인의 사건과 관련 있는 모든 증거를 찾아야 하고, 의사는 특정 조건을 충족하는 모든 환자를 검토하여 임상 시험을 위한 코호트(Cohort)를 선택해야 하며, 변리사 혹은 특허 심사관은 특허 출원 시에 기존에 출원된 특허 중 유사한 특허가 있는지 면밀히 조사를 해야 하고, 또한 기업은 제품을 개발 전 본격적인 투자에 앞서서 관련 선행특허 검색에 대한 철저한 검토가 필요하다. 만약 철저한 선행기술 검색이 이루어지지 않아 소송이 발생할 경우 막대한 손실이 발생한다.

[0005] 특허 검색은 특허 출원 당사자, 변리사, 사내 기술사업팀을 비롯한 사용자 혹은 특허 심사관이 특정 특허의 출원, 특허성 판단, 등록 무효를 실시함에 있어서, 관련된 선행 특허를 놓치지 않기 위하여 수행되며, 매우 신중한 검토가 요구된다. 더불어, 검색 과정에서 높은 상호작용 요구와 복잡한 절차를 필히 거치도록 한다.

[0006] 재현을 기반의 질의 검색은 특허 검색, 법률 검색 등에서 관련된 모든 문서를 찾는 데 주안점을 둔다. 그러나, 대부분의 사용자(reviewer)들은 해당 분야에서 전문가이며 이들에게 막대한 분량의 문서 검토를 할당하기에는 엄청나게 고비용이 든다. 즉, 전문가가 수행하는 문제의 특성상 검색 비용이 높으며, 100% 재현율을 달성하지 못하였을 때 발생하는 리스크가 크다.

[0007] 이러한 이유로 재현율을 최대화하는 동시에 검토해야 할 문서의 수를 최소화 하는 것은 매우 중요하다. 이것을 100% 재현율 문제(Total Recall Problem)라 하며, 최근 이를 목적으로 하는 검색 방법에 대한 연구는 이루어지고 있으나, 각각의 검색 시스템을 제대로 평가할 수 있는 방법은 존재하지 않는다.

[0008] 또한, 평가 방법으로는 전통적으로 정보 검색(IR)에서 사용되는 정확도(Precision) 기반 평가 방법들(MAP, nDCG, F-Measure)을 사용하기 때문에 100% 재현율 환경을 평가하기에는 근본적인 한계가 존재한다.

[0009] 더욱이, 기존의 평가 방법들의 경우, 문서 검색 정보를 그대로 사용함으로써 문서의 검색 도출 순위는 달라도 정밀도 내지 재현율이 동일하게 산출되므로, 어떠한 검색 주체가 더욱 효율적으로 검색을 수행하는지 평가할 수 없다. 따라서 동일한 재현율을 가진 검색 시스템이라 하더라도, 검토 시간(검토 비용) 및 정확도를 고려해야 할 필요성이 있음을 알 수 있다.

[0010] 이처럼, 기존의 평가 방법은 높은 재현율 획득에만 초점을 맞추고 있어 명백한 한계점이 따르고, 관련도 높은 문서가 검색 목록의 하위에 위치할 경우 사용자는 검색 결과를 목록 끝까지 모두 확인해야 하므로 사용자의 문서 검색(검토) 시간이 증가하는 문제가 있고 이는 매우 비효율적이다.

[0011] 따라서, 검색의 정확도는 물론, 사용자의 검색 시간 비용까지 단축하는 검색 엔진을 더욱 높이 평가할 수 있는 새로운 평가 기법이 요구된다.

[0012] 본 발명의 배경이 되는 기술은 한국공개특허 제2002-0038966호(2002.05.25 공개)에 개시되어 있다.

발명의 내용

해결하려는 과제

[0013] 본 발명은, 비재현을 기반의 특허 검색 엔진 평가 시스템 및 그 방법에 관한 것으로서, 검색 엔진의 검색 정확도 및 사용자의 문서 검색 시간을 모두 고려하여 검색 엔진의 검색 능력을 더욱 현실적으로 평가할 수 있는 비재현을 기반의 특허 검색 엔진 평가 시스템 및 그 방법을 제공하는데 목적이 있다.

과제의 해결 수단

[0014] 본 발명은, 비재현을 기반의 특허 검색 엔진 평가 시스템의 특허 검색 엔진 평가 방법에 있어서, 소정 질의에 대응하는 복수의 타겟 문서의 정보를 설정하는 단계와, 평가 대상이 되는 검색 엔진에서 소정 질의에 대응하여 도출된 검색 리스트를 획득하는 단계와, 상기 검색 리스트 내에서 검색된 문서들의 총 개수, 상기 검색된 문서들 중 상기 타겟 문서의 검출 개수, 상기 검색 리스트 상의 상기 타겟 문서의 표출 순번에 대한 정보를 획득하는 단계, 및 상기 설정된 타겟 문서의 총 개수, 상기 검색된 문서들의 총 개수, 상기 타겟 문서의 검출 개수, 상기 타겟 문서의 표출 순번을 이용하여 상기 검색 엔진에 대한 비재현율을 연산하는 단계를 포함하는 비재현을 기반의 특허 검색 엔진 평가 방법을 제공한다.

[0015] 또한, 상기 정보를 획득하는 단계는, 상기 검색된 문서들과 상기 설정된 복수의 타겟 문서의 정보를 비교하여 상기 검색된 문서들 중 상기 타겟 문서의 검출 개수를 획득할 수 있다.

[0016] 또한, 상기 비재현율은, 상기 검색 리스트 내에 포함된 상기 타겟 문서의 표출 순번이 크거나, 상기 설정된 타겟 문서의 총 개수 대비 상기 타겟 문서의 검출 개수가 작을수록 상승할 수 있다.

[0017] 또한, 상기 비재현율의 연산 시에, 상기 타겟 문서의 표출 순번에 대해 제곱 가중치를 부여할 수 있다.

[0018] 또한, 상기 비재현율은 0과 1 사이의 값을 가질 수 있다.

[0019] 또한, 상기 비재현율은, 아래의 수학식을 통해 연산될 수 있다.

$$\text{비재현율} = 1 - \frac{\sum_{i=1}^n \beta_i}{\sum_{i=1}^{|r|} \beta_{r_i} + (n - |r|) \cdot \beta_N}$$

[0020]

[0021] 여기서, $\beta_x = x^2$ 의 함수 규칙을 따르며, N은 상기 검색된 문서들의 총 개수, n은 상기 설정된 타겟 문서의 총 개수, $|r|$ 는 상기 타겟 문서의 검출 개수, i는 상기 타겟 문서의 인덱스, r_i 는 i번째 타겟 문서의 표출 순번($1 \leq r_i \leq N$)을 나타낸다.

[0022] 그리고, 본 발명은, 소정 질의에 대응하여 복수의 타겟 문서의 정보를 설정하는 정보 설정부와, 평가 대상이 되는 검색 엔진에서 소정 질의에 대응하여 도출된 검색 리스트를 획득하는 검색 결과 획득부와, 상기 검색 리스트 내에서 검색된 문서들의 총 개수, 상기 검색된 문서들 중 상기 타겟 문서의 검출 개수, 상기 검색 리스트 상의 상기 타겟 문서의 표출 순번에 대한 정보를 획득하는 평가 인자 획득부, 및 상기 설정된 타겟 문서의 총 개수, 상기 검색된 문서들의 총 개수, 상기 타겟 문서의 검출 개수, 상기 타겟 문서의 표출 순번을 이용하여 상기 검색 엔진에 대한 비재현율을 연산하는 비재현율 연산부를 포함하는 비재현을 기반의 특허 검색 엔진 평가 시스템을 제공한다.

발명의 효과

[0023] 본 발명에 따른 비재현을 기반의 특허 검색 엔진 평가 시스템 및 그 방법에 따르면, 검색 엔진의 검색 정확도 및 사용자의 검색 시간을 고려하여 검색 엔진의 검색 능력의 척도를 현실적으로 평가할 수 있는 이점이 있다.

도면의 간단한 설명

[0024] 도 1은 본 발명의 실시예에 따른 비재현을 기반의 특허 검색 엔진 평가 시스템의 구성을 나타낸 도면이다.

도 2는 도 1의 시스템을 이용한 비재현을 기반의 특허 검색 엔진 평가 방법을 설명하는 도면이다.

도 3은 본 발명의 실시예에서 비재현율의 연산 개념을 설명하는 도면이다.

도 4는 도 3에서 도시된 계단형 그래프를 이용한 비재현율 연산 원리를 설명하기 위한 도면이다.

도 5는 도 3에서 각 경우에 대한 비재현율 연산 실시예를 나타낸 도면이다.

발명을 실시하기 위한 구체적인 내용

- [0025] 그러면 첨부한 도면을 참고로 하여 본 발명의 실시 예에 대하여 본 발명이 속하는 기술 분야에서 통상의 지식을 가진 자가 용이하게 실시할 수 있도록 상세히 설명한다.
- [0026] 본 발명은 비재현율 기반의 특허 검색 엔진 평가 시스템으로, 검색 엔진에서 관련 문서의 검색을 누락하지 않는 정도인 재현율(recall)의 개념과는 반대로, 관련 문서의 검색을 누락하는 정도인 비재현율(non-recall)을 기초로 특허 검색 엔진의 검색 능력을 평가할 수 있는 시스템을 제안한다.
- [0027] 일반적으로 검색 엔진에서 관련 문서를 단 하나도 누락하지 않고 검색하는 경우를 완전 재현율(100% 재현율)로 정의한다. 하지만, 관련 문서의 검색을 누락하지는 않았으나 검색 리스트 상에서 말단에 관련 문서가 존재하면 사용자가 검색 리스트의 말단까지 확인해야 하는 노력이 필요하며 이는 사용자의 검색 시간 및 노력을 포함한 검색 비용을 증가시키는 요인이 된다.
- [0028] 따라서 단순한 검색 문서의 누락 여부나 누락 개수만으로 특허 검색 엔진을 평가하는 시스템은 사용자의 검색 노력을 전혀 고려하지 않은 것으로, 본 발명의 실시예는 이러한 사용자의 검색 노력까지 고려한 특허 검색 엔진 평가 시스템을 제안한다.
- [0029] 특히, 본 발명은 재현율과는 반대 개념인 비재현율에 기반한 기술로, 검색 엔진을 통해 소정 질의에 대하여 문서를 검색한 결과, 검색 리스트 내에 관련 문서(이하, 타겟 문서)의 누락 정도가 클수록, 그리고 관련 문서가 말단에 랭킹될수록 비재현율이 상승하도록 설계되어 있다.
- [0030] 따라서, 비재현율이 낮은 검색 엔진일수록 검색 능력이 우수한 것으로 평가할 수 있으며, 사용자의 검색 시간과 노력까지 절감할 수 있는 최적의 검색 엔진을 확인할 수 있다. 예를 들어, 키프리스, 웹스, 윈텔립스 각각에 대한 비재현율을 분석하여 검색 성능의 척도를 평가하고 이로부터 사용자의 검색 노력을 최소화하는 최적의 검색 엔진을 판정할 수 있다.
- [0031] 도 1은 본 발명의 실시예에 따른 비재현율 기반의 특허 검색 엔진 평가 시스템의 구성을 나타낸 도면이고, 도 2는 도 1의 시스템을 이용한 비재현율 기반의 특허 검색 엔진 평가 방법을 설명하는 도면이다.
- [0032] 도 1에 나타난 것과 같이, 본 발명의 실시예에 따른 비재현율 기반의 특허 검색 엔진 평가 시스템(100)은 정보 설정부(110), 검색 결과 획득부(120), 평가 인자 획득부(130) 및 비재현율 연산부(140)를 포함한다.
- [0033] 먼저, 정보 설정부(110)는 소정 질의에 대응하는 복수의 타겟 문서의 정보를 설정한다(S210). 여기서 질의란, 통상의 질의어나 검색어 등을 의미한다.
- [0034] 예를 들어, 자동차 핸들과 관련한 특허 검색이 필요한 경우에, "자동차", "핸들"과 같은 적어도 하나의 키워드로 질의어가 구성될 수 있다. 일반적으로 특허 검색 엔진에서는 적어도 하나의 키워드, 연산자, 검색 범주(발명 명칭, 청구항, 출원인, 국가 등) 등에 의해 구성된 질의어가 검색 창에 입력되면 질의어와 관련된 문서들을 탐색하여 검색 리스트를 제공한다.
- [0035] 또한, 본 발명의 실시예에서 복수의 타겟 문서란, 해당 질의와 밀접히 관련된 특허 문서들로서, 해당 질의 입력 시 검색 리스트에 반드시 포함되어야 하는 문서들을 의미할 수 있다. 따라서, 검색 결과 내에 타겟 문서의 도출 여부, 타겟 문서의 도출 개수, 그리고 검색 리스트 내 도출 순번(랭킹)은 특허 검색 엔진의 검색 능력을 평가하는데 중요한 자료로 사용된다.
- [0036] 정보 설정부(110)는 특정 질의어와 관련하여 기 설정된 복수의 타겟 문서의 정보를 질의어와 함께 저장할 수 있다. 이때, 타겟 문서의 정보로는 발명 명칭, 출원번호, 출원일, 출원인, 공개번호 등을 포함할 수 있다.
- [0037] 물론, 정보 설정부(110)는 소정 질의에 대응하는 복수의 타겟 문서의 정보를 사용자로부터 입력받아 설정할 수 있다. 즉, 사용자에 의해 사전 검토된 복수의 타겟 문서들의 정보를 사용자로부터 등록받을 수 있다.
- [0038] 검색 결과 획득부(120)는 평가 대상이 되는 검색 엔진에서 해당 질의에 대응하여 도출된 검색 리스트를 획득한다(S220).

- [0039] 검색 리스트는 주로 테이블 형태로 제공되며 검색 리스트 상에서 복수의 검색 문서들에 대한 정보가 표출 순번(랭킹)에 따라 나열되어 제공된다. 예를 들어, 자동차 핸들과 관련하여 총 10건의 특허 문서가 검색된 경우, 1번부터 10번까지 총 10개의 문서들이 상하 배열되어 제공된다. 물론, 각 문서에 대한 요약 정보들(발명명칭, 출원번호, 출원인, 발명자 등)이 해당 열에 함께 제공된다.
- [0040] 여기서, 총 10개의 문서 중 타겟 문서가 리스트 상의 마지막 순번에 위치할 경우, 사용자가 맨 아래의 리스트까지 검토해야 하므로, 관련도 높은 타겟 문서들을 상위에 배열하여 제공하는 검색 엔진 즉, 검색 효율이 높은 검색 엔진을 사용할 경우에 사용자의 검토 시간과 노력을 최소화할 수 있음을 알 수 있다. 본 발명의 경우 이러한 점까지 고려하여 특허 검색 능력을 평가한다.
- [0041] 검색 결과 획득부(120)는 검색 리스트 내에서, 검색된 문서들의 총 개수(N), 검색된 문서들 중 타겟 문서의 검출 개수($|r|$), 검색 리스트 상의 타겟 문서의 표출 순번(r_i)($1 \leq r_i \leq N$)에 대한 정보를 획득한다(S230).
- [0042] 여기서, 검색된 문서들의 총 개수(N)는, 검색 리스트 내 존재하는 문서들의 총 개수를 의미한다. 이는 Raw Data의 개수를 의미할 수 있다.
- [0043] 그리고, 검색된 N 개의 문서들 중 타겟 문서의 검출 개수($|r|$)는 N 개의 문서 중에 존재하는 타겟 문서의 개수로, 최악의 성능을 가진 검색 엔진의 경우 $|r|=0$ 일 것이다. 물론, $|r|$ 의 경우, $0 \leq |r| \leq N$ 의 조건을 가질 수 있다.
- [0044] 여기서, $|r|$ 값이 N 값과 같거나 근접하다고 해서 무조건 성능이 좋은 검색 엔진으로 볼 수는 없는데 그 이유는 다음과 같다. 소정 질의에 대해 기 설정된 타겟 문서의 개수는 5개인 경우를 예를 들면, 검색 엔진에서 해당 질의로 검색한 결과, 검색된 문서의 총 개수는 2개($N=2$)이면서 이들 2개 문서가 모두 타겟 문서($|r|=2$)에 해당하면, $|r|=N$ 이 된다. 하지만, 검색되어야 하는 5개의 타겟 문서 중에 2개만 검색하였고 나머지 3개의 타겟 문서는 검색 결과로 도출하지 못했기 때문에, 해당 검색 엔진을 성능이 우수한 검색 엔진으로 보기 어려울 수 있다.
- [0045] 타겟 문서의 검출 개수($|r|$)는, 검색된 N 개의 문서들과 기 설정된 n 개의 타겟 문서의 정보를 상호 비교하는 것을 통해 획득될 수 있다. 즉, N 개의 문서들 중에서 기 설정된 n 개의 타겟 문서의 정보와 일치하는 문서들의 개수를 구하면 된다. 이때, $|r|$ 은 n 보다 클 수는 없으므로, $|r|$ 과 n 의 관계를 보면 $|r| \leq n$ 이 된다.
- [0046] 검색 리스트 상의 타겟 문서의 표출 순번(r_i)의 경우, 해당 타겟 문서가 검색 리스트 상에서 어느 순번에 도출된 것인지, 즉 검색 리스트 상에서 얼마나 상위 부분에 존재하는지를 의미한다. 여기서 물론, 표출 순번의 숫자가 작은 타겟 문서일수록 상위에 위치한 것에 해당한다.
- [0047] r_i 는 $|r|$ 개의 타겟 문서 중 i 번째 타겟 문서의 표출 순번을 나타낸다. 검색 리스트 내 문서들은 문서의 총 개수(N) 만큼 1번부터 N 번의 순번을 가지므로, r_i 값은 1부터 N 의 범위($1 \leq r_i \leq N$)를 가질 수 있다. 검색되어야 하는 타겟 문서가 검색 리스트 상의 상위 부분에 도출될수록 사용자의 검색 노력을 줄일 수 있음은 앞서 설명한 바 있다.
- [0048] 비재현율 연산부(140)는 설정된 타겟 문서의 총 개수(n), 검색된 문서들의 총 개수(N), 타겟 문서의 검출 개수($|r|$), 타겟 문서의 표출 순번(r_i)을 이용하여, 검색 엔진에 대한 비재현율을 연산한다(S240).
- [0049] 여기서, 비재현율은 검색 리스트 내에 포함된 타겟 문서의 표출 순번(r_i)이 클수록 상승하고, 설정된 타겟 문서의 총 개수(n) 대비 타겟 문서의 검출 개수($|r|$)가 작을수록 상승하도록 설계된다.
- [0050] 여기서, n 값은 미리 설정되어 있으며, n 값 대비 $|r|$ 값이 작을수록 n 과 $|r|$ 간의 편차 즉, $(n - |r|)$ 은 커지게 된다. 이를 달리 표현하면, $(n - |r|)$ 값이 커질수록 비재현율이 증가하게 된다. 간단히, $(n - |r|)$ 값은

타겟 문서의 검출 누락 개수를 의미하며 누락 수가 많을수록 비재현율이 높아진다.

[0051] 그리고, 본 발명의 실시예에서, 비재현율 값의 연산 시에, 타겟 문서의 표출 순번(r_i)에 대해서는 제곱 가중치를 부여할 수 있다.

[0052] 만일, 소정 질의에 대응하여 검색된 문서의 개수가 총 10개($N=10$)이며, 이 중에서 타겟 문서가 4개($n=4$) 존재하고, 검색 리스트 내 1~10번 문서 중 타겟 문서는 4,5,6,7번째 문서로 도출된 경우, $r_i=\{4,5,6,7\}$ 이 된다.

[0053] 본 발명의 실시예에서, 비재현율 연산시에는 타겟 문서의 표출 순번 값 r_i 을 그대로 사용하지 않고, 이것의 제곱 값을 사용하여, 표출 순번에 대하여 가중치를 부여한다. 예를 들어, 4,5,6,7 대신에 $4^2, 5^2, 6^2, 7^2$ 값을 실제 연산에 사용한다. 물론, 이를 위해 $f(x) = x^2$ 함수를 활용한다.

[0054] 물론, 본 발명의 실시예는 이러한 점에 따라, 문서 도출 순위 이외에도 문서의 총 개수(N)에도 제곱 가중치를 적용한다. 만일 $N=10$ 인 경우, 10^2 값을 사용한다.

[0055] 본 발명의 실시예에서 비재현율 값은 0과 1 사이의 값을 가지며, 아래의 수학적 식 1을 통해 연산된다.

수학적 식 1

$$\text{비재현율} = 1 - \frac{\sum_{i=1}^n \beta_i}{\sum_{i=1}^{|r|} \beta_{r_i} + (n - |r|) \cdot \beta_N}$$

[0056]

[0057] 여기서, N 은 검색된 문서들의 총 개수, n 은 상기 설정된 타겟 문서의 총 개수, $|r|$ 는 타겟 문서의 검출 개수, i 는 타겟 문서의 인덱스, r_i 는 i 번째 타겟 문서의 표출 순번을 나타낸다. 이들은 앞서 정의한 바 있다.

[0058] 수학적 식 1에서, $\beta_x = x^2$ 의 함수 규칙을 따른다. 따라서, 수학적 식 1의 분수 항의 분모 및 분자에서 $\beta_{r_i} = r_i^2$, $\beta_N = N^2$, $\beta_i = i^2$ 가 적용된다.

[0059] 수학적 식 1의 분모 부분을 참조하면, 앞서 설명한 바와 같이 $(n - |r|)$ 값이 클수록 분수 항은 작아지면서, '1 - 분수 항'에 따라 결과적으로 비재현율 값은 높아짐을 알 수 있다.

[0060] 또한, 수학적 식 1의 분모 부분에서, 타겟 문서의 표출 순번(r_i)과 관련한 함수인 β_{r_i} 값이 클수록 분수 항은 작아지면서 비재현율이 높아짐을 알 수 있다. 특히, $\beta_{r_i} = r_i^2$ 의 성립에 따라, 단순히 r_i 값을 적용하는 경우보다 r_i 값의 '1' 증가 시마다 비재현율이 급격히 가중되어 상승하는 것을 알 수 있다.

[0061] 이와 같이, 본 실시예는 비재현율 연산 시에 단순히 검색 리스트 상에 타겟 문서가 검출된 개수만 고려하지 않으며, 검출된 타겟 문서가 리스트 상에 위치한 순번, 그리고 검출되어야 하는 총 타겟 문서의 수와 실제 검출된 타겟 문서의 수 간의 차이 등을 고려한다. 본 발명의 실시예는 사용자가 질의와 관련된 문서들을 리스트 상에서 색출해내는데 필요한 시간을 고려한 평가 기법임을 알 수 있다.

[0062] 도 3은 본 발명의 실시예에서 비재현율의 연산 개념을 설명하는 도면이다.

[0063] 도 3에서 가로축은 사용자의 검토 노력(Review effort)에 대응하는 것으로, 리스트 상에서 타겟 문서가 검출된 순번을 의미할 수 있다. 타겟 문서가 검출된 순번이 클수록, 즉 가로축 값이 커질수록 검색 노력이 증가하는 것을 의미한다. 도 3은 검출된 문서의 총 개수가 10인 상황을 예시한 것이므로, 그 안에서 타겟 문서가 검출되는 순번 역시 10 이내의 값을 가진다.

[0064] 세로축은 1-Recall(비재현율)로서 수학적 식 1의 개념과 대응된다. 비재현율은 0과 1 사이의 값을 가지는데, 1에

가까울수록 비재현율이 높기 때문에 검색 성능이 좋은 것을 의미한다

- [0065] 도 3은 수학적 식 1의 도출 개념을 설명하기 위한 것이며, 특정 질의어(ex, 차량 핸들)에 대응하여, 검색 엔진에서 검색된 문서들의 총 개수 $N=10$ 이고, 검색된 10개 문서 중 타겟 문서의 개수 $|r|=4$ 이고, 질의어에 대응하여 기 설정된 타겟 문서의 총 개수 $n=5$ 이고, i 번째 타겟 문서의 표출 순번인 $r_i=\{4,5,6,7\}$ 인 상황(도 3의 actual 상황)을 가정한다.
- [0066] 총 검출된 타겟 문서가 4개이므로, $i=1\sim 4$ 이며, $r_i=\{4,5,6,7\}$ 라는 것은, 검출된 4개의 타겟 문서 각각이 총 10개 문서 중 4번째, 5번째, 6번째, 7번째로 표출된 것을 나타낸다. 즉, 검색 리스트에 순서대로 제공된 1번~10번 문서 중 4,5,6,7번 문서에 차례로 표출된 경우를 나타낸다.
- [0067] 미리 설정된 타겟 문서가 총 5개이기 때문에, 도 3에서 '1-Recall' 값은, 검색된 총 10개 문서 중에서 타겟 문서가 1개씩 검출될 때마다 0.2씩 계단식으로 차감된다.
- [0068] best 상황(점선)의 경우, actual 상황(실선)과의 비교를 위해 도시한 것이며, 모든 5개의 타겟 문서가 10개 문서 중 상위에 1번째부터 5번째까지 차례로 도출된 가장 이상적 경우를 나타낸다. 이 경우 가로축의 1,2,3,4,5 지점에 대응하여 세로축의 값은 0.2씩 단계적으로 차감되어 최종적으로 0의 값을 가진다.
- [0069] 다만, actual 상황(실선)의 경우, 5개 중 4개의 타겟 문서가 검출된 경우이므로 세로축 값은 최종적으로 0.2에서 멈춘 것을 볼 수 있다. 또한, 4개의 타겟 문서가 4,5,6,7번째에 검출되었으므로, 가로축의 4,5,6,7의 지점에 대응하여 세로축의 값은 1에서 0.2씩 단계적으로 차감되어 최종적으로 0.2에 수렴한 것을 알 수 있다.
- [0070] 본 발명의 실시예는 이와 같이 도출된 계단형 그래프의 내부 면적 또는 크기를 기초로 비재현율을 구하고자 한 것이다. 도 4는 도 3에서 도시된 계단형 그래프를 이용한 비재현율 연산 원리를 설명하기 위한 도면이다.
- [0071] 현재의 예시에서, $n=5$ 로 미리 설정된 값이므로, 수학적 식 1의 분자 값($\sum \beta_i$)은 늘 고정된 값($\sum \beta_i = \beta_1 + \beta_2 + \beta_3 + \beta_4 + \beta_5 = 1^2 + 2^2 + 3^2 + 4^2 + 5^2 = 55$)을 가진다. 이러한 분자 값($\sum \beta_i$)은 도 4에서 best 상황에 대한 계단형 라인(점선)의 내부 면적(음영 참조)과 대응되는 개념으로 볼 수 있다.
- [0072] 즉, $1^2, 2^2, 3^2, 4^2, 5^2$ 값은 각각 맨 위층의 계단(가로축의 0~1 구간), 그 아래의 두 번째 계단(0~2 구간), 세번째 계단(0~3 구간), 네번째 계단(0~4 구간), 다섯번째 계단(0~5 구간)에 대한 사각형 면적에 대응된다 볼 수 있다.
- [0073] 다만, 실제 도 4에서 각 계단의 세로 길이는 0.2이고 이를 계단의 면적 계산에 직접적으로 사용한 것은 아니지만, 상술한 바와 같이, 본 발명의 실시예의 경우 각 층의 가로 길이에 대한 제공 값의 개념을 사용함으로써 면적에 상응하는 값(개념)을 얻는다고 볼 수 있다.
- [0074] 수학적 식 1의 분모 부분을 보면, $N=10$ 이므로 β_N 항의 경우 $\beta_N = \beta_{10} = 100$ 으로 늘 고정된 값을 가진다. 분모에서 나머지 값은 케이스 별로 달라질 수 있다.
- [0075] best 상황의 경우, 수학적 식 1의 분모 부분을 보면, $n=|r|=5$ 로서 $(n-|r|)=0$ 이 되고, 분모의 시그마항도 분자의 시그마항과 동일하게 되어, 최종적으로 분자/분모 값은 1이 된다. 또한, $1-\text{분자}/\text{분모}=1/1=0$ 으로 수학적 식 1은 최종적으로 0이 되며, 비재현율이 0이므로 가장 이상적인 상황임을 확인할 수 있다.
- [0076] 하지만, 도 4의 actual 상황의 경우, 수학적 식 1에서 분자/분모 값은 1보다 작은 값을 가지게 된다. 이는 best 상황(점선)보다 actual 상황(실선)의 그래프가 더욱 넓은 면적을 가지는 것과 관련된다.
- [0077] actual 상황의 경우, $n=5$, $|r|=4$ 로서 $(n-|r|)=1$ 이 된다. 수학적 식 1의 분모 부분을 보면, $\beta_N = \beta_{10} = 100$ 으로 늘 고정된 값을 가진다. 분모에서 시그마 항의 경우, $r_i=\{4,5,6,7\}$ 을 대입하면, $\beta_4 + \beta_5 + \beta_6 + \beta_7 = 4^2 + 5^2 + 6^2 + 7^2 = 126$ 이 된다. 결론적으로 분모 값은 $126+100$, 분자 값은 55로서, 비재현율은 $1-(55/226)=0.76$ 이 된다.
- [0078] 이처럼 비재현율이 높게 나온 것은, 총 5개 중 4개의 타겟 문서만이 검출되고 1개의 타겟 문서를 누락한 것도 원인이 되겠지만, 실질적으로 이 4개의 타겟 문서들이 10개 문서 중 후반 부분 즉, 4,5,6,7번째 문서로 검색된 것에 따른 영향이 매우 큰 것을 알 수 있다. 이는 물론 타겟 문서의 표출 순번(r_i)에 대해 제공 가중치를 부여한 것과도 관련된다.

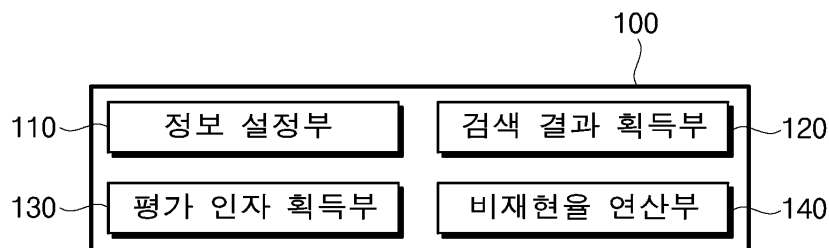
- [0079] 도 4를 통해 알 수 있는 것은 수학적 식 1의 분자/분모는 결국 best/actual의 면적의 개념에 대응하고, actual 그래프가 best에 비해 옆으로 퍼질수록(표출 순번이 뒤로 밀릴수록) 비재현율이 높아지는 것을 알 수 있다.
- [0080] 물론 수학적 식 1에서 시그마 항은 면적과 관련된 것이지만, 분자 부분에는 시그마 항 이외에도 다른 항이 존재하므로 비재현율은 단순 면적의 비로만 연산되지 않는다. 즉, 수학적 식 1의 분모를 보면, 타겟 문서의 검출 개수가 작을 때도 패널티가 부가되도록 $(n - |r|)$ 항이 시그마항에 더해져 면적이 커지는 효과를 부가하며, 특히 $(n - |r|)$ 값에 β_N 의 가중치가 부가되어 패널티 효과를 더욱 극대화한다. 따라서, 타겟 문서를 검색하지 못한 개수, 즉 타겟 문서의 누락 개수에 대한 패널티가 분자 부분에 추가로 부가된 것을 알 수 있다.
- [0081] 도 5는 도 3에서 각 경우에 대한 비재현율 연산 실시예를 나타낸 도면이다.
- [0082] 도 5의 (a)는 5개의 타겟 문서 중 4개를 4,5,6,7번째 순으로 찾은 경우(도 3의 actual)이며, (b)는 5개 타겟 문서 중 5개 모두를 1~5번째 순으로 찾은 이상적(도 3의 best) 경우로서 앞서 설명한 바 있다.
- [0083] 도 5의 (c)는 5개의 타겟 문서 모두 검색 결과에 없는 경우($|r|=0$)로서, 분모의 시그마 항은 0이 되고 비재현율을 값은 0.89가 나온다. 물론, (d)와 같이, 검색된 문서의 총 개수 $N=1000$ 인 경우는 분모 값이 매우 커지면서 분수 항은 거의 0으로 수렴하여, 비재현율은 1과 같아진다. 그 밖에도, $N=100$ 인 경우 역시 비재현율은 0.99로서 거의 1과 같아진다.
- [0084] 이상과 같은 본 발명에 따르면, 검색 엔진의 검색 정확도 및 사용자의 검색 시간을 고려하여 검색 엔진의 검색 능력의 척도를 현실적으로 평가할 수 있는 이점이 있다.
- [0085] 본 발명은 도면에 도시된 실시 예를 참고로 설명되었으나 이는 예시적인 것에 불과하며, 본 기술 분야의 통상의 지식을 가진 자라면 이로부터 다양한 변형 및 균등한 다른 실시 예가 가능하다는 점을 이해할 것이다. 따라서, 본 발명의 진정한 기술적 보호 범위는 첨부된 특허청구범위의 기술적 사상에 의하여 정해져야 할 것이다.

부호의 설명

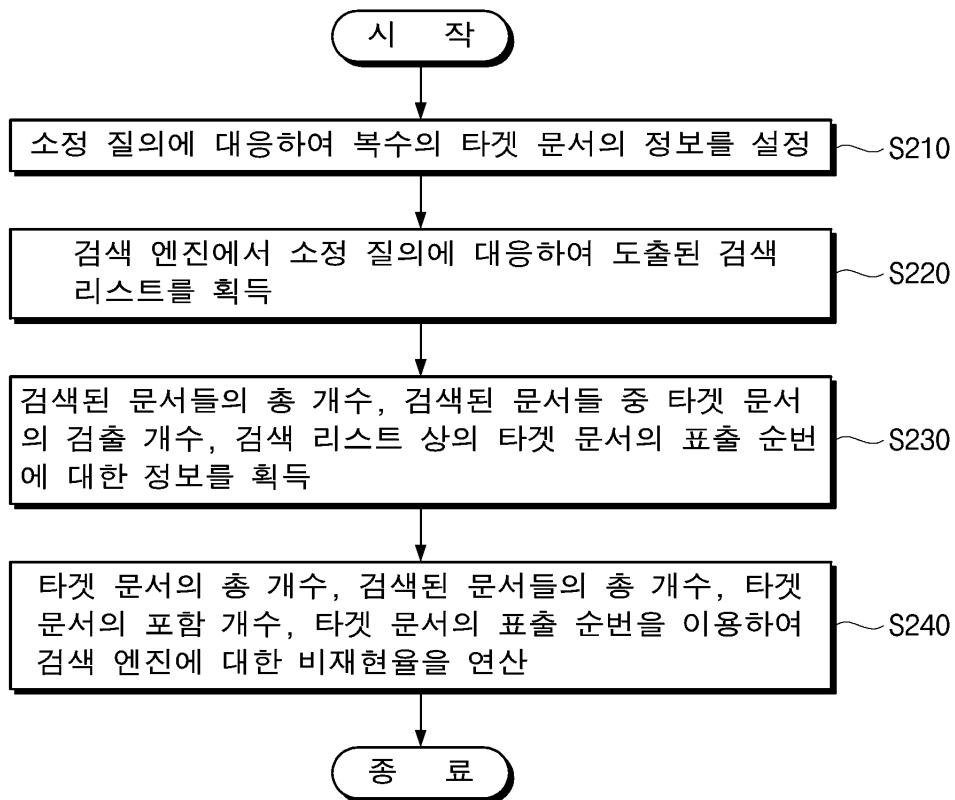
- [0086] 100: 특허 검색 엔진 평가 시스템
110: 정보 설정부
120: 검색 결과 획득부
130: 평가 인자 획득부
140: 비재현율 연산부

도면

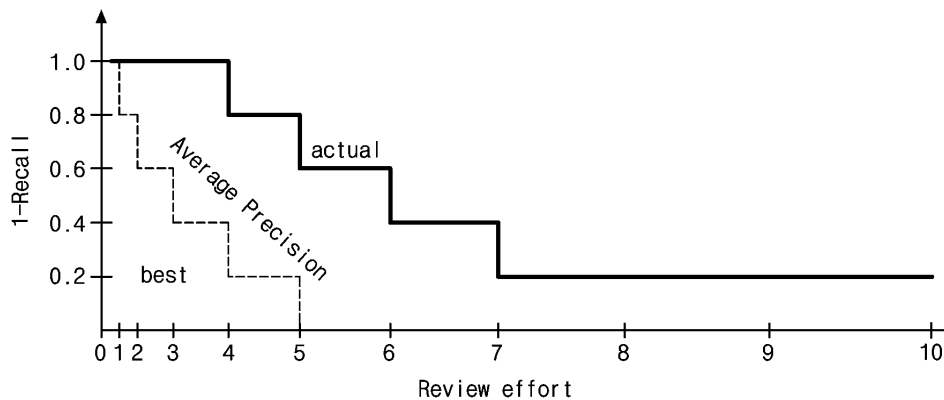
도면1



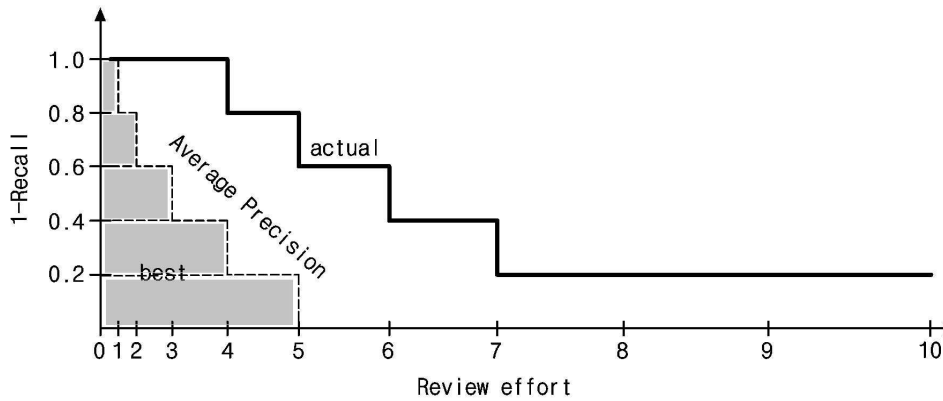
도면2



도면3



도면4



도면5

<전체 N=10개의 검색 결과 중, n=5 개의 관련문서를 찾는 상황> ($N = 10, n = 5, \beta_x = x^2$)

$$\text{비재현율} = 1 - \frac{\sum_{i=1}^n \beta_i}{\sum_{i=1}^{|r|} \beta_{r_i} + (n - |r|) \cdot \beta_N}$$

(a) 4,5,6,7 순으로 검색 결과가 나오는 경우(actual) ; $r_i = 4,5,6,7 \quad |r| = 4$

$$1 - \frac{1^2 + 2^2 + 3^2 + 4^2 + 5^2}{(4^2 + 5^2 + 6^2 + 7^2) + (5 - 4) \cdot 10^2} = 1 - \frac{1 + 4 + 9 + 16 + 25}{(16 + 25 + 36 + 49) + 100} = 1 - \frac{55}{226} = 0.76$$

(b) 1,2,3,4,5 순으로 검색 결과가 나오는 경우(best)(이상적) ; $r_i = 1,2,3,4,5 \quad |r| = 5$

$$1 - \frac{1^2 + 2^2 + 3^2 + 4^2 + 5^2}{(1^2 + 2^2 + 3^2 + 4^2 + 5^2) + (5 - 5) \cdot 10^2} = 1 - \frac{1^2 + 2^2 + 3^2 + 4^2 + 5^2}{1^2 + 2^2 + 3^2 + 4^2 + 5^2 + 0} = 0$$

(c) 한 개도 검색 결과에 없는 경우; $r_i = \emptyset \quad |r| = 0$

$$1 - \frac{1^2 + 2^2 + 3^2 + 4^2 + 5^2}{5 \cdot 10^2} = 0.89$$

(d) 검색 결과의 수 N= 1000, 한 개도 검색 결과에 없는 경우(worst)

$$1 - \frac{1^2 + 2^2 + 3^2 + 4^2 + 5^2}{5 \cdot 1000^2} \approx 1$$