(19)대한민국특허청(KR)
(12) 공개특허공보(A)

(51) 。Int. Cl.
G06F 17/30 (2006.01)

(11) 공개번호 10-2006-0047306
(43) 공개일자 2006년05월18일

(21) 출원번호 10-2005-0033008
(22) 출원일자 2005년04월21일

(30) 우선권주장 JP-P-2004-00127122 2004년04월22일 일본(JP)

(71) 출원인 휴렛-팩커드 디벨롭먼트 컴퍼니, 엘 피
미국 텍사스주 77070 휴스턴 스테이트 하이웨이 249 20555

(72) 발명자 오다 히로미
일본 가나가와켄 가와사키시 다마쿠 스게센고쿠 3-1 츠구미다이1-203

(74) 대리인 김창세
김원준

심사청구 : 없음

(54) 문서 검색 및 분류 방법 및 그 시스템, 문서 처리 방법 및그 시스템 또는 메모리

요약

전문가 영역과 소인 영역에서 이용가능한 어휘 쌍 또는 문서 쌍이 없는 경우에, 소인 영역 내의 용어에 대응하는 전문가 영역에 사용된 용어(예, 단어)가 검출된다. 동일한 주제에 대한 설명인 것으로 알려져 있으며 전문가 영역과 소인 영역으로 기록된 문서는 인터넷을 검색함으로써 수집된다. 이들 문서에 발생하는 용어의 빈도수가 계수된다. 그 계수는 전문가 언어 표현의 어휘와 소인 언어 표현의 어휘 간의 대응을 계산하는데 사용된다.

대표도

도 1

명세서

도면의 간단한 설명

도 1은 본 발명의 바람직한 실시예를 실시하기 위한 전체 시스템에 대한 도면,

도 2는 도 1의 시스템에 포함된 장치에 대한 도면,

도 3은 도 1의 시스템에 의해 수행되는 알고리즘에 대한 흐름도,

도 4는 도 2의 장치에 의해 사용되며 도 1의 시스템의 검색 문서로부터 "노이즈" 문서를 제거하기 위한 방법의 흐름도,

도 5는 도 2의 장치에 의해 사용되며 문서의 순위 상관 계수와 유의도를 계산하기 위한 방법의 흐름도,

도 6은 도 2의 장치에 의해 사용되며 문서를 전문가 문서와 소인(naive) 문서로서 분류하기 위한 방법의 흐름도,

도 7은 도 2의 장치에 의해 사용되며 MLR 방법을 이용하여 사전적 매핑을 수행하기 위한 방법의 흐름도,

도 8a는 전문가 용어 행렬에 대한 도면,

도 8b는 소인 용어 행렬에 대한 도면,

도 8c는 사전적 매핑 행렬에 대한 도면,

도 9는 도 2의 장치에 의해 사용되며 도 8c의 사전적 매핑 행렬을 계산하는 알고리즘.

도면의 주요 부분에 대한 부호의 설명

110 : 유저 PC 120 : 사이트 서버(1)

130 : 사이트 서버(2) 140 : 인터넷 네트워크

210 : 기억 장치 220 : 메인 메모리

230 : 출력 장치 240 : 중앙 제어 장치(CPU)

250 : 조작 장치 260 : 네트워크 I/O

발명의 상세한 설명

발명의 목적

발명이 속하는 기술 및 그 분야의 종래기술

본 발명은 공통의 제목을 가진 복수 세트의 문서의 처리에 관한 것이다.

동일 언어로 복수로 기술되고 동일 내용을 공유하는 문서는, 그 제목에 대해 저자가 가지고 있는 전문 지식과, 저자가 속하는 상이한 사회 계층, 예를 들어, 성별 또는 나이에 따라서 상이한 명세서의 용어를 종종 사용한다. 그 명세서가 공통의 제목에 관한 것일지라도, 비전문가가 사용한 용어와 전문가가 사용한 용어는 그들의 각각의 표현 영역에서 상당히 다를 수 있다.

본 발명의 목적은, 이러한 상이한 영역에서, 전문가 사용 용어에 의해 그것이 무엇을 의미하는가에 해당하는 비전문가 사용 용어를 검출하고, 역으로, 비전문가 사용 용어에 의해 그것이 무엇을 의미하는가에 해당하는 전문가 사용 용어를 검출하는 장치 및 다른 필요한 기술의 신규하고 개선된 방법을 제공하는 것이다.

문서를 상이한 영역으로 변환하는 기술의 전형적인 예는 변환 기계이다. 컴퓨터가 변환 기계의 작업을 수행하게 하는 기술은 잠시동안 알려졌다. 변환 기계는 용어 데이터베이스를 이용하는 컴퓨터 프로그램, 문법적 규칙, 어법 및 문장예의 데이터베이스, 및 다른 시스템 특정 성분을 이용하여, 원어로 기록된 문서를 다른 원어로 자동 변환한다. 이러한 기술은 이미 실제 이용되고 있으며, 퍼스널 컴퓨터용의 시판중인 언어 변환 소프트웨어 제품이 있다. 몇몇 변환 서비스가 또한 인터넷상에서 제공된다. 추가로, 단어별 변환용의 소형 휴대용 장치가 광범위하게 이용가능하다. 단어별 변환 기계는 어떤 언어의 하나의 단어를 동일한 의미의 다른 언어의 단어로 변환한다. 기본적으로, 사전편집된 사전이 저장 장치에 저장되어 있으며, 입력 단어는 다른 언어의 대응 단어로 변환된다. 이들 종래의 기술은 하나의 영역으로부터 다른 영역으로 문서를 변환하기 위한 전제 조건을 가지고 있으며, 즉, 하나의 영역 내의 문장은 다른 영역의 문장에 대응하는 것으로 알고 있어야 하며, 하나의 영역 내의 단어는 다른 영역의 단어에 대응하는 것으로 알고 있어야 한다.

어려운 표현을 동일한 언어의 쉬운 표현으로 변환하기 위한 패러프레이징 연구(paraphrasing research)가 이미 발표되었다. 예를 들어, Atsushi Fujita 외 다수(2003) 및 Masahiro Murayama 외 다수(2003)에 의한 연구에 보고되어 있다. "패러프레이징"에 관한 연구에서, 기본적인 기술은 패턴 일치 규칙에 따라서 소정의 표현 패턴으로 대체될 표현 패턴을 찾는 것이다. 언어 변환의 다른 접근 방식은 통계적 및/또는 개연적 모델을 이용한다. 이들 모델 기반의 접근 방식은 동일한 것으로 알고 있는 내용을 가진 상이한 언어의 한 쌍의 데이터 세트를 초기에 마련한다. 다음에, 각각의 데이터 세트의 문장 길이와 같은 정보에 근거하여, 언어 A와 언어 B의 대응 문장이 결정된다. 최종적으로, 단어 간의 대응 관계가 데이터 세트 내의 그들의 동시 발생(co-occurrence) 관계에 근거하여 결정된다. 이러한 종래 기술의 경우 및 다른 종래 기술의 경우에, 언어 B의 단어 Wb가 적당한 의미 정확도를 가진 언어 A의 단어 Wa에 대응한다는 전제가 있다.

(특허 문헌 1) "Daily Language Computing and its Method" JP 2002-236681A

(특허 문헌 2) "Association Method for Words in Paginal Translation Sentences" JP 2002-328920A

(비특허 문헌 1)

http://www2.crl.go.jp/it/a133/kuma/mrs_li/midisearch.htm.

(비특허 문헌 2)

Atsushi Fujita, Kentaro Inui, Yuji Matsumoto. "Text Correction Processing necessary for Paraphrasing into Plain Expressions". Collection of Lecture Theses in 65th National General Meeting of Information Processing Society of Japan, 5th Separate Volume, 1T6-4, pp. 99-102, 2003.03.

(비특허 문헌 3)

Masahiro Murayama, Masahiro Asaoka, Masanori Tsuchiya, Satoshi Sato. "Normalization of Terms and Support for Paraphrasing Declinable words based on the Normalization", Language Processing Society, 9th Annual General Meeting, pp 85-88, (2003.3).

(비특허 문헌 4)

Dunning, T. (1993). Accurate methods for the statistics of surprise and coincidence. Computational Linguistics. 19(1):61-74

상술한 바와 같이, 통상적인 기계 변환에서, 질문에 2개 언어의 대응 단어가 있다고 가정하고, 대응 문서 세트가 하나의 언어에서 다른 언어로의 변환에 이용가능하다고 가정한다.

본 발명의 목적은, (1) 타겟 영역에서 서로 대응하는 기지의 단어 쌍, (2) 사전에 서로 대응하는 것으로 알고 있는 문서 세트 쌍, 및/또는 (3) 질문에서 영역의 매핑에 도움이 되는 사전 또는 백과 사전이 없는 경우에도, 다른 영역의 용어에 대략 대응하는 하나의 영역의 용어를 검출하는 신규하고 개선된 방법 및 장치를 제공하는 것이다.

발명이 이루고자 하는 기술적 과제

본 발명의 일측면에 따르면, 상술한 문제를 해결하기 위해서,

(1) 동일 주제에 대해서 기술된 것으로 알고 있는 2개의 상이한 종류의 언어 표현으로 기록된 문서(이하에서, 이들 문서를 영역 A의 문서와 영역 B의 문서라 함)가 검색되며,

(2) 2개의 상이한 종류의 언어 표현의 이러한 문서 세트가 주어질 때, 영역 A의 문서에 나타난 용어와 영역 B의 문서에 나타난 용어 간의 연관성이 확립된다.

이러한 목적으로, 2개의 상이한 종류의 언어 표현으로 기록된 문서 세트를 마련하기 위해서, 검색 도구를 이용한 소정의 키워드 리스트를 이용하여 후보 문서가 수집된다. 그러나, 검색 도구를 이용하여 검색된 후보 문서는 상당수의 소위 "노이

즈" 문서를 포함하고 있기 때문에, 대부분의 경우에, 검색 결과는 있는 그대로 사용될 수 없다. 따라서, 본 발명의 일측면은 수집된 문서로부터 "노이즈" 문서를 제거하는 초기 단계를 포함한다. 이러한 초기 단계 후에, 문서는 용어 빈도수와 문서의 다른 정보에 근거하여, 상이한 형태의 언어 표현을 포함하는 전문가 문서와 소인 문서로 분류된다. 타겟의 전문가 문서와 타겟의 소인 문서에 나타난 용어가 항상 동일하지 않기 때문에, 2개의 상이한 영역 내의 용어 간의 상관 관계가 다음에 계산된다. 기본적인 개념은 전문가 또는 소인 영역에 나타나는 하나의 용어 또는 용어 세트와, 다른 영역에 나타나는 하나의 용어 또는 용어 세트와의 연관성이, 동일한 대상(object)에 대해서 기술된 것으로 알고 있는 전문가 문서 세트와 소인 문서 세트의 용어 간의 동시 발생 관계에 근거하여 획득된다는 개념이다.

본 발명의 적용예는 몇몇 제품 또는 물건을 막 구매하고자 하는 고객을 위한 추천 시스템이다. 문서가 상품과 같은 동일 대상에 대해 기술하고 있을지라도, 일반적으로, 대상에 대해 깊은 지식을 가진 전문가가 사용하는 용어와 그 대상에 대해 약간의 지식을 가진 비전문가가 사용하는 용어 간에는 상당한 차이가 있다. 전문가는 전문 용어와 대상에 대한 구체적인 지식을 이용하여 대상을 기술하지만, 이러한 지식이 없는 비전문가는 지각력에 근거한 표현으로 대상을 기술하거나 유사한 대상 또는 예에 의해서 기술하는 것을 제외하고는 할 수 없다. 전문가는 제조 회사 및/또는 구성 물질에 대한 자신의 지식을 이용하여 제품을 상세히 설명하고자 하는 반면에, 비전문가는 마음에 떠오르는 지각력 기반의 용어를 이용하여 동일 제품을 기술하고자 한다. 모든 관심 분야에서 일반적인 소비자가 제품에 대한 상세한 지식 및 제품에 관한 적절한 명칭을 가지고 있는 것은 거의 불가능하다. 따라서, 사실상 광범위하게 선택하기 위해서 전문 지식을 필요로 하는 특정 제품을 전문가가 비전문가에게 설명하고 추천할지라도, 비전문가는 구매 전에 그 설명을 충분히 이해할 수 없을 것으로 생각된다.

본 발명을 적용함으로써, 판매자는 소비자가 이해할 수 없는 어휘로 제품에 대한 충분한 정보를 소비자에게 제공할 수 있으며, 역으로, 일반적인 소비자는 제품에 대한 정보를 쉽게 이해하여 자신의 기호 및 취향에 맞는 정보를 선택할 수 있다.

발명의 구성 및 작용

도 1은 네트워크(140)에 접속되어 있는 사용자 PC(110), 사이트 서버(1)(120) 및 사이트 서버(2)(130)를 포함하는 시스템에 대한 도면이다. 사용자는 PC(110)의 동작을 통한 몇몇 검색 도구를 이용하여 필요한 정보를 획득하기 위해서, 사이트 서버(1)(120) 및 사이트 서버(2)(130)에 액세스한다. 인터넷상에서의 검색은 도 1의 실시예와 같이 기술된다. 그러나, 시스템이 필요한 정보를 검색할 수 있는 한, 다른 검색 시스템이 사용될 수 있다. 사용자는 획득된 정보를 사용자 PC(110) 상의 컴퓨터 프로그램을 이용하여 처리함으로써 원하는 결과를 얻을 수 있다.

도 2는 저장 장치(210), 메인 메모리(220), 출력 장치(230), 중앙 처리 장치(CPU)(240), 조작 장치(250) 및 네트워크 I/O(260)를 구비한 하우징(200)을 포함하는 사용자 PC에 대한 도면이다. 사용자는 조작 장치(250)를 동작시켜 네트워크 I/O(260)를 통해 인터넷 상의 각각의 사이트로부터 필요한 정보를 획득한다. 중앙 처리 장치(240)는 저장 장치(210)에 저장된 문서 정보에 기초하여 인터넷으로부터 검색된 정보에 대한 소정의 데이터 처리를 수행하고, 출력 장치(230) 상에 결과를 디스플레이한다.

도 3은 도 1의 시스템과 도 2의 PC에 의해 수행되며, 소인 문서와 전문가 문서 간의 대응하는 용어를 검출하는 동작(즉, 단계)에 대한 흐름도이다. 그 단계는 다음과 같다.

단계 310 : 지정된 용어를 이용하여 후보 문서를 취득

단계 320 : 후보 문서를 전처리

단계 330 : "노이즈" 문서를 제거

단계 340 : 각 문서의 특징값을 산출

단계 350 : 판별 분석을 이용하여 문서를 분류

단계 360 : 소인 문서와 전문가 문서 간의 대응 용어를 검출

각각의 단계는 이하에 상세히 설명된다.

(1) 지정된 용어를 이용하여 후보 문서를 취득

대응 용어(310)를 검출함에 있어서의 제 1 단계는 동일 내용을 기술하는 소인 문서(비전문가에 의해 기술된 문서, 이하, N 문서라 함)와 전문가 문서(전문가에 의해 기술된 문서, 이하, E 문서라 함)의 쌍을 포함하는 데이터 세트를 준비하는 것이다. 데이터 세트는 용어 리스트를 이용함으로써 준비된다.

용어 리스트는 임의의 주어진 영역에서의 키워드로서 사용될 수 있는 용어 세트이다. 예를 들어, "와인(wine)"의 영역이 선택될 때, 용어 리스트는 "와인의 (제품) 이름"을 포함한다. 사용자는 용어 리스트에 기술된 와인 이름에 따라서 인터넷 상의 검색 도구를 이용하여 와인에 대한 정보를 수집한다. "아우스레제 - 샤토 큐어 본 - 샤토 마고 - 빈 산토 토스카노 (Auslese - Chateau Cure Bon - Chateau Margaux - Vin Santo Toscano)"와 같은 와인 이름이 지정된다. 이들 용어를 키워드로서 가진 데이터베이스로부터 후보 문서가 검색된다. 데이터베이스는 데이터베이스가 이러한 정보를 저장하고 있는 한 사용될 수 있다. 인터넷 상의 탐색 엔진을 이용하여 후보 문서를 검색하는 방법이 설명된다.

사용자는 상술한 용어 리스트에서 탐색 키로서 정의된 와인 이름을 이용하여 검색한다. 와인 이름은 시판 제품 또는 프리 소프트웨어로서 이용가능한 탐색 엔진을 이용함으로써 검색된다. 일반적으로, 와인 이름이 탐색 키로서 지정된 경우에, 상당수의 후보 문서가 검색된다. 그러나, 소정수의 후보가 몇몇 순위에 따라서 선택될 수 있다. 용어 리스트를 이용함으로써 모든 원하는 용어에 대해 후보 문서를 자동으로 획득하는 것이 가능하다.

(2) 후보 문서를 전처리(단계 320)

이러한 방식으로 인터넷 상의 웹 페이지로부터 자동으로 획득된 문서는 여러 종류의 정보를 포함하며, 대부분의 경우에, 있는 그대로 사용할 수 없다. 가비지(garbage) 문서, 리스트 문서 및 일기 문서에 대응하는 문서가 자동으로 획득된 문서로부터 "노이즈" 문서로서 제거된다. "노이즈" 문서를 제거하기 전에, 웹 페이지로부터 추출된 문서를 전처리한다. 전처리의 제 1 단계에서, 문서로 간주될 수 있는 부분이 웹 페이지의 정보로부터 추출되어 문서를 분석한다. 다음에, 내용어, 불변화사(particle), 보조어 등을 추출하기 위해 단어로 분절하여 이들 문서에 대한 특징값, 즉, 내용어의 개수값, 소인어의 비율, 적절한 명사의 비율, 추가의 적절한 명사의 비율, 및 불변화사/보조어의 비율을 산출할 수 있다. 이들 특징값을 산출하기 위해서 본 명세서에 사용된 개념적인 용어가 이하에 기술된다.

(i) 내용어의 개수

이는 웹 페이지의 문서에 포함된 내용어의 개수이다. 내용어는 불변화사/보조어를 제외한 명사, 동사, 형용사 및 부사를 포함한다.

(ii) 소인어의 비율 = 소인어의 개수/내용어의 개수

소인어는 관련 분야의 비전문가에 의해 사용된 소정의 언어이다. 소인어의 비율은 내용어의 개수에 대한 하나의 웹 페이지에 나타나는 소정의 소인어(이하, "마스터 소인어"라 함)의 개수의 비율이다.

(iii) 적절한 명사의 비율 = 적절한 명사의 개수/내용어의 개수

이 항목에서의 적절한 명사는 적절한 명사로서 일반적으로 알려진 명사이다. 적절한 명사의 비율은 내용어의 개수에 대한 웹 페이지에 나타난 적절한 명사의 개수의 비율이다.

(iv) 추가의 적절한 명사의 비율 = 추가의 적절한 명사의 개수/내용어의 개수

추가의 적절한 명사는 적절한 명사로서 일반적으로 인식되지 않지만, 대응 용어를 검출하기 위해서 적절한 명사로서 추가될 필요가 있는 명사이다. 추가의 적절한 명사의 비율은 내용어의 개수에 대한 하나의 웹 페이지에 나타난 추가의 적절한 명사의 개수의 비율이다.

(v) 불변화사/보조어의 비율 = 불변화사의 개수/보조어의 개수/내용어의 개수

불변화사/보조어의 비율은, 하나의 웹 페이지에 나타난 보조어의 개수에 대한 불변화사의 개수의 비율을 산출하고, 그 비율을 내용어의 개수로 제산하여 그 비율을 표준화함으로써, 계산된다.

(vi) 내용어 n 기록

문서 간의 상관 관계는 내용어 단일 기록, 내용어 이중 기록, 내용어 3중 기록 및 내용어 스킵 이중 기록을 이용함으로써 검사된다.

내용어 이중 기록은 하나의 워드(또는 용어)의 빈도수에 기초하여 문서 간의 상관 관계를 결정하는데 사용된다. 와인 영역의 경우에, "와인", "맛", 및 "마시다"와 같은 단어의 빈도수가 사용될 수 있다.

내용어 이중 기록은 2개의 연속적인 단어의 빈도수에 기초하여 문서 간의 상관 관계를 결정하는데 사용된다. 와인 영역의 경우에, "알코올 퍼센트", "이러한 와인" 및 "제조국 - 연도"와 같은 2개의 연속적인 단어의 빈도수가 사용된다.

내용어 3중 기록은 3개의 연속적인 단어의 빈도수에 기초하여 문서 간의 상관 관계를 결정하는데 사용된다. 와인 영역의 경우에, "와인-고기-식음 방법", "화이트-프랑스-1990", 및 "레드-독일-아우스레제"와 같은 3개의 연속적인 단어의 빈도수가 사용된다.

내용어 스킵 이중 기록은 3개의 연속적인 단어 중에서 첫 단어와 최종 단어를 이용하여, 이들 단어의 빈도수에 기초하여 문서 간의 상관 관계를 결정한다. 예를 들면, "고품질" 및 "제조"는 이중 기록 패턴의 첫 단어와 최종 단어로서 표시된다. 결과적인 패턴이 "고품질 - XXX - 제조"를 필요로 하기 때문에, "고품질 - 과일 묶음 - 제조" 또는 "고품질 - 라인산 - 제조"와 같은 시퀀스는 그 조건을 만족한다. XXX는 임의의 단어를 표시한다.

(vii) 불변화사/보조어 n 기록

유사하게, 불변화사/보조어 단일 기록, 이중 기록, 3중 기록, 불변화사/보조어 이중 기록, 불변화사/보조어 3중 기록, 및 불변화사/보조어 스킵 이중 기록이 사용된다.

불변화사/보조어 단일 기록의 예는, "노(no)", "가(ga)" 및 "니(ni)"를 포함한다. 불변화사/보조어 이중 기록의 예는, "노-가", "노-노" 및 "노-니"를 포함한다. 불변화사/보조어 3중 기록의 예는 "노-가-가", "노-노-가" 및 "노-니-가"를 포함한다.

불변화사/보조어 스킵 이중 기록의 예는, "노 - X - 가", "노 - X - 가" 및 "노 - X - 가"를 포함한다. "X"는 임의의 불변화사 또는 보조어임을 알아야 한다.

(viii) 순위 상관 계수 및 그들의 유의도

이러한 실시예에서, 순위 상관 계수 및 유의도는 스피어먼(Spearman) 공식을 이용하여 계산된다. 이는 예를 들어 내용어 단일 기록을 이용하여 설명될 것이다. 먼저, 마스터 소인 문서에 사용된 "술(liquor)", "맛(flavor)", "마시다(drink)", "풍미(taste)", "느낌(feel)", 및 "평가(think)"와 같은 단어의 빈도수가 결정된다. 유사하게, 특정 웹 사이트로부터 획득된 문서에 사용된 "술(liquor)", "맛(flavor)", "마시다(drink)", "풍미(taste)", "느낌(feel)" 및 "평가(think)"와 같은 단어의 빈도수가 결정된다. 이들 단어의 빈도수의 순위는 각각의 문서에 있어서 계산된다. 스피어먼의 순위 상관 계수가 이들 순위 정보 부분에 기초하여 계산되며, 상관 계수의 유의도가 계산된다.

(ix) 마스터 소인 문서 세트(또는 마스터 전문가 문서 세트)

마스터 소인 문서 세트는 특정 영역에서 비전문가가 사용하는 용어를 포함하는 문서의 집합체이다. 마스터 전문가 문서 세트는 특정 영역에서 전문가가 사용하는 용어를 포함하는 문서의 집합체이다.

(3) "노이즈" 문서를 제거

인터넷 상의 웹 페이지로부터 검색된 문서로부터 "노이즈" 문서로서의 가비지 문서, 리스트 문서, 및 일기 문서를 제거하는 것이 필요하다. 다른 영역에서 사용된 용어에 대략 대응하는 하나의 영역에 사용된 용어를 검출하는데 필요한 정보가 "노이즈" 문서에 포함되어 있지 않다고 생각된다. 도 4는 "노이즈" 문서의 제거를 위한 도 1의 시스템에 의해 수행되는 단계의 흐름도이다.

410 : 가비지 문서를 제거

420 : 리스트 문서를 제거

430 : 일기 문서를 제거

440 : 모든 문서가 제거되었음을 확인

450 : 다음 문서를 지정

가비지 문서, 리스트 문서 및 일기 문서의 제거가 이하에 설명된다.

(A) 가비지 문서

이하의 조건 모두를 만족하는 문서가 가비지 문서로서 정의된다. 가비지 문서는 글자 그대로 가비지이여서 하나의 영역에서 다른 영역으로의 용어 검출에 사용될 수 없다. 가비지 문서의 선택 기준은 아래와 같이 정의된다.

(a) 내용어의 개수가 적다.

(b) 소인어의 비율이 낮다.

(c) 적절한 명사의 비율이 낮다.

(d) "마스터 소인 문서"와의 상관 계수가 낮다.

"마스터 소인 문서 세트"는, 비전문가에 의해 기록된 문서로서 미리 선택된 문서 세트이다. 대안으로, 전문가에 의해 문서로서 미리 선택된 문서 세트는 "마스터 전문가 문서 세트"로서 사용될 수 있다.

(B) 리스트 문서

이하의 조건 모두를 만족하는 문서는 리스트 문서로서 정의된다. 특정 영역에서의 대상에 대한 정보가 인터넷 상의 사이트에서 하나의 리스트로서 단순히 저장된 경우에 발생한다.

(a) 적절한 명사의 비율이 높다.

(b) 내용어와 불변화사/보조어에 기초한 "마스터 소인 문서"와의 상관 계수가 낮다.

(C) 일기 문서

이하의 조건 모두를 만족하는 문서는 일기 문서로서 정의된다. 일기 문서는 예를 들어, 술과 와인에 대한 정보가 기술되어 있지만 다른 주제 또는 정보가 주로 논의되는 문서 형태이다. 이러한 문서는 술 또는 와인을 다루면서 여러 다른 주제를 포함하는 개인의 일기 또는 온라인 백화점의 인터넷 사이트 상에 나타날 수 있다.

(a) 특정 영역에 관한 적절한 명사의 비율은 낮다.

(b) 내용어 n 기록에 기초한 마스터 문서와의 상관 관계가 낮다.

(c) 불변화사/보조어 n 기록에 기초한 마스터 문서와의 상관 관계가 낮다.

상술한 정의에 기초하여, 가비지 문서, 리스트 문서 및 일기 문서는 그들이 "노이즈" 문서인 것으로 생각되기 때문에 용어 영역 검출 과정의 고려에서 제거된다.

(4) 판별 분석으로 문서를 분류

"노이즈" 문서의 제거 후에, 판별 분석이 적용되어 남은 문서를 소인 문서 또는 전문가 문서로 분류한다. 판별 분석을 행하기 위해서, 특징값이 각각의 입력 문서로부터 추출된다. 사용된 특징값은 5종류의 비율, 즉, 내용어의 개수, 소인어의 비율,

적절한 명사의 비율, 추가의 적절한 명사의 비율, 및 불변화사/보조어의 비율을 가지고 있다. 또한, 내용어 n 기록으로부터 산출된 스피어먼의 상관 계수와 유의도, 및 불변화사/보조어 n 기록으로부터 계산된 스피어먼의 순위 상관 계수 및 그 유의도가 사용된다.

스피어먼 공식에 기초한 순위 상관 계수와 그 유의도의 계산이 이하에 설명된다. 도 5는 도 2의 컴퓨터가 스피어먼 공식에 기초하여 순위 상관 계수와 그 유의도를 계산하기 위해 수행하는 동작의 흐름도이다.

510 : 마스터 소인 문서에서의 N 기록의 빈도수(Y)

520 : 입력 문서에서의 N 기록의 빈도수(X)

530 : X와 Y에 따라서 스피어먼의 순위 상관 계수(ri)와 유의도(ei)를 계산

540 : 모든 n 기록에 대한 계산을 확인

550 : 다음 n 기록을 지정

560 : 모든 n 기록에 대한 순위 상관 계수와 그 유의도를 획득

순위 상관 계수/유의도가 이하에 상세히 설명된다.

내용어 단일 기록이 설명을 위해 예로서 사용된다. 이들 문서는 단일 단어의 빈도수에 기초하여 문서 간의 상관 관계를 계산하는데 사용된다. 와인 영역의 경우에, "와인", "맛" 및 "마시다"와 같은 단어의 빈도수는 선택된 문서와 마스터 소인 문서 세트(또는 마스터 전문가 문서 세트)로부터 계산된다. 이들 빈도수는 Y(y1, y2, y3,...yh)로서 표시된다(단계 510).

다음에, 유사한 특징값이 입력 문서로부터 계산되며, 유사한 특징값은 X(x1, x2, x3,...xh)로서 표시된다(단계 520). 여기서, h는 빈도수가 계산되는 데이터 또는 단어 유형의 개수를 나타낸다. 순위 상관 계수 및 유의도는 이들 데이터로부터 스피어먼 공식에 기초하여 계산된다.

$$r1=F(X,Y)$$

$$e1=G(X,Y)$$

여기서, r1은 스피어먼의 상관 계수 공식에 따라서 계산된 순위 상관 계수이며, e1은 스피어먼의 유의도 공식에 따라서 계산된 순위 상관 계수의 유의도이다(단계 530). 동일한 방식으로, r2, e2가 내용어 이중 기록, 유사하게 다른 n 기록에 대해서 계산된다. 또한, 순위 상관 계수 및 유의도는 동일 방식으로 불변화사/보조어 n 기록에 대해서 계산된다(단계 540 및 550). 결과적으로, R = (r1, r2, ...rd) 및 E = (e1, e2,...ed)가 계산된다(단계 560). 여기서, d는 내용어 n 기록와 불변화사/보조어 n 기록의 총 개수를 나타낸다.

이러한 실시예에서, 스피어먼의 상관 계수와 그들의 유의도는 4 종류의 내용어 n 기록, 즉, 내용어 단일 기록, 내용어 이중 기록, 내용어 3중 기록, 및 내용어 스킵 이중 기록에 대해서 계산된다. 따라서, 8개의 특징값이 스피어먼의 상관 계수 및 그들의 유의도로서 계산된다. 유사하게, 불변화사/보조어에 기초하여 8개의 특징값이 스피어먼의 상관 계수 및 그들의 유의도로서 계산된다. 상술한 5개의 특징값을 추가하면, 21(= 5 + 8 + 8) 특징값이 모두 사용된다.

다음에, 입력 문서를 구별하여 입력 문서를 소인 문서 또는 전문가 문서로 분류하기 위해서 마할라노비스(Mahalanobis) 거리 함수가 사용된다. 도 6은 도 2의 컴퓨터가 입력 문서를 소인 문서, 전문가 문서 또는 다른 문서로 분류하기 위해서 수행하는 동작의 흐름도이다.

610 : 마스터 소인 문서와 마스터 전문가 문서에 대한 특징값을 계산

620 : 각각의 입력 문서에 대한 특징값을 계산

630 : 입력 문서와 마스터 소인 문서 간의 거리(Db) 및 입력 문서와 마스터 전문가 문서 간의 거리(Da)를 계산

640 : 입력 문서와 마스터 소인 문서 간의 거리가 임계값보다 작으면, 입력 문서는 소인 문서로서 분류된다.

650 : 입력 문서와 마스터 전문가 문서 간의 거리(Da)가 임계값보다 작으면, 입력 문서는 전문가 문서로서 분류된다.

660 : 마스터 소인 문서 또는 마스터 전문가 문서에 해당하지 않는 문서는 "다른" 문서로서 분류된다.

670 : 모든 문서가 분류됨을 확인

680 : 다음 문서를 지정

각각의 단계가 이하에 상세히 설명된다. 먼저, 마스터 소인 문서 및 마스터 전문가 문서에 대한 특징값이 계산된다. 이들 특징값은, 문서를 판별하기 위해 판별식이 사용된 경우에 각각의 세트에 대한 기본 모집단을 구성한다. 마스터 소인 문서는 "마스터 소인 문서 세트"로부터 선택된 마스터 소인 문서로서 현저한 특징을 가진 문서 세트이다. 마스터 소인 문서를 구성하는 각각의 문서의 특징값이 계산되며, 특징값의 평균값이 계산된다. 마스터 전문가 문서가 또한 "마스터 전문가 문서 세트"로부터 선택되며, 각각의 문서의 특징값이 계산되며, 특징값의 평균값이 동일한 방식으로 계산된다(단계 610).

다음에, 입력 문서의 특징값이 계산된다(단계 620). 입력 문서와 마스터 소인 문서 간의 거리(Db)가, 입력 문서의 특징값과 마스터 소인 문서의 특징값을 사용함으로써, 마할라노비스 공식(식 1)을 이용하여 계산된다. 유사하게, 입력 문서와 마스터 전문가 문서 간의 거리(Dc)가, 입력 문서의 특징값과 마스터 전문가 문서의 특징값을 이용하여 마할라노비스 공식(식 2)으로 계산된다(단계 630).

(표현식 1) $Db=(A-B)^t\sum b^{-1}(A-B)$

(표현식 2) $Dc=(A-C)^t\sum c^{-1}(A-C)$

여기서, A는 각각의 문서로부터 획득된 특징값을 나타내고, A^t = (a1, a2, ...ap)로서 표현되며, B는 소인 문서의 특징값의 평균값을 나타내고, B^t = (b1, b2, ...bp)로서 표현되며, C는 전문가 문서의 특징값의 평균값을 나타내고, C^t = (c1, c2, ...cp)로서 표현되며, p는 특징 벡터의 거리의 개수를 나타내며, t는 행렬의 전치를 나타낸다. $\sum b$ 및 $\sum c$은 각각의 세트의 공분산(covariance) 행렬을 나타내며, $\sum b^{-1}$ 및 $\sum c^{-1}$은 공분산 행렬의 역행렬을 나타낸다.

Db가 소정의 임계값보다 작으면, 문서는 소인 문서로서 분류된다(단계 640). Dc가 소정의 임계값보다 작으면, 문서는 전문가 문서로서 분류된다(단계 650).

소인 문서 또는 전문가 문서 어느 것으로도 분류되지 않은 문서는 분류 불가능으로 간주되어 "다른" 문서로서 간주된다(단계 660).

상술한 단계는 모든 문서에 대해서 실행된다(단계 670 및 680).

(6) 소인 문서와 전문가 문서 간의 대응 용어를 검출

상술한 처리의 결과로서, 특정의 공통 주제를 기술하는 N 문서와 E 문서로 구성된 문서 쌍이 획득될 수 있다. N(소인) 문서와 E(전문가) 문서에 사용되는 용어 간의 상관 관계가 이하에 설명된다.

소인 문서(N 문서)와 전문가 문서(E 문서)에는 상이한 용어가 사용된다. 그러나, 이들 문서는 공통의 주제를 기술하고 있기 때문에, 유사한 의미를 가진 대응 용어가 사용됨을 짐작할 수 있다. 따라서, 유사한 의미를 가진 단어 쌍의 E 문서 및 N 문서로부터의 식별 방법이 전개될 것이다. 그 방법은 E 문서 내의 r번째 단어(Er)에 해당하는 소인 단어의 리스트를 검출하고, N 문서의 i번째 단어(Ni)에 해당하는 전문가 단어의 리스트를 검출한다.

(I) 최대 우도 비율 시험

먼저, 최대 우도 비율 시험을 이용한 계산 방법이 기술된다. 도 7은 도 2의 컴퓨터가 최대 우도 비율(MLR) 시험과 결합하여 수행하는 동작의 흐름도이다.

710 : 소인 문서로서 분류된 문서에 대한 각각의 용어의 빈도수를 계산

720 : 전문가 문서로서 분류된 문서에 대한 각각의 용어의 빈도수를 계산

730 : P(A) = Prob(Ni AND Er)를 계산

740 : P(B) = Prob(NOT(Ni) AND Er)를 계산

750 : P(A) 및 P(B)에 기초하여 MLR을 계산

760 : 임계값을 초과하는 MLR으로 (Ni)와 (Er)의 조합을 추출

770 : 모든 조합에 대해서 계산됨을 확인

780 : 다음 조합을 지정

790 : 양 방향으로부터 대응 용어를 검출

도 1의 시스템이 최대 우도 비율을 검출하기 위해 사용하는 방법이 도 7의 흐름도를 기준으로 상세히 설명된다.

다음의 상황을 고려한다. (1) m 용어가 문서 N으로부터 추출되며, N의 i번째 용어는 Ni라고 가정하고, (2) n 용어가 문서 E로부터 추출되며, 그 r번째 용어는 Er이며, (3) Ni 및 Er는 빈번하게 함께 발생한다고 가정한다. 환언하면, Ni가 빈번하게 발생할 때 Er이 빈번하게 발생하며, Ni가 드물게 발생할 때 Er이 드물게 발생한다고 가정한다. 이러한 상황의 확률이 너무 높아서 동시 발생으로 간주되지 않는다고 판단되는 조건이 기술된다. 추가로, 수식값으로 확률의 신뢰도를 표현하는 방법이 기술될 것이다.

소인 용어(N 문서의 용어)에 대한 대응 전문가 용어(E 문서의 용어)를 찾는 방법이 아래에 기술된다.

하나의 주제에 기초하여 추출되어 소인 문서 또는 전문가 문서로서 분류된 문서 쌍을 고려한다. 소인 문서와 전문가 문서의 모든 용어를 처리하기 보다는, 처리하고자 하는 용어가 미리 결정된다. 이러한 용도로 준비된 소인 용어 리스트와 전문가 용어 리스트는 각각의 영역에 대응하는 용어를 저장한다. 소인 용어 리스트는 사람의 감각 및 주관적인 판단에 관련된 표현을 저장한다.

전문가 용어 리스트는 다음의 기준을 만족하는 용어를 저장한다.

(a) 용어 리스트에 포함된 용어 및 이들 용어에 관련된 용어

(b) 소인 용어 리스트에 포함되지 않은 용어

(c) 소정의 빈도수와 같거나 높은 빈도수로 나타나는 용어

소인 문서에 나타나는 소인 문서 리스트로부터 n 용어가 있으며, 소인 용어 리스트의 i번째 용어는 Ni(i = 1 내지 m)이다고 가정한다. i번째 용어의 빈도수가 계수된다(단계 710). 유사하게, 소인 용어 리스트 내의 용어 중에서 전문가 문서에 m 용어가 있으며, 전문가 리스트의 r번째 용어가 Er(r = 1 내지 n)이다고 가정한다. 전문가 용어 리스트의 r번째 용어의 빈도수가 계수된다(단계 720). 빈도수를 계수하는 단위는 용어 단일 기록, 용어 이중 기록 또는 용어 3중 기록 중 하나이다. Ni 및 Er P(A)의 동시 발생의 확률(단계 730) 및 Ni의 발생 및 Er P(B)의 미발생의 확률(단계 740)은 각각의 문서에서의 Ni 및 Er의 빈도수에 기초하여 다음과 같이 정의된다.

$$P(A)=Prob\,(Ni|Er)$$

P(B)=Prob (Not (Ni)|Er)

다음에, 최대 우도 비율(MLR)이 계산된다(단계 750). MLR은 (1) P(A)와 P(B)사이에 차이가 없다고 가정(귀무가설)되는 경우의 확률인 확률 P(HO)와, (2) 차이가 있다고 가정(대립가설)되는 경우의 확률인 확률 P(H1)의 비율로서 계산된다. MLR은, 질문의 용어 쌍(Ni 및 Er)이 이항 분포에 따른 2개의 랜덤 처리인 것으로 고려함으로써 계산된다. 하나의 랜덤 변수에 대한 이항 분포 확률을 계산하기 위한 표현이 이하에 주어진다.

$$b(p,k,n)=\binom{n}{k}p^k(1-p)^{(n-k)} \quad \text{(공식 3)}$$

여기서, k는 특정 단어의 실제 출현 회수를 나타내며, n은 단어의 최대 가능 출현 회수를 나타내며, p는 기본적인 출현 확률을 나타낸다. H0의 경우(귀무가설)에서의 가정된 확률이 p0이면, H1의 경우(대립가설)의 P(A)의 가정된 최대 확률은 p1이며, P(B)의 가정된 최대 확률은 p2이며, P(H0)와 P(H1)의 비율은 다음과 같이 표현된다.

$$\lambda=\frac{P(H0)}{P(H1)}=\frac{b(p0,k1,n1)b(p0,k2,n2)}{b(p1,k1,n1)b(p2,k2,n2)} \quad \text{(공식 4)}$$

k1, n1, k2 및 n2의 값은 단어의 출현 회수로부터 용이하게 계산된다. 확률비에 대한 MLR은 다음과 같다.

$$MLR=-2\log\lambda \quad \text{(공식 5)}$$

MLR은 1의 자유도를 가진 카이 제곱 분포(chi-squared distribution)를 실질적으로 따르는 것으로 일반적으로 알려져 있다. 이것을 이용하면, 임계값을 지정하는 것이 용이하다. 환언하면, MLR의 값이 특정의 수치값을 초과하면, 너무 높아서 동시 발생인 것으로 간주되지 않는 확률에서 2개의 용어 Ni 및 Er이 동시 발생한다고 말할 수 있다(단계 760).

상술한 이론을 이용하면, 도 2의 컴퓨터는 다음의 방법을 이용하여 사전적 매핑을 위한 후보를 선택한다. 모든 타겟 용어의 조합, 즉, {(Ni, Er) i = 1 내지 m, r = 1 내지 n}과 관련하여 MLR을 계산한 후에(단계 770, 780), 수치값의 내림 차수로 소정의 임계값, 예를 들어, 5%의 레벨을 초과하는 쌍을 선택한다. 임계값을 초과하는 MLR의 값을 가진 N의 i번째 용어에 대응하는 전문가 리스트 용어가 검색되며, 그 용어 중에서 소정수의 용어가 MLR의 내림 차수로 선택되어, 소인 용어에 대응하는 전문가 용어가 획득된다(단계 780).

전문가 용어(E 문서의 용어)로부터 대응 소인 용어(N 문서의 용어)를 찾기 위해서 도 2의 컴퓨터가 이용하는 방법이 기술된다.

상술한 바와 유사한 방식으로, 임계값을 초과하는 MLR의 값을 가진 E의 r번째 용어에 대응하는 N의 용어는 그 저장된 리스트로부터 검색되며, 그 용어 중에서 소정수의 용어가 MLR의 값의 내림 차수로 선택되어, 전문가 용어에 해당하는 소인 용어가 획득된다(단계 780).

(ii) 사전적 매핑 행렬의 계산에 기초한 방법

다음에, 문서의 길이 및 용어 빈도수에 따라서 가중 조절을 이용한 사전적 매핑 행렬(T)의 계산에 기초한 방법이 기술된다.

도 9는 도 1의 시스템이 사전적 매핑 행렬과 관련해서 수행하는 동작의 흐름도이다.

810 : s ×n 전문가 용어 행렬(P)을 형성

820 : s ×m 소인 용어 행렬(Q)을 형성

830 : m ×n 사전적 매핑 행렬(T)을 계산

840 : 소인 용어를 전문가 용어로 변환 및 전문가 용어를 소인 용어로 변환

각각의 단계 810 - 840가 이하에 상세히 설명된다. 먼저, 전문가 용어 행렬(P)이 전문가 문서로서 분류된 문서 세트로부터 형성된다. 키워드로서 용어 리스트의 k번째 용어(k = 1 내지 s)로 검색된 문서가 여기서 고려된다. 전문가 문서로서 분류된 이들 문서는 처리되어 문서에 사용된 용어의 빈도수를 계산한다.

처리된 용어는 상술한 전문가 용어 리스트 내의 용어이다. 용어 리스트 내의 모든 용어에 대해 검색되고 전문가 문서로서 분류된 문서에 상술한 동작이 적용되어, 전문가 용어 리스트 내의 용어에 대응하는 용어의 빈도수가 계산된다. n이 전문가 문서 내의 용어의 개수이다고 가정하면, 전문가 용어의 빈도수를 나타내는 s ×n 행렬(P_0)(도시 생략)이 계산된다.

유사하게, m은 소인 문서 내의 용어의 개수인 것으로 가정하면, 소인 용어의 빈도수를 나타내는 s ×m 행렬(Q_0)(도시 생략)이 계산된다.

서로 동시 발생된 2개의 단어 간의 결합 강도가 높지만, 높은 빈도수의 단어가 종종 여러 다른 단어와 동시 발생한다. 이러한 이유로, 사전적 매핑을 위해서 후보로서의 이러한 단어의 중요성을 감하는 것이 중요하다. 유사하게, 하나의 문서가 장문이고 상당수의 단어를 포함하고 있을 때, 이러한 문서에 발생한 단일 단어의 중요성은 감해져야 한다.

따라서, 다음과 같이, 행렬(P_0)의 원소를 변환함으로써 s ×n 전문가 용어 행렬(도 8a)이 형성된다(단계 810).

$$We(k,i) = \frac{Exp(k,i)}{(\mathrm{Etf(i)} * \mathrm{Ewf(k)})}$$

여기서, 전문가 문서의 k번째 문서에 나타난 단어의 빈도수는 Exp(k, i)이며, 모든 문서에서의 그 단어의 빈도수는 Etf(i)이며, k번째 문서에 나타난 단어의 총 개수는 Ewf(k)이다.

유사하게, 다음과 같이, 행렬(Q_0)의 원소를 변환함으로써 s ×m 소인 용어 행렬(Q)(도 8b)이 형성된다(단계 820).

$$Wn(k,i) = \frac{\mathrm{Naive(k,r)}}{(\mathrm{Ntf(r)} * \mathrm{Nwf(k)})}$$

여기서, 소인 문서의 k번째 문서에 나타난 단어의 빈도수는 Naive(k, r)이며, 모든 문서에서의 그 단어의 빈도수는 Ntf(r)이며, k번째 문서에 나타난 단어의 총 개수는 Nwf(k)이다.

s ×n 행렬(P)과 s ×m 행렬(Q)을 형성하는 목적은 이들 각각의 단어의 조합의 강도를 표시하는 가중값을 계산하여 m ×n 사전적 매핑 행렬(T)을 획득하기 위한 것이다. 따라서, 행렬(T)은 다음과 같이 정의된다.

$T = Q^tP$

여기서, t는 행렬의 전치이며, 사전적 매핑 매트릭스(T)의 각각의 가중값은 다음과 같이 정의된다.

$$W(r,i) = \sum_{k=1}^{s} \left[\frac{Exp(k,i)}{(\mathrm{Etf(i)} * \mathrm{Ewf(k)})} \frac{\mathrm{Naive(k,r)}}{(\mathrm{Ntf(r)} * \mathrm{Nwf(k)})}\right]$$

매핑의 후보 단어는 사전적 매핑 행렬로부터 추출된다. 예를 들어, i번째 소인 용어의 Ni에 대응하는 전문가 용어의 후보를 추출하기 위해서, 사전적 매핑 행렬(T)의 i번째 행을 지칭하여 가중값의 내림 차수로 원하는 수의 용어를 선택하기에 충분하다(단계 840).

한편, r번째 전문가 용어에 대응하는 소인 용어의 후보를 추출하기 위해서, 사전적 매핑 행렬(T)의 r번째 행을 지칭하여 가중값의 내림 차수로 원하는 수의 용어를 선택하기에 충분하다(단계 840). 둘 다의 경우에, 0의 값을 가진 단어를 제외한, 최상위 가중값을 가진 10개의 단어가 바람직하게 후보 단어로서 선택된다.

그러나, 10개의 선택된 후보 단어는 불필요한 정보를 포함할 수 있기 때문에, 본 방법이 반드시 실용적인 것은 아니다. 따라서, 용어 리스트에 포함된 용어를 이용하여 후보 용어를 추가로 필터링하는 방법이 사용될 수 있다. 예를 들어, 용어 리스트에 기술된 "와인 이름"에 대한 데이터만이 출력될 것이다. 또한, 비전문가의 선호 정보를 만족하는 소인 용어 후보를 선택하는 것이 또한 가능하다. 예를 들어, "드라이(dry)", "좋은 감촉(good texture)" 및 "감식력이 있는(tasteful)"과 같은 단일 문서를 가진 선호 정보를 나타내는 비전문가 용어 또는 비전문가를 나타내는 용어의 이중 기록 조합에 대응하는 "와인 이름"을 출력하는 것이 가능하다. 결과적으로, 비전문가 선호 정보에 기초하여 비전문가의 선호에 일치하는 "와인 이름"을 찾을 수 있다. 이러한 필터링의 적용 후의 출력의 예가 이하에 논의된다.

검색의 샘플 결과가 이하에 표시된다.

다음의 예는 소인 용어에 대응하는 용어로서 검색된 전문가 용어의 샘플이다. 일본 인터넷 사이트에서 영역 "니혼슈(nihonshu : Japanese rice wine)"를 조사하였을 때, 다음의 소인(비전문가) 용어, 즉, "아츠이(atsui : heavy)", "유타카(yutaka : rich)", "탄레이(tanrei : light and fine)", "사라리토(sararito : smooth)", "비미(bimi : tasty)", "후카미(fukami : depth)" 등이 검색되었다. 이들 소인 용어에 대응하는 전문가 용어가 이하의 와인 이름, 즉 "heavy" 및 "rich"에 대해서 "이소지만(Isojiman)", "light and fine" 및 "smooth"에 대해서 "고시노칸바이(Koshinokanbai)" 및 "tasty" 및 "depth"에 대해서 "가모미도리(Kamomidori)" 각각으로 검색되었다.

"와인" 영역이 일본 인터넷 사이트에서 조사되었을 때, 다음의 비전문가 용어, 즉 "비미(bimi : tasty)", "코이(koi : thick)", "우마미(umami : tastiness)", "스파이(suppai : sour)", "시타자와리(shitazawari : texture)", "기레(kire : sharpness)", "피타리(pittari : exact fit)", "후카미(fukami : depth)", "사와야카(sawayaka : fresh)", "야와라카(yawaraka : soft)", "마로야카(maroyaka : smooth and soft)" 등이 검출되었다. 이들 소인 용어에 대응하는 전문가 용어가 다음의 와인 이름, 즉 "tasty", "thick", "tastiness", "sour" 등에 대해서 "Au Bon Climat", 및 "texture", "sharpness", "fit", "depth", "fresh", "soft", "smooth and soft" 등에 대해서 "Zonnebloem"으로 검색되었다.

다음의 예는 전문가 용어에 대응하는 용어로서 검출된 샘플의 소인 용어이다.

일본 인터넷 사이트에서 "니혼슈(nihonshu : Japanese rice wine)의 영역이 조사되었을 때, 와인 이름인 다음의 전문가 용어, "가가토비(Kagatobi)", "하나노마이(Hananomai)", "가쿠부토(Kakubuto)" 등이 검출되었다. 이들 와인 이름에 대응하는 용어로서 검색된 소인 용어는 "가가토비"에 대해 "오이시이(oishii : delicious), 미즈미즈시이(mizumizushii : refreshing)", "하나노마이"에 대해 "조힌(johin : elegant), 탄레이(tanrei : light and fine)", 및 "가쿠부토"에 대해 "나메라카(nameraka : soft and mellow), 사와야카(sawayaka : cool and fresh), 스바라시이(subarashii : wonderful)"를 포함한다.

일본 사이트에서 "와인" 영역이 조사되었을 때, 와인 이름인 다음의 전문가 용어, "골타살라(Coltassala)", "산소니어레(Sansoniere)" 등이 검출되었다. 이들 와인 이름에 대응하는 용어로서 검색된 소인 용어는 "골타살라"에 대해 "아와이(awai : translucent), 기힌(kihin : elegance), 호노카(honoka : faint), 가루이(karui : light), 고코치요이(kokochiyoi : comfortable)", 및 "산소니어레"에 대해 "호로니가이(horonigai : slightly bitter), 가라이(karai : dry), 조힌(johin : elegant), 유유가(yuuga : grace)"가 검출되었다.

발명의 효과

상술한 사전적 매핑 방법 둘 다를 이용하면, 그들의 가중값의 내림 차수로 용어를 선택함으로써, N->E(비전문가 - 전문가) 및 E->N(전문가 - 비전문가)의 양 방향으로 특정 용어에 대응하는 후보 용어를 선택하는 것이 가능하다.

(57) 청구의 범위

청구항 1.

공통의 주제를 가진 문서를 검색하고, 상기 문서를 제 1 특징값 세트를 가진 제 1 문서 세트와 제 2 특징값을 가진 제 2 문서 세트로 분류하는 방법에 있어서,

사전결정된 용어 리스트에 기초하여 관련 제 3 문서 세트를 검색하는 단계와,

상기 제 3 문서 세트에서 각각의 문서에 대한 특징값을 계산함으로써, 제 3 특징값 세트를 구성하는 단계와,

(a) 상기 제 1 특징값 세트와 상기 제 3 특징값 세트를 이용한 판별과, (b) 상기 제 2 특징값 세트와 상기 제 3 특징값 세트를 이용한 판별에 따라서, 상기 제 3 문서 세트 내의 문서를 상기 제 1 문서 세트와 제 2 문서 세트로 분류하는 단계

를 포함하는 방법.

청구항 2.

제 1 항에 있어서,

상기 특징값 세트로서, 다음의 아이템, 즉, 내용어의 개수, 소인(naive) 용어의 비율, 적절한 명사의 비율, 추가의 적절한 명사의 비율, 불변화사/보조어의 비율, 내용어와 불변화사/보조어에 관련된 n 기록 패턴의 빈도수로부터 계산된 스피어먼 상관 계수/유의도로부터 임의의 아이템 세트를 선택하는 단계를 더 포함하는 방법.

청구항 3.

제 2 항에 있어서,

상기 제 3 문서 세트의 검색은 가비지 문서, 리스트 문서 및 일기 문서 중 적어도 하나에 속하는 문서를 제거하는 단계를 더 포함하는 방법.

청구항 4.

제 1 항에 있어서,

상기 제 3 문서 세트의 검색은 가비지 문서, 리스트 문서 및 일기 문서 중 적어도 하나에 속하는 문서를 제거하는 단계를 더 포함하는 방법.

청구항 5.

(제 1 특징값을 가진 제 1 문서 세트와 제 2 특징값을 가진 제 2 문서 세트로부터) 상기 제 1 및 제 2 문서 세트가, (a) 공통의 주제, (b) 상기 제 1 문서 세트 내의 특정 용어에 대응하는 상기 제 2 문서 세트 내의 용어, 또는 (c) 상기 제 2 문서 세트 내의 특정 용어에 대응하는 상기 제 1 문서 세트 내의 용어 중 적어도 하나를 가지고 있음을 검출하는 방법에 있어서,

사전 결정된 용어 리스트에 기초하여 관련 제 3 문서 세트를 검색하는 단계와,

상기 제 3 문서 세트 내의 각각의 문서에 대한 특징값을 계산함으로써, 제 3 특징값 세트를 구성하는 단계와,

상기 제 1 특징값 세트와 상기 제 3 특징값을 이용한 판별과, 상기 제 2 특징값과 상기 제 3 특징값을 이용한 판별에 따라서 상기 제 3 문서 세트 내의 문서를 상기 제 1 문서 세트 또는 상기 제 2 문서 세트로 분류하는 단계와,

상기 제 1 문서 세트로 분류된 문서로부터 컴파일된 제 1 용어 리스트에 리스트된 각 용어의 빈도수와, 상기 제 2 문서 세트로 분류된 문서로부터 컴파일된 제 2 용어 리스트에 리스트된 각 용어의 빈도수를 계산하는 단계와,

상기 제 1 및 제 2 용어 리스트에 리스트된 용어의 빈도수에 기초하여, 상기 제 1 문서 세트 내의 특정 용어에 대응하는 상기 제 2 문서 세트 내의 용어를 검출하는 단계와,

상기 제 1 및 제 2 용어 빈도수에 기초하여 상기 제 2 문서 세트 내의 특정 용어에 대응하는 상기 제 1 문서 세트 내의 용어를 검출하는 단계

를 포함하는 방법.

청구항 6.

(공통의 주제를 가진 제 1 문서 세트와 제 2 문서 세트로부터) (a) 상기 제 1 문서 세트 내의 특정 용어에 대응하는 상기 제 2 문서 세트 내의 용어, 또는 (b) 상기 제 2 문서 세트 내의 특정 용어에 대응하는 상기 제 1 문서 세트 내의 용어를 검출하는 방법에 있어서,

상기 제 1 문서 세트로부터 컴파일된 제 1 용어 리스트에 리스트된 각 용어의 빈도수와, 상기 제 2 문서 세트로부터 컴파일된 제 2 용어 리스트에 리스트된 각 용어의 빈도수를 계산하는 단계와,

상기 제 1 및 제 2 용어 리스트에 리스트된 용어의 상기 빈도수에 기초하여, 상기 제 1 문서 세트 내의 특정 용어에 대응하는 상기 제 2 문서 세트 내의 용어를 검출하는 단계와,

상기 제 1 및 제 2 용어 리스트에 리스트된 용어의 상기 빈도수에 기초하여, 상기 제 2 문서 세트 내의 특정 용어에 대응하는 상기 제 1 문서 세트 내의 용어를 검출하는 단계

를 포함하는 방법.

청구항 7.

(용어 리스트에 기초하여 검색되며 공통의 주제를 가진 제 1 문서 세트와 제 2 문서 세트로부터) (a) 상기 제 1 문서 세트 내의 특정 용어에 대응하는 상기 제 2 문서 세트 내의 용어, 및 (b) 상기 제 2 문서 세트 내의 특정 용어에 대응하는 상기 제 1 문서 세트 내의 용어를 검출하는 방법에 있어서,

상기 제 1 문서 세트로부터의 용어와 상기 제 2 문서 세트로부터의 용어를 포함하는 특정 용어 쌍의 동시 발생의 확률 P(A)을 계산하는 단계와,

상기 제 1 문서 세트에는 발생하는 질문의 상기 용어 쌍의 제 1 용어와 상기 제 2 문서 세트에는 발생하지 않는 상기 용어 쌍의 제 2 용어의 동시 발생의 부족의 확률 P(B)를 계산하는 단계와,

P(A) 및 P(B)에 기초하여 최대 우도 비율을 계산하는 단계와,

사전 설정된 임계값을 초과하는 최대 우도 비율을 가진 모든 용어 쌍의 조합을 추출하는 단계와,

상기 제 2 문서 세트 내의 특정 용어에 대응하는 상기 제 1 문서 세트 내의 용어로부터 상기 최대 우도 비율 값의 내림 차수로 사전 설정된 수의 용어를 선택하고, 상기 선택된 용어를 상기 제 2 문서 세트 내의 특정 용어에 대응하는 상기 제 1 문서 세트의 후보 용어로서 채택하는 단계와,

상기 제 1 문서 세트 내의 특정 용어에 대응하는 상기 제 2 문서 세트 내의 용어로부터 상기 최대 우도 비율 값의 내림 차수로 사전 설정된 수의 용어를 선택하고, 상기 선택된 용어를 상기 제 1 문서 세트 내의 특정 용어에 대응하는 상기 제 2 문서 세트의 후보 용어로서 채택하는 단계

를 포함하는 방법.

청구항 8.

(공통의 주제를 가진 제 1 문서 세트와 제 2 문서 세트로부터) (a) 상기 제 1 문서 세트 내의 특정 용어에 대응하는 상기 제 2 문서 세트 내의 용어, 및/또는 (b) 상기 제 2 문서 세트 내의 특정 용어에 대응하는 상기 제 1 문서 세트 내의 용어를 검출하되, 상기 제 1 및 상기 제 2 문서 세트는 용어 리스트에 기초하여 검색되는 방법에 있어서,

제 1 용어 리스트에 리스트된 각 용어의 빈도수에 기초하여 상기 제 1 문서 세트로부터 제 1 용어 행렬을 형성하는 단계와,

제 2 용어 리스트에 리스트된 각 용어의 빈도수에 기초하여 상기 제 2 문서 세트로부터 제 2 용어 행렬을 형성하는 단계와,

상기 제 1 용어 행렬과 상기 제 2 용어 행렬의 곱으로부터 사전적 매핑 행렬을 계산하는 단계와,

원소값의 내림 차수로 상기 사전적 매핑 행렬의 특정 행 내의 사전 설정된 수의 용어를 선택하여, 상기 선택된 용어를 상기 제 2 문서 세트 내의 특정 용어에 대응하는 상기 제 1 문서 세트 내의 용어로서 채택하는 단계와,

원소의 내림 차수로 상기 사전적 매핑 행렬의 특정 열 내의 사전 설정된 수의 용어를 선택하여, 상기 선택된 용어를 상기 제 1 문서 세트 내의 특정 용어에 대응하는 상기 제 2 문서 세트 내의 용어로서 채택하는 단계

를 포함하는 방법.

청구항 9.

제 8 항에 있어서,

(a) 상기 용어 리스트 내의 용어의 개수는 s이며, (b) 상기 제 1 문서 세트로부터 선택된 용어의 개수는 n이며, (c) 상기 제 1 용어 행렬은 s ×n 행렬(P)로 표현되며, (d) 상기 제 1 문서 세트의 k번째 문서 내의 i번째 용어의 빈도수는 Exp(k, i)이며, (e) 상기 i번째 용어의 전체 빈도수는 Etf(i)이며, (f) 상기 k번째 문서 내의 용어의 총 개수는 Ewf(k)이며, 상기 행렬(P)의 원소는

$$We(k,i) = \frac{Exp(k,i)}{(\mathrm{Etf(i)} * \mathrm{Ewf(k)})}$$

이며, (g) 상기 제 2 문서 세트로부터 선택된 용어의 개수는 m이며, (h) 상기 제 2 용어 행렬은 s ×m 행렬(Q)로 표현되며, (i) 상기 제 2 문서 세트의 k번째 문서에 나타나는 r번째 용어의 빈도수는 Naive(k, r)이며, (j) 상기 r번째 용어의 전체 빈도수는 Ntf(r)이며, 상기 k번째 문서 내의 용어의 총 개수는 Nwf(k)이며, 상기 행렬(Q)의 원소는

$$Wn(k,i) = \frac{\mathrm{Naive(k,r)}}{(\mathrm{Ntf(r)} * \mathrm{Nwf(k)})}$$

로 주어지는 방법.

청구항 10.

청구항 1의 방법을 수행하기 위한 문서 검색 및 분류 시스템.

청구항 11.

청구항 2의 방법을 수행하기 위한 문서 검색 및 분류 시스템.

청구항 12.

청구항 3의 방법을 수행하기 위한 문서 검색 및 분류 시스템.

청구항 13.

청구항 4의 방법을 수행하기 위한 문서 검색 및 분류 시스템.

청구항 14.

청구항 5의 방법을 수행하기 위한 문서 처리 시스템.

청구항 15.

청구항 6의 방법을 수행하기 위한 문서 처리 시스템.

청구항 16.

청구항 7의 방법을 수행하기 위한 문서 처리 시스템.

청구항 17.

청구항 8의 방법을 수행하기 위한 문서 처리 시스템.

청구항 18.

청구항 9의 방법을 수행하기 위한 문서 처리 시스템.

청구항 19.

청구항 1의 방법을 컴퓨터가 수행하게 하는 메모리 또는 컴퓨터 판독가능 저장 매체.

청구항 20.

청구항 2의 방법을 컴퓨터가 수행하게 하는 메모리 또는 컴퓨터 판독가능 저장 매체.

청구항 21.

청구항 3의 방법을 컴퓨터가 수행하게 하는 메모리 또는 컴퓨터 판독가능 저장 매체.

청구항 22.

청구항 4의 방법을 컴퓨터가 수행하게 하는 메모리 또는 컴퓨터 판독가능 저장 매체.

청구항 23.

청구항 5의 방법을 컴퓨터가 수행하게 하는 메모리 또는 컴퓨터 판독가능 저장 매체.

청구항 24.

청구항 6의 방법을 컴퓨터가 수행하게 하는 메모리 또는 컴퓨터 판독가능 저장 매체.

청구항 25.

청구항 7의 방법을 컴퓨터가 수행하게 하는 메모리 또는 컴퓨터 판독가능 저장 매체.

청구항 26.

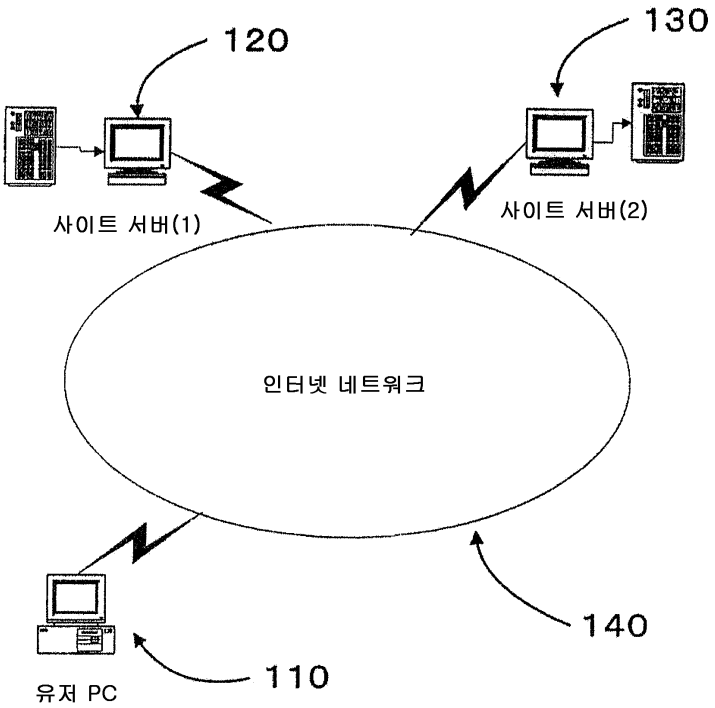
청구항 8의 방법을 컴퓨터가 수행하게 하는 메모리 또는 컴퓨터 판독가능 저장 매체.

청구항 27.

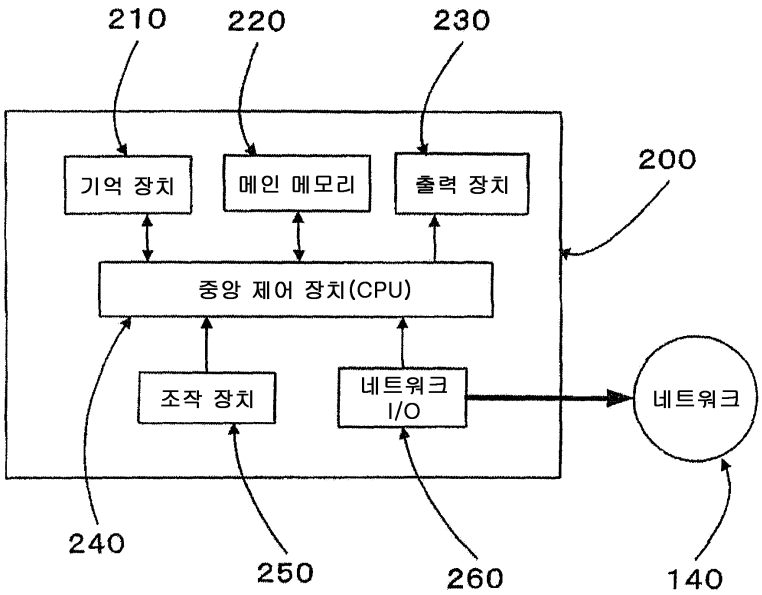
청구항 9의 방법을 컴퓨터가 수행하게 하는 메모리 또는 컴퓨터 판독가능 저장 매체.

도면

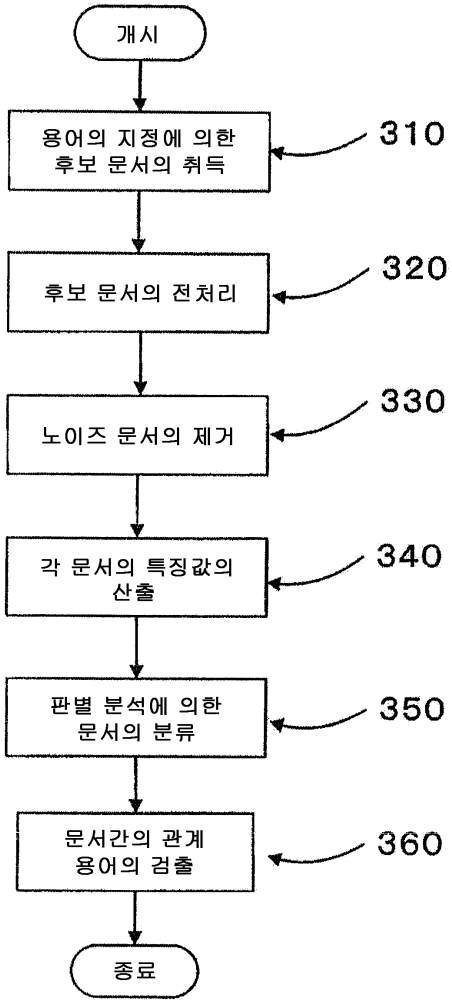
도면1


120
130
사이트 서버(1)
사이트 서버(2)
인터넷 네트워크
140
110
유저 PC

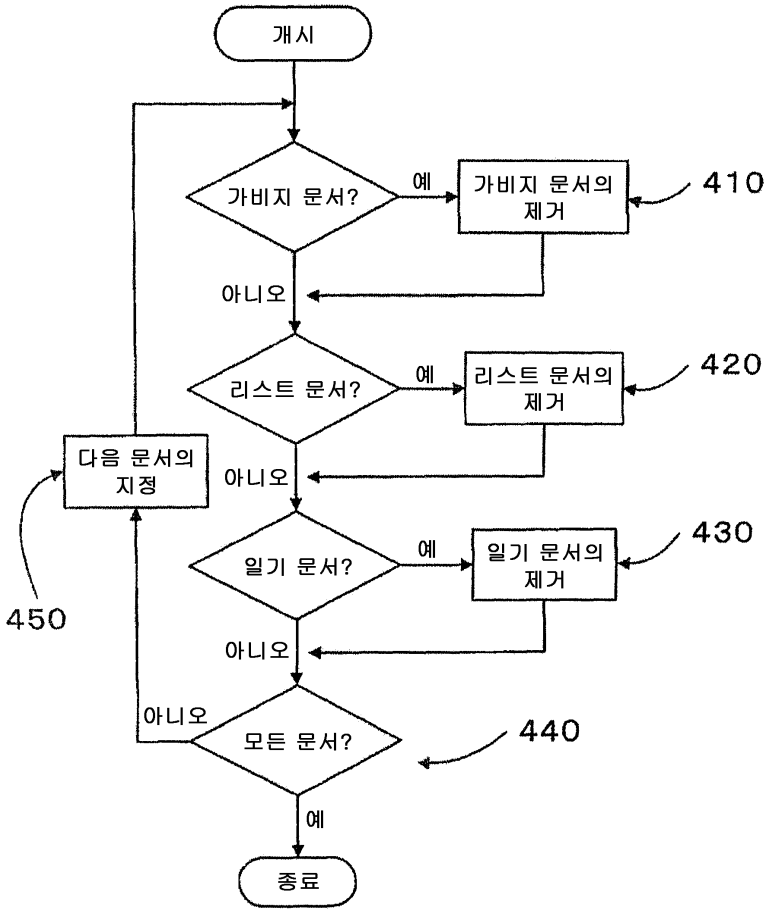
도면2


210
220
230
기억 장치
메인 메모리
출력 장치
200
중앙 제어 장치(CPU)
조작 장치
네트워크 I/O
네트워크
240
250
260
140

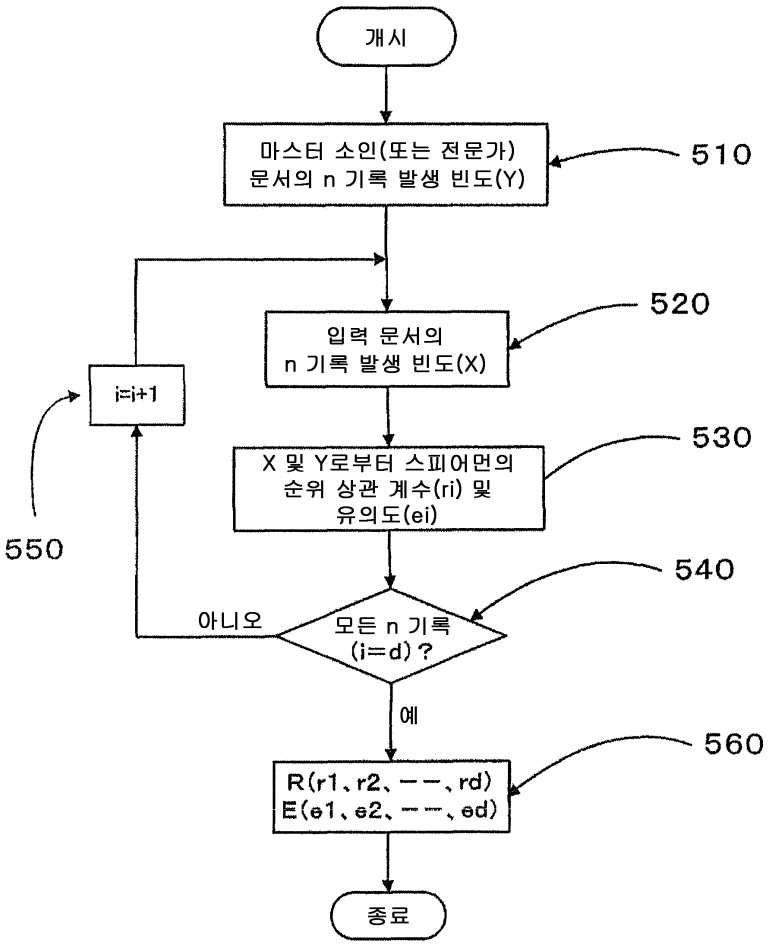
도면3


개시
용어의 지정에 의한 후보 문서의 취득
310
후보 문서의 전처리
320
노이즈 문서의 제거
330
각 문서의 특징값의 산출
340
판별 분석에 의한 문서의 분류
350
문서간의 관계 용어의 검출
360
종료

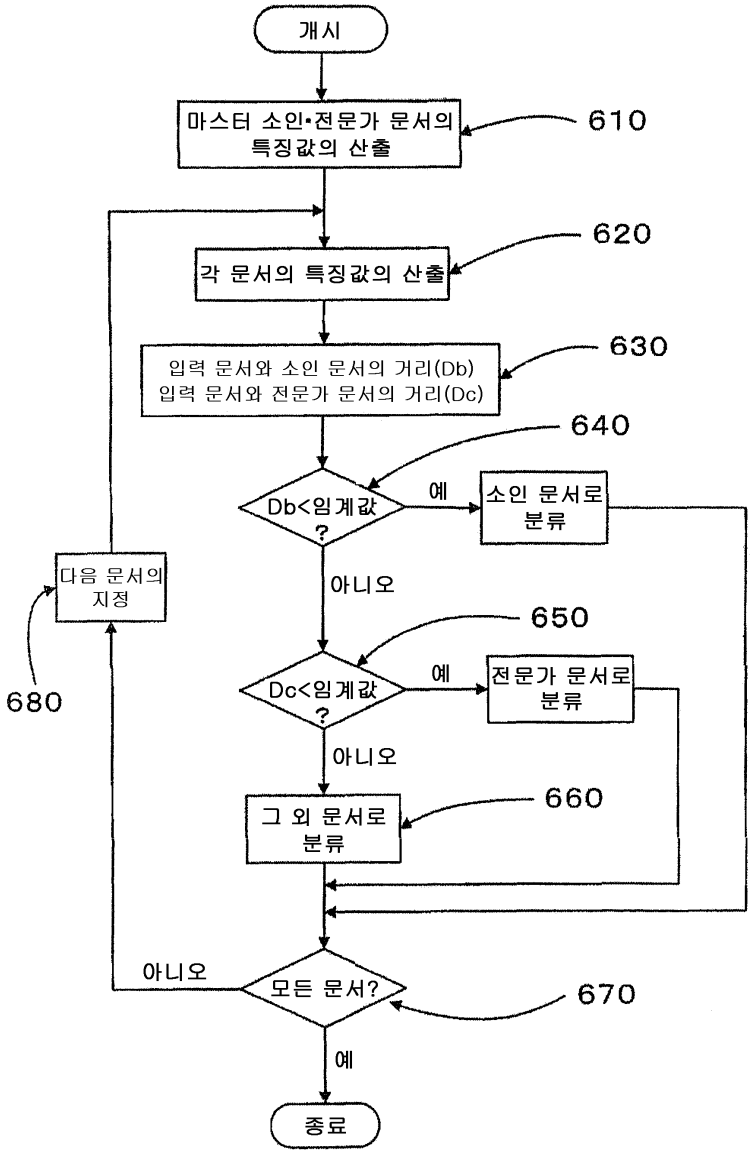
도면4


개시
가비지 문서?
예
가비지 문서의 제거
410
아니오
리스트 문서?
예
리스트 문서의 제거
420
다음 문서의 지정
아니오
일기 문서?
예
일기 문서의 제거
430
450
아니오
아니오
모든 문서?
440
예
종료

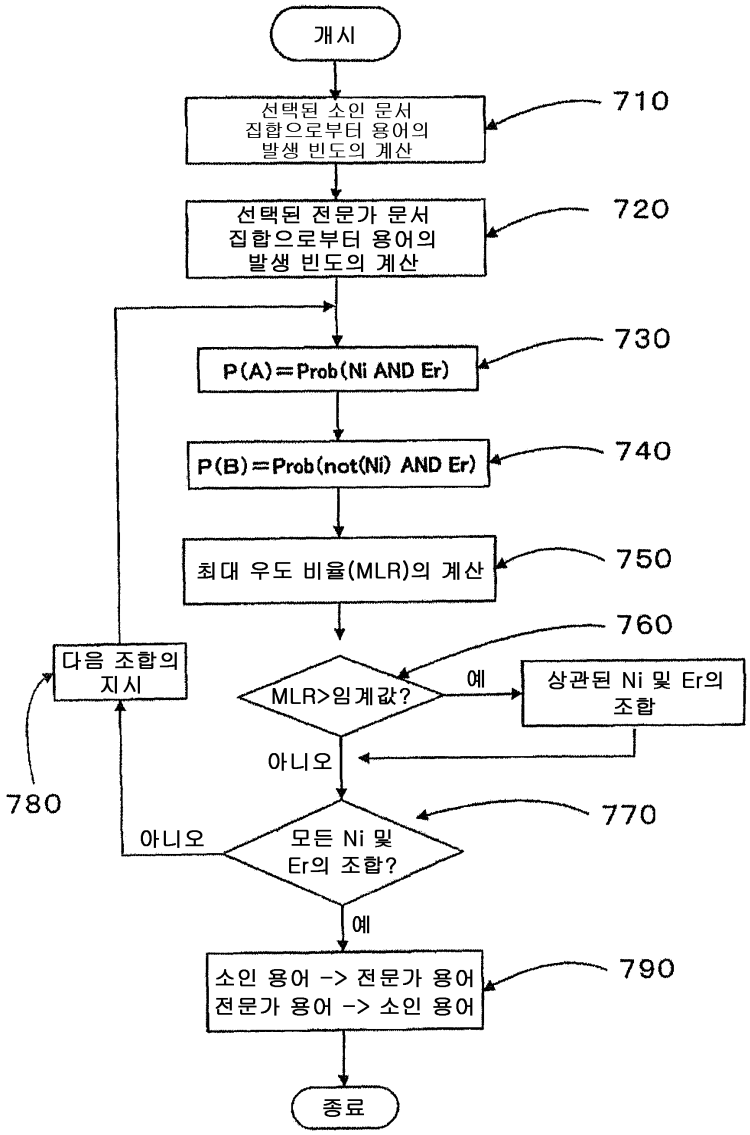
도면5


개시
마스터 소인(또는 전문가)
문서의 n 기록 발생 빈도(Y)
510
입력 문서의
n 기록 발생 빈도(X)
520
X 및 Y로부터 스피어먼의
순위 상관 계수(ri) 및
유의도(ei)
530
모든 n 기록
(i=d)?
540
아니오
예
i=i+1
550
R(r1、r2、--、rd)
E(e1、e2、--、ed)
560
종료

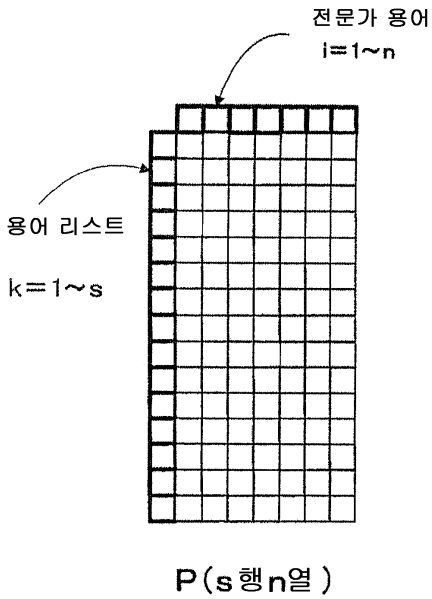
도면6


개시
마스터 소인·전문가 문서의 특징값의 산출
610
각 문서의 특징값의 산출
620
입력 문서와 소인 문서의 거리(Db)
입력 문서와 전문가 문서의 거리(Dc)
630
640
Db<임계값 ?
예
소인 문서로 분류
아니오
650
Dc<임계값 ?
예
전문가 문서로 분류
아니오
660
그 외 문서로 분류
680
다음 문서의 지정
아니오
모든 문서?
670
예
종료

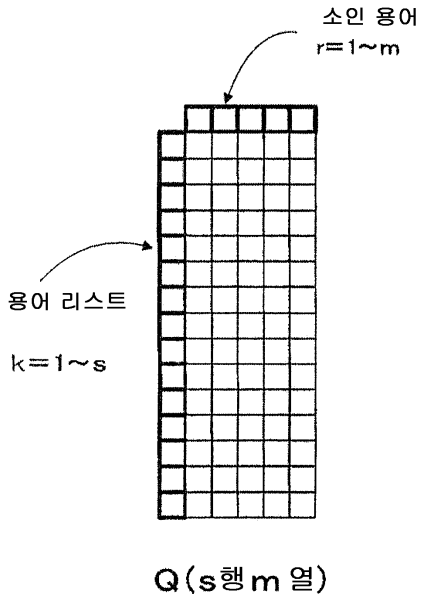
도면7


개시
선택된 소인 문서 집합으로부터 용어의 발생 빈도의 계산
710
선택된 전문가 문서 집합으로부터 용어의 발생 빈도의 계산
720
P(A)=Prob(Ni AND Er)
730
P(B)=Prob(not(Ni) AND Er)
740
최대 우도 비율(MLR)의 계산
750
760
MLR>임계값?
예
상관된 Ni 및 Er의 조합
아니오
다음 조합의 지시
780
아니오
모든 Ni 및 Er의 조합?
770
예
소인 용어 -> 전문가 용어
전문가 용어 -> 소인 용어
790
종료

도면8a

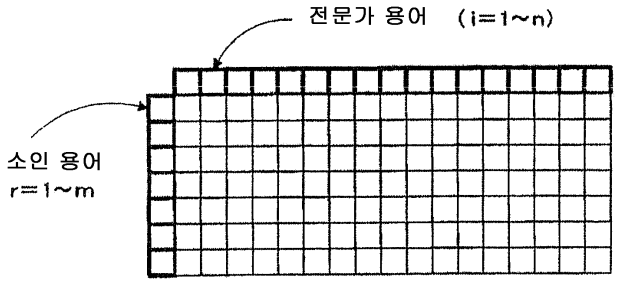

전문가 용어
i=1~n
용어 리스트
k=1~s
P(s행n열)

도면8b



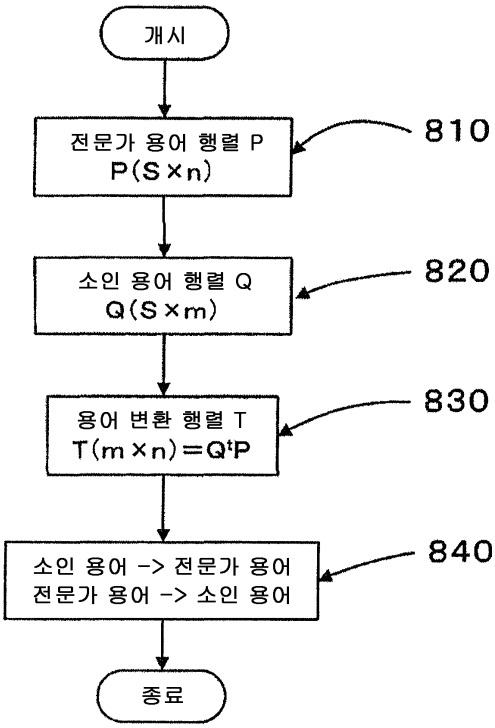
Q(s행 m 열)

도면8c



$T(m행\ n\ 열) = Q^tP$

도면9


개시
전문가 용어 행렬 P
P(S×n)
810
소인 용어 행렬 Q
Q(S×m)
820
용어 변환 행렬 T
T(m×n)=Q^tP
830
소인 용어 -> 전문가 용어
전문가 용어 -> 소인 용어
840
종료