(19) 대한민국특허청(KR)
(12) 등록특허공보(B1)

(45) 공고일자 2008년12월31일
(11) 등록번호 10-0876319
(24) 등록일자 2008년12월22일

(51) Int. Cl.
G06F 17/10 (2006.01)
(21) 출원번호 10-2007-0081006
(22) 출원일자 2007년08월13일
심사청구일자 2007년08월13일
(56) 선행기술조사문헌
KR100703193 B1*
KR1020020046558 A*
US20070050356 A1
US20030018604 A1
*는 심사관에 의하여 인용된 문헌

(73) 특허권자
인하대학교 산학협력단
인천 남구 용현동 253 인하대학교
(72) 발명자
이주홍
인천 남구 용현3동 인하대학교 컴퓨터공학부, 하이테크관 1406호
박선
전북 전주시 완산구 풍남동3가 3 17/5
(뒷면에 계속)
(74) 대리인
유완식, 이은철

전체 청구항 수 : 총 16 항 심사관 : 안병일

(54) 용어 가중치 재계산을 이용한 문서 군집 장치

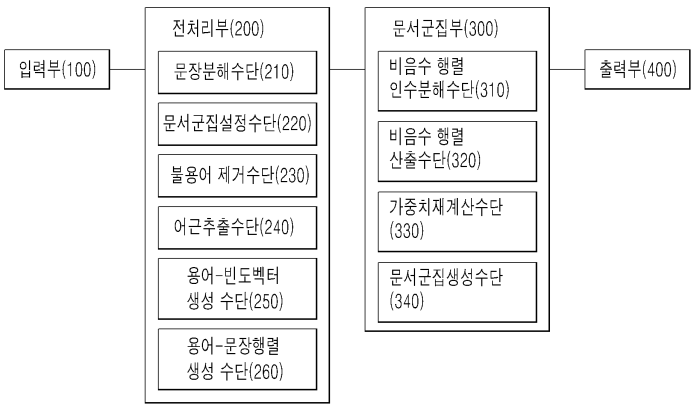
(57) 요 약

사용자 피드백 및 비음수 행렬 인수분해를 적용하여 사용자의 요구에 적합한 군집 결과를 도출하는 용어 가중치 재계산을 이용한 문서 군집 장치가 개시된다.

개시되는 용어 가중치 재계산을 이용한 문서 군집 장치는 상위 수준의 사용자 요구 및 하위 수준의 시스템 군집 간의 의미적 차이를 최소화하기 위하여 사용자 피드백에 의해 가중치가 재계산된 용어를 이용한다. 또한 문서 군집 과정에서 문서 집합의 내부 구조를 나타내는 의미 특징 행렬 및 의미 변수 행렬을 사용한다.

본 발명에 따르면, 사용자가 원하는 군집과 시스템에 의한 군집 사이의 의미적 차이를 최소화할 수 있으며, 군집의 내부 구조와 의미 특징의 분포를 쉽게 파악할 수 있고, 문서 군집의 정확도를 향상시킬 수 있는 효과가 있다.

대 표 도 - 도1



(72) 발명자

김덕환

인천 남구 용현동 인하대학교 하이테크관 814호

안찬민

인천 남구 주안2동 497-9번지

특허청구의 범위

청구항 1

용어 가중치 재계산을 이용한 문서 군집 장치에 있어서,

용어의 가중치 재계산을 이용한 문서 군집의 전단계로서, 입력받은 문서 정보를 각각의 문장 정보로 분해한 후, 불용어를 제거하고, 어근을 추출하며, 용어-문장행렬을 생성함으로써, 비음수 행렬 인수분해(NMF : Non-Negative Matrix Factorization)를 기반으로 문서를 군집할 수 있게 하는 기능을 수행하는 전처리부(200); 및

상기 용어-문장행렬을 비음수 행렬 인수분해하며, 용어의 가중치를 재계산하고, 문서 군집을 생성하는 문서군집부(300); 를 포함하되,

상기 비음수 행렬 인수분해는 다음의 수학식에 의한 목적 함수가 최소값을 가지도록 수행되는 것을 특징으로 하는 용어 가중치 재계산을 이용한 문서 군집 장치.

$$J = \frac{1}{2}\left\| A - WH^T \right\|$$

(상기 수학식에서, J는 목적 함수이고, A는 용어-문장행렬로서 $A \approx WH^T$ 이며, W 는 비음수 의미특징 행렬(NSFM, non-negative semantic feature matrix)로서 $W=[w_{ij}]$이고, H^T는 비음수 의미변수 행렬(NSVM, non-negative semantic variable matrix)로서 $H=[h_{ij}]$이다.)

청구항 2

제 1 항에 있어서,

사용자로부터 용어의 가중치 재계산을 위한 문서 정보를 입력받고, 입력받은 문서를 분류하는 입력부(100); 및

상기 문서군집부(300)가 생성한 문서 군집 결과를 상기 사용자가 확인할 수 있도록 출력하는 출력부(400); 를 더 포함하는 것을 특징으로 하는 용어 가중치 재계산을 이용한 문서 군집 장치.

청구항 3

제 1 항에 있어서,

상기 전처리부(200)는,

입력받은 문서 정보를 각각의 문장 정보로 분해하는 문장 분해수단(210);

입력받은 문서 정보에 있어서 문서 군집의 개수를 설정하는 문서군집 설정수단(220);

분해된 각각의 문장에서 불용어를 제거하는 불용어 제거수단(230);

불용어가 제거된 각각의 문장에서 어근을 추출하는 어근 추출수단(240);

불용어가 제거된 각각의 문장에서 용어의 사용빈도에 따른 용어-빈도벡터를 생성하는 용어-빈도벡터 생성수단(250); 및

입력받은 문서 정보에서 용어-문장행렬을 생성하는 용어-문장행렬 생성수단(260); 을 포함하는 것을 특징으로 하는 용어 가중치 재계산을 이용한 문서 군집 장치.

청구항 4

제 1 항에 있어서,

상기 문서군집부(300)는,

상기 용어-문장행렬을 비음수 의미특징 행렬(NSFM, non-negative semantic feature matrix) 및 비음수 의미변

수 행렬(NSVM, non-negative semantic variable matrix)로 인수분해하는 비음수 행렬 인수분해 수단(310);

상기 비음수 의미특징 행렬 및 비음수 의미변수 행렬을 산출하는 비음수 행렬 산출수단(320);

상기 비음수 의미특징 행렬 및 비음수 의미변수 행렬을 이용하여 용어의 가중치를 재계산하는 가중치 재계산 수단(330); 및

재계산된 용어 가중치를 이용하여 문서를 군집하는 문서군집 생성수단(340); 을 포함하는 것을 특징으로 하는 용어 가중치 재계산을 이용한 문서 군집 장치.

청구항 5

제 1 항에 있어서,

상기 불용어 제거는 리츠베르젠(rijsbergen) 불용어 목록을 통해 수행되는 것을 특징으로 하는 용어 가중치 재계산을 이용한 문서 군집 장치.

청구항 6

삭제

청구항 7

제 1 항에 있어서,

상기 W 및 H^T의 원소 값을 갱신하기 위하여 상기 목적 함수값이 미리 설정된 수렴 허용 오차보다 작아지거나 미리 설정된 반복 횟수를 초과할 때까지 다음의 수학식에 의한 계산을 반복하는 것을 특징으로 하는 용어 가중치 재계산을 이용한 문서 군집 장치.

$$w_{ij} \leftarrow w_{ij} \frac{(AH)_{ij}}{(WH^T H)_{ij}}$$

$$h_{ij} \leftarrow h_{ij} \frac{(A^T H)_{ij}}{(HW^T W)_{ij}}$$

청구항 8

제 7 항에 있어서,

상기 수렴 허용 오차는 0.0001 내지 0.01인 것을 특징으로 하는 용어 가중치 재계산을 이용한 문서 군집 장치.

청구항 9

제 7 항에 있어서,

상기 반복 횟수는 10 내지 100인 것을 특징으로 하는 용어 가중치 재계산을 이용한 문서 군집 장치.

청구항 10

제 1 항에 있어서,

상기 비음수 행렬 인수분해의 결과, 비음수 의미특징 행렬 및 비음수 의미변수 행렬을 생성하고, 다음의 수학식을 이용하여 생성한 비음수 의미특징 행렬 및 비음수 의미변수 행렬을 정규화하는 것을 특징으로 하는 용어 가중치 재계산을 이용한 문서 군집 장치.

$$w_{ij} \leftarrow \frac{w_{ij}}{\sqrt{\sum_i w_{ij}^2}}$$

$$h_{ij} \leftarrow h_{ij}\sqrt{\sum_i w_{ij}^2}$$

(상기 수학식에서, w_{ij}는 비음수 의미특징 행렬의 원소를 나타내며, h_{ij}는 비음수 의미변수 행렬의 원소를 나타낸다.)

청구항 11

제 10 항에 있어서,

$x = \arg\max_j h_{ji}$ 이면 해당 문서를 군집 x에 할당하는 것을 특징으로 하는 용어 가중치 재계산을 이용한 문서 군집 장치.

청구항 12

제 1 항에 있어서,

상기 용어의 가중치 재계산은 다음의 수학식에 의하는 것을 특징으로 하는 용어 가중치 재계산을 이용한 문서 군집 장치.

$$H_{ki}^{new} = H_{ki}^{old} + \Delta H_{ki}$$

$$\Delta g_a = E(\Delta g_a^i) = \frac{1}{n}\sum_{i=1}^{n} \Delta g_a^i = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{A_{ai}} \sum_{k \in I_i} \Delta H_{ki} W_{ak}$$

$$g_a^{new} = g_a^{old} + \Delta g_a$$

(상기 수학식에서, Δg_a는 a행의 전체 원소에 대한 평균 가중치의 변화량이고, Δg_a^i는 i번째 문서의 a번째 용어 가중치의 변화량이며, n은 전체 문서의 개수이다. 또한, g_a는 a번째 용어의 가중치 값이며, A_{ai}는 a번째 용어와 i번째 문서의 용어의 빈도이고, W 및 H는 각각 비음수 행렬 인수분해로 생성된 비음수 의미특징 행렬 및 비음수 의미변수 행렬이며, I_i는 H의 i번째 문서 H_{*i}에서 $\Delta H_{ki}=0$이 아닌 k들의 집합이고, H_{ki}^{old}는 원래 값, H_{ki}^{new}는 학습 문서에 의해 수정된 값이다.)

청구항 13

제 12 항에 있어서,

상기 g_a^{old}의 초기 값은 1로 설정되는 것을 특징으로 하는 용어 가중치 재계산을 이용한 문서 군집 장치.

청구항 14

제 12 항에 있어서,

H_{ki}^{new}는 다음의 수학식에 의해 계산되는 것을 특징으로 하는 용어 가중치 재계산을 이용한 문서 군집 장치.

$$\Delta\overline{H}_{ji_k} = \begin{cases} if, b_j^c > \overline{H}_{ji_k^c}^{old}, \Delta\overline{H}_{ji_k^c} = +\zeta \cdot l_{ji_k^c} \\ if, b_j^c < \overline{H}_{ji_k^c}^{old}, \Delta\overline{H}_{ji_k^c} = -\zeta \cdot l_{ji_k^c} \end{cases}$$

$$\overline{H}_{ji_k^c}^{new} = \overline{H}_{ji_k^c}^{old} + \Delta\overline{H}_{ji_k^c}$$

$$H_{ji_k^c}^{new} = \overline{\overline{H}}_{ji_k^c}^{new} \times \sum_{p=1}^{r} H_{pi_k^c}^{old}$$

$$\overline{\overline{H}}_{ji_k^c}^{new} = \frac{\widetilde{H}_{ji_k^c}^{new}}{\sum_{p=1}^{r} \widetilde{H}_{pi_k^c}^{new}}$$

(상기 수학식에서, c는 군집의 일련번호로서 c=1,2,...,e이고, e는 군집의 개수를 나타내며, $l_{ji_k^c}$ 은 c번째 군집 내에서 의미특징의 중요도의 차이를 나타내고, ζ는 조절 상수이다.)

청구항 15

제 14 항에 있어서,

조절 상수 ζ는 0.5인 것을 특징으로 하는 용어 가중치 재계산을 이용한 문서 군집 장치.

청구항 16

제 1 항에 있어서,

상기 용어의 가중치 재계산은 다음의 수학식에 의하는 것을 특징으로 하는 용어 가중치 재계산을 이용한 문서 군집 장치.

$$\widetilde{A} = GA$$

(상기 수학식에서, A는 용어-문장행렬이고, $\widetilde{A}$ 는 재계산된 가중치가 적용된 용어-문장행렬이며, G는 재계산된 용어의 가중치 행렬이다.)

청구항 17

제 16 항에 있어서,

상기 $\widetilde{A}$ 행렬에 K-평균(K-means) 방법을 이용하여 상기 문서 군집의 생성을 수행하는 것을 특징으로 하는 용

어 가중치 재계산을 이용한 문서 군집 장치.

청구항 18

삭제

청구항 19

삭제

청구항 20

삭제

청구항 21

삭제

청구항 22

삭제

명 세 서

발명의 상세한 설명

기 술 분 야

<1> 본 발명은 용어 가중치 재계산을 이용한 문서 군집 장치에 관한 것으로서, 더욱 상세하게는 상위 수준의 사용자 요구 및 하위 수준의 시스템 군집 간의 의미적 차이를 최소화하기 위하여 사용자 피드백에 의해 가중치가 재계산된 용어를 이용하며, 문서 군집 과정에서 문서 집합의 내부 구조를 나타내는 의미 특징 행렬 및 의미 변수 행렬을 사용하여 문서 군집 방법의 성능을 향상시킬 수 있는 용어 가중치 재계산을 이용한 문서 군집 장치에 관한 것이다.

배 경 기 술

<2> 문서 군집은 군집 알고리즘에 의하여 문서 집합으로부터 유사한 특성을 가진 문서들의 그룹을 발견하는 것이다. 문서 군집 방법은 자료의 조직화, 웹 검색 결과의 브라우징 또는 다중 문서 요약 등 정보 검색의 많은 응용 분야에 사용되는 중요한 문서 분석 방법이다.

<3> 문서 군집에 있어서는 자료 집합의 분포, 내부 구조 또는 사용자가 원하는 군집 형태 등이 군집 결과에 상당한 영향을 미치게 되는데, 사용자에게 제한된 정보만을 제공하여 군집의 질을 향상시키는 연구들이 있어 왔다.

<4> 이러한 연구의 하나로서, 문서 군집 분석에 군집의 구성원에 대한 사전 지식(Prior Knowledge of Cluster Membership)을 통합한 준지도 문서 군집 모델(Semi-Supervised Document Clustering Model)이 제안된 바 있다.

<5> 상기 준지도 문서 군집 모델은 사용자가 그룹을 원하는 클러스터를 부호화하고, 부호화된 사용자의 사전 지식을 군집의 코스트 함수(Cost Function)에 벌점 용어(Penalty Term)로 부가한다.

<6> 다른 연구로서, 준지도 K-평균(K-means) 방법이 있는데, 이 방법은 분류 표시가 된 자료(Labeled Data)를 이용하여 초기 시드(Seed) 클러스터를 생성하고, 분류 표시가 된 자료로부터 제약 사항을 생성하여 군집한다.

<7> 또한, 비지도 학습 방법인 웹 검색 결과의 군집을 지도 학습 방법으로 변환하는 방법이 제안된 바 있다. 이 방법은 주어진 질의 및 검색 결과의 순위 리스트로부터 여러 개의 속성을 계산하고, 이 속성 및 학습 자료를 회귀 모델 학습에 적용하여 검색 결과에 대한 사용자의 브라우징(Browsing)을 목적으로 한다.

<8> 또 다른 연구로서, 군집의 이름과 관련이 없는 일반적인 특질로부터 군집에 필요한 문서의 특질을 분류하고, 제안 모델에 다양한 유형의 사용자 피드백을 적용하여 매개 변수를 조정할 수 있는 수단을 제공하는 방법이 있다.

<9> 그러나, 상술한 종래의 방법들은 사용자 피드백이 반영되지 않거나 부족하여 사용자가 원하는 군집 결과와 시스템에 의한 군집 결과의 의미적 차이를 유발하는 문제점이 있었다.

<10> 그리고, 상술한 종래의 방법들은 문서 집합의 내부 구조를 문서 군집에 반영하기가 곤란하여 문서 군집의 정확도를 담보할 수 없는 문제점이 있었다.

발명의 내용

해결 하고자하는 과제

<11> 본 발명이 해결하고자 하는 과제는, 사용자가 요구하는 군집과 시스템이 제공하는 군집 간의 의미적 차이를 최소화하는 군집 장치 및 방법을 제공하는 것이다.

<12> 본 발명이 해결하고자 하는 다른 과제는, 군집의 내부 구조 및 의미 특징의 분포를 용이하게 파악할 수 있는 군집 장치 및 방법을 제공하는 것이다.

<13> 본 발명이 해결하고자 하는 또 다른 과제는, 용어의 가중치를 재계산함으로써 문서 군집의 정확도를 향상시키는 군집 장치 및 방법을 제공하는 것이다.

<14> 본 발명이 해결하고자 하는 또 다른 과제는, 종래의 문서 군집 방법에 적용하여 군집 방법의 성능을 향상시킬 수 있는 용어 가중치 추정 방법을 제공하는 것이다.

과제 해결수단

<15> 본 발명은, 용어 가중치 재계산을 이용한 문서 군집 장치에 관한 것으로서, 용어의 가중치 재계산을 이용한 문서 군집의 전단계로서, 입력받은 문서 정보를 각각의 문장 정보로 분해한 후, 불용어를 제거하고, 어근을 추출하며, 용어-문장행렬을 생성함으로써, 비음수 행렬 인수분해(NMF : Non-Negative Matrix Factorization)를 기반으로 문서를 군집할 수 있게 하는 기능을 수행하는 전처리부; 및 상기 용어-문장행렬을 비음수 행렬 인수분해하며, 용어의 가중치를 재계산하고, 문서 군집을 생성하는 문서군집부; 를 포함하되, 상기 비음수 행렬 인수분해는 다음의 수학식에 의한 목적 함수가 최소값을 가지도록 수행되는 것을 특징으로 한다.

$$J = \frac{1}{2}\left\| A - WH^T \right\|$$

(상기 수학식에서, J는 목적 함수이고, A는 용어-문장행렬로서 $A \approx WH^T$ 이며, W 는 비음수 의미특징 행렬로서 $W=[w_{ij}]$이고, H^T는 비음수 의미변수 행렬로서 $H=[h_{ij}]$이다.)

<16> 바람직하게는, 사용자로부터 용어의 가중치 재계산을 위한 문서 정보를 입력받고, 입력받은 문서를 분류하는 입력부; 및 상기 문서군집부가 생성한 문서 군집 결과를 상기 사용자가 확인할 수 있도록 출력하는 출력부; 를 더 포함하는 것을 특징으로 한다.

<17> 또한 바람직하게는, 상기 전처리부는, 입력받은 문서 정보를 각각의 문장 정보로 분해하는 문장 분해수단; 입력받은 문서 정보에 있어서 문서 군집의 개수를 설정하는 문서군집 설정수단; 분해된 각각의 문장에서 불용어를 제거하는 불용어 제거수단; 불용어가 제거된 각각의 문장에서 어근을 추출하는 어근 추출수단; 불용어가 제거된 각각의 문장에서 용어의 사용빈도에 따른 용어-빈도벡터를 생성하는 용어-빈도벡터 생성수단; 및 입력받은 문서 정보에서 용어-문장행렬을 생성하는 용어-문장행렬 생성수단; 을 포함하는 것을 특징으로 한다.

<18> 또한 바람직하게는, 상기 문서군집부는, 상기 용어-문장행렬을 비음수 의미특징 행렬(NSFM, non-negative semantic feature matrix) 및 비음수 의미변수 행렬(NSVM, non-negative semantic variable matrix)로 인수분해하는 비음수 행렬 인수분해 수단; 상기 비음수 의미특징 행렬 및 비음수 의미변수 행렬을 산출하는 비음수 행렬 산출수단; 상기 비음수 의미특징 행렬 및 비음수 의미변수 행렬을 이용하여 용어의 가중치를 재계산하는 가중치 재계산 수단; 및 재계산된 용어 가중치를 이용하여 문서를 군집하는 문서군집 생성수단; 을 포함하는 것을 특징으로 한다.

<19> 또한 바람직하게는, 상기 불용어 제거는 리츠베르젠(rijsbergen) 불용어 목록을 통해 수행되는 것을 특징으로 한다.

<20> 삭제

<21> 삭제

<22> 삭제

<23> 또한 바람직하게는, 상기 W 및 H^T의 원소 값을 갱신하기 위하여 상기 목적 함수값이 미리 설정된 수렴 허용 오차보다 작아지거나 미리 설정된 반복 횟수를 초과할 때까지 다음의 수학식에 의한 계산을 반복하는 것을 특징으로 한다.

$$w_{ij} \leftarrow w_{ij} \frac{(AH)_{ij}}{(WH^T H)_{ij}}$$

$$h_{ij} \leftarrow h_{ij} \frac{(A^T H)_{ij}}{(HW^T W)_{ij}}$$

<24>

<25> 또한 바람직하게는, 상기 수렴 허용 오차는 0.0001 내지 0.01인 것을 특징으로 한다.

<26> 또한 바람직하게는, 상기 반복 횟수는 10 내지 100인 것을 특징으로 한다.

<27> 또한 바람직하게는, 상기 비음수 행렬 인수분해의 결과, 비음수 의미특징 행렬 및 비음수 의미변수 행렬을 생성하고, 다음의 수학식을 이용하여 생성한 비음수 의미특징 행렬 및 비음수 의미변수 행렬을 정규화하는 것을 특징으로 한다.

$$w_{ij} \leftarrow \frac{w_{ij}}{\sqrt{\sum_i w_{ij}^2}}$$

$$h_{ij} \leftarrow h_{ij} \sqrt{\sum_i w_{ij}^2}$$

<28>

<29> (상기 수학식에서, w_{ij}는 비음수 의미특징 행렬의 원소를 나타내며, h_{ij}는 비음수 의미변수 행렬의 원소를 나타낸다.)

$$x = \arg\max_j h_{ji}$$

<30> 또한 바람직하게는, 이면 문서 d_i를 군집 x에 할당하는 것을 특징으로 한다.

<31> 또한 바람직하게는, 상기 용어의 가중치 재계산은 다음의 수학식에 의하는 것을 특징으로 한다.

$$H_{ki}^{new} = H_{ki}^{old} + \Delta H_{ki}$$

<32>

$$\Delta g_a = E(\Delta g_a^i) = \frac{1}{n}\sum_{i=1}^{n} \Delta g_a^i = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{A_{ai}} \sum_{k \in I_i} \Delta H_{ki} W_{ak}$$

<33>

<34> $$g_a^{new} = g_a^{old} + \Delta g_a$$

<35> (상기 수학식에서, Δg_a는 a행의 전체 원소에 대한 평균 가중치의 변화량이고, Δg_a^i는 i번째 문서의 a번째 용어 가중치의 변화량이며, n은 전체 문서의 개수이다. 또한, g_a는 a번째 용어의 가중치 값이며, A_{ai}는 a번째 용어와 i번째 문서의 용어의 빈도이고, W 및 H는 각각 비음수 행렬 인수분해로 생성된 비음수 의미특징 행렬 및 비음수 의미변수 행렬이며, I_i는 H의 i번째 문서 H_{*i}에서 $\Delta H_{ki}=0$이 아닌 k들의 집합이고, H_{ki}^{old}는 원래 값, H_{ki}^{new}는 학습 문서에 의해 수정된 값이다.)

<36> 또한 바람직하게는, 상기 g_a^{old}의 초기 값은 1로 설정되는 것을 특징으로 한다.

<37> 또한 바람직하게는, H_{ki}^{new}는 다음의 수학식에 의해 계산되는 것을 특징으로 한다.

$$\Delta \overline{H}_{ji_k} = \begin{cases} if, b_j^c > \overline{H}_{ji_k^c}^{old}, \Delta \overline{H}_{ji_k^c} = +\zeta \cdot l_{ji_k^c} \\ if, b_j^c < \overline{H}_{ji_k^c}^{old}, \Delta \overline{H}_{ji_k^c} = -\zeta \cdot l_{ji_k^c} \end{cases}$$

$$\overline{H}_{ji_k^c}^{new} = \overline{H}_{ji_k^c}^{old} + \Delta \overline{H}_{ji_k^c}$$

<38>

$$H_{ji_k^c}^{new} = \approx\!\!\approx_{ji_k^c}^{new} \times \sum_{p=1}^{r} H_{pi_k^c}^{old}$$

$$\approx\!\!\approx_{ji_k^c}^{new} = \frac{\widetilde{H}_{ji_k^c}^{new}}{\sum_{p=1}^{r} \widetilde{H}_{pi_k^c}^{new}}$$

<39>

<40> (상기 수학식에서, c는 군집의 일련번호로서 c=1,2,...,e이고, e는 군집의 개수를 나타내며, $l_{ji_k^c}$은 c번째 군집 내에서 의미특징의 중요도의 차이를 나타내고, ζ는 조절 상수이다.)

<41> 또한 바람직하게는, 조절 상수 ζ는 0.5인 것을 특징으로 한다.

<42> 또한 바람직하게는, 상기 용어의 가중치 재계산은 다음의 수학식에 의하는 것을 특징으로 한다.

<43> $$\widetilde{A} = GA$$

<44> (상기 수학식에서, A는 용어-문장행렬이고, $\widetilde{A}$는 재계산된 가중치가 적용된 용어-문장행렬이며, G는 재계산된 용어의 가중치 행렬이다.)

<45> 그리고 바람직하게는, 상기 $\tilde{A}$ 행렬에 K-평균(K-means) 방법을 이용하여 상기 문서 군집의 생성을 수행하는 것을 특징으로 한다.

<46> 삭제

<47> 삭제

<48> 삭제

<49> 삭제

<50> 삭제

효 과

<51> 본 발명에 따르면, 사용자에 의해서 미리 분류된 문서를 이용하여 사용자가 원하는 군집과 시스템에 의한 군집 사이의 의미적 차이를 최소화할 수 있는 효과가 있다.

<52> 본 발명에 따르면, 의미 특징과 의미 변수를 사용하여 군집의 내부 구조 및 의미 특징의 분포를 쉽게 파악할 수 있는 효과가 있다.

<53> 본 발명에 따르면, 의미 특징과 의미 변수를 이용하여 용어의 가중치를 용이하게 재계산함으로써 문서 군집의 정확도를 향상시킬 수 있는 효과가 있다.

<54> 본 발명에 따르면, 가중치 추정은 비음수 행렬 인수분해에 기반한 문서 군집 방법뿐만 아니라 단수값 분해(SVD : Singular Value Decomposition)도 적용하여 계산할 수 있어서 적용 범위가 확장되는 효과가 있다.

<55> 본 발명에 따르면, 종래의 문서 군집 방법에도 추정된 용어의 가중치를 적용할 수 있으므로 종래의 문서 군집 방법의 성능을 향상시키는 효과도 있다.

발명의 실시를 위한 구체적인 내용

<56> 본 발명의 실시를 위한 구체적인 내용을 설명하기에 앞서, 본 발명의 기술적 요지와 직접적 관련이 없는 구성에 대하여는 본 발명의 기술적 요지를 흐뜨리지 않는 범위 내에서 생략하였음을 유의하여야 할 것이다.

<57> 또한, 본 명세서 및 청구범위에 사용된 용어나 단어는 발명자가 그 자신의 발명을 가장 최선의 방법으로 설명하기 위해 용어의 개념을 적절하게 정의할 수 있다는 원칙에 입각하여 본 발명의 기술적 사상에 부합하는 의미와 개념으로 해석되어야 할 것이다.

<58> 본 발명은 지도 학습 방법을 이용하여 비음수 행렬 인수분해(NMF : Non-Negative Matrix Factorization)에 기반한 용어의 가중치를 추정하고, 추정된 용어의 가중치를 군집 대상 집합에 적용하여 문서를 군집하는 장치 및 방법을 제안한다.

<59> 상기 비음수 행렬 인수분해는 인간이 객체를 인식할 때 객체의 부분 정보의 조합으로 인식하는 것에 착안하여, 객체 정보를 기초 특징(Base Feature) 및 부호 특징(Encoding Feature)으로 분해하여 부분 정보(Part-Base)로 표현하는 알고리즘이다.

<60> 이러한 부분 정보의 조합으로 전체 객체를 표현하는 방법은 대량의 정보를 효율적으로 표현할 수 있으며, 본 발명의 바람직한 실시예에 있어서는 주어진 양의 행렬로부터 양의 인수를 찾아낸다.

<61> 이하, 본 발명의 바람직한 실시예에 따른 용어 가중치 재계산을 이용한 문서 군집 장치에 관하여 도 1 을 참조하여 설명한다.

<62> 도 1 은 본 발명의 바람직한 실시예에 따른 용어 가중치 재계산을 이용한 문서 군집 장치에 관한 전체 구성도이다.

<63> 상기 도 1 에 도시된 바와 같이, 본 발명의 바람직한 실시예에 따른 용어 가중치 재계산을 이용한 문서 군집 장치는 입력부(100), 전처리부(200), 문서군집부(300) 및 출력부(400)를 포함한다.

<64> 상기 입력부(100)는 사용자 측으로부터 용어의 가중치 재계산을 위한 문서 정보를 입력받고, 입력받은 문서를 군집 레이블(Label)에 일치하도록 분류한다.

<65> 또한, 상기 전처리부(200)는 용어의 가중치 재계산을 통한 문서 군집의 전단계로서, 입력받은 문서 정보를 각각의 문장 정보로 분해한 후, 불필요한 문자열을 제거하고, 어근을 추출하며, 비음수 행렬 인수분해를 기반으로 문서를 군집할 수 있게 하는 기능을 수행하는 바, 문장 분해수단(210), 문서군집 설정수단(220), 불용어(Stop Word) 제거수단(230), 어근 추출수단(240), 용어(Term)-빈도(Frequency)벡터 생성수단(250), 용어-문장행렬 생성수단(260)을 포함한다.

<66> 먼저, 상기 문장 분해수단(210)은 입력받은 문서 정보를 각각의 문장 정보로 분해하는 기능을 수행한다.

<67> 다음으로, 상기 문서군집 설정수단(220)은 입력받은 문서 정보에 있어서 문서 군집의 개수를 설정하는 기능을 수행하는 바, 본 실시예에서 문서 군집의 개수는 k로 설정하였으나, 본 발명이 이에 한정되지 아니한다.

<68> 다음으로, 상기 불용어 제거수단(230)은 상기 문장 분해수단(210)을 통해서 분해된 각각의 문장들에 대하여 불용어 목록(예컨대, 리츠베르젠-rijsbergen 등)을 통해 불용어를 제거하는 기능을 수행한다.

<69> 참고적으로, 불용어(Stop-words 또는 Noise Word)는 검색시에 무시해버리는 단어 또는 검색엔진이 데이터베이스를 구축할 때, 색인에서 제외해 버리는 단어를 뜻하며, 예를 들어 한글에서 '~를, ~을, ~에서, ~와' 등을 포함하는 조사, 접미사, 접속사, 어미 등이 있고, 그리고 영어에서 'a, an, is, are, if, for, but, this, may' 등을 포함하는 동사, 조동사, 인칭대명사, 지시대명사, 전치사 등이 있다.

<70> 다음으로, 상기 어근 추출수단(240)은 각각의 문장에 대하여 어근을 추출하는 기능을 수행한다.

<71> 참고적으로, 어근은 단어를 이루는 형태소 중에서 그 단어의 실질적인 뜻을 나타내는 형태소로서, 합성어의 경우에 합성된 낱낱의 실질 형태소이고, 파생어의 경우에 접사를 제외한 나머지 부분이 해당한다. 예를 들면, '덮개'에서 '덮', '사람답다'에서 '사람', '넉넉하다'에서 '넉넉' 등이 어근에 해당한다.

<72> 다음으로, 상기 용어-빈도벡터 생성수단(250)은 각각의 문장에 대하여 불용어를 제외한 용어의 사용빈도에 따른 용어-빈도벡터를 생성하는 기능을 수행한다.

<73> 마지막으로, 상기 용어-문장행렬 생성수단(260)은 문서에서 총 m개의 용어와 n개의 문장으로 이루어진 m×n으로 구성된 행렬 A를 생성하는 기능을 수행한다.

<74> 이때, 행렬 A는 $A=[A_{*1},A_{*2},...,A_{*n}]$로 나타낼 수 있으며, 각 행 벡터 A_{*i}는 i번째 문서의 용어-빈도벡터이다.

<75> 또한, 상기 문서군집부(300)는 상기 용어-문장행렬 생성수단(260)을 통해서 생성된 행렬 A를 비음수 행렬 인수분해함으로써, 용어의 가중치를 재계산하고, 문서 군집을 생성하는 기능을 수행하는 바, 비음수 행렬 인수분해 수단(310), 비음수 행렬 산출수단(320), 가중치 재계산 수단(330) 및 문서군집 생성수단(340)을 포함한다.

<76> 상기 비음수 행렬 인수분해 수단(310)은 행렬 A를 비음수 의미특징 행렬(NSFM, non-negative semantic feature matrix)과 비음수 의미변수 행렬(NSVM, non-negative semantic variable matrix)을 이용하여 인수분해하는 기능을 수행한다.

<77> 이하, 상기 비음수 행렬 인수분해(NMF : Non-Negative Matrix Factorization)를 이용한 용어-문장행렬의 분해에 관하여 설명한다.

<78> 문서 집합이 k개의 군집으로 구성된다고 가정할 때, 용어-문장행렬 A를 다음의 [수학식 2]가 나타내는 목적 함수가 최소값을 가지도록 다음의 [수학식 1]과 같이 m×k 비음수 의미특징 행렬 W 및 k×n 비음수 의미변수 행렬 H^T로 분해한다. 본 실시예에서 행렬 A의 i번째 행과 j번째 열의 원소는 A_{ij}로 표시하기로 한다.

수학식 1

<79> $$A \approx WH^T$$

수학식 2

<80> $$J = \frac{1}{2}\left\|A - WH^T\right\|$$

<81> 상기 [수학식 1] 및 [수학식 2]에서, W=[w_{ij}]이고, H=[h_{ij}]이며, W=[w_1,w_2,...,w_k]이다.

<82> 이때, W와 H^T의 원소 값을 갱신하기 위하여 상기 [수학식 2]에 의한 목적함수 J값이 미리 설정한 수렴 허용 오차보다 작아지거나 지정한 반복 횟수를 초과할 때까지 다음의 [수학식 3]을 반복한다.

<83> 미리 설정한 수렴 허용 오차는 0.0001 내지 0.01인 것으로 설정할 수 있으며, 바람직하게는 0.001이다. 또한, 지정한 반복 횟수는 10 내지 100인 것으로 설정할 수 있으며, 바람직하게는 50이다.

수학식 3

<84> $$w_{ij} \leftarrow w_{ij} \frac{(AH)_{ij}}{(WH^T H)_{ij}}$$

$$h_{ij} \leftarrow h_{ij} \frac{(A^T H)_{ij}}{(HW^T W)_{ij}}$$

<85> 주어진 문서 집합으로부터 구성된 용어-문장행렬 A에 상기 [수학식 3]에 의한 비음수 행렬 인수분해를 수행하여 비음수 행렬 W와 H를 얻는다.

<86> 그리고 다음의 [수학식 4]를 이용하여 행렬 W와 H를 정규화하며, 행렬 H를 이용하여 각 데이터 포인트(Data Point)의 군집 레이블을 결정한다. 예를 들어, $x = \arg\max_j h_{ji}$ 이면 문서 d_i를 군집 x에 할당한다.

수학식 4

<87> $$w_{ij} \leftarrow \frac{w_{ij}}{\sqrt{\sum_i w_{ij}^2}}$$

$$h_{ij} \leftarrow h_{ij}\sqrt{\sum_i w_{ij}^2}$$

<88> 다음은 상기 [수학식 3]을 이용하여 비음수 행렬 인수분해를 적용한 결과, A 행렬이 W 및 H^T 행렬로 분해된 예이다. 다음의 예에서, k=2이고, 수렴할 반복 횟수는 50이며, 수렴 허용 오차는 0.001로 설정하였고, W 및 H^T 행렬의 초기값은 각각 0.5이었다.

$$\underset{A}{\begin{bmatrix}1 & 2 & 3\\ 4 & 5 & 6\\ 7 & 8 & 9\\ 10 & 11 & 12\end{bmatrix}} \approx \underset{W}{\begin{bmatrix}0.15 & 1.60\\ 0.66 & 0.97\\ 1.15 & 0.57\\ 1.61 & 0.41\end{bmatrix}} \times \underset{H^T}{\begin{bmatrix}6.11 & 6.68 & 7.18\\ 0.09 & 0.60 & 1.22\end{bmatrix}} = \underset{\tilde{A}}{\begin{bmatrix}1.05 & 1.94 & 3.03\\ 4.12 & 4.98 & 5.93\\ 7.07 & 8.01 & 8.95\\ 9.90 & 11.01 & 12.08\end{bmatrix}}$$

<89>

<90> 비음수 행렬 인수분해의 수행 결과, 행렬 A의 j번째 열벡터 A_{*j}는 행렬 W의 l번째 열벡터 W_{*l}와 행렬 H^T의 요소 H_{kj}^T가 선형 조합을 이루며 다음의 [수학식 5]와 같다.

수학식 5

$$A_{*j} = \sum_{l=1}^{k} H_{kj}^{T} W_{*l}$$

<91>

<92> 다음으로, 상기 비음수 행렬 산출수단(320)은 비음수 의미특징 행렬 W 및 비음수 의미변수 행렬 H를 산출하는 기능을 수행한다.

<93> 상기 비음수 행렬 인수분해 수단(310)에 의하여 행렬 A로부터 분해된 비음수 의미특징 행렬을 W라 하고, 비음수 의미변수 행렬을 H라 할 때, 행렬 W의 의미특징 벡터 W_{*l}는 문장의 내부 특징을 나타내며, 행렬 H의 원소인 의미변수 H_{kj}는 문장 내에서의 의미특징의 중요도를 나타낸다.

<94> 다음으로, 상기 가중치 재계산 수단(330)은 군집 대상이 되는 문서 집합에 적용하기 위하여, 상기 비음수 행렬 인수분해의 결과를 이용하여 용어의 가중치를 재계산하는 기능을 수행한다.

<95> 이하, 상기 가중치 재계산 수단(330)의 용어 가중치 재계산에 관하여 상세히 설명한다.

<96> 상기 용어 가중치 재계산은 다음의 [수학식 9] 및 [수학식 10]에 의하여 이루어질 수 있다. 상기 [수학식 9]는 다음의 [수학식 6] 내지 [수학식 8]에 의하여 유도된다.

수학식 6

$$\tilde{A}_{ai} = g_a A_{ai}$$

<97>

수학식 7

$$\Delta g_a^{\,i} A_{ai} = \sum_{k \in I_i} \Delta H_{ki} W_{ak}$$

<98>

수학식 8

$$H_{ki}^{\,new} = H_{ki}^{\,old} + \Delta H_{ki}$$

<99>

<100> 상기 [수학식 6] 내지 [수학식 8]에서, g_a는 a번째 용어의 가중치 값, A_{ai}는 a번째 용어와 i번째 문서의 용어의 빈도, I_i는 H의 i번째 문서 H_{*i}에서 $\Delta H_{ki}=0$이 아닌 k들의 집합, H_{ki}^{old}는 원래 값, H_{ki}^{new}는 학습 문서에 의해 수정

된 값이다.

수학식 9

$$\Delta g_a = E(\Delta g_a^{\,i}) = \frac{1}{n}\sum_{i=1}^{n}\Delta g_a^{\,i} = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{A_{ai}}\sum_{k\in I_i}\Delta H_k W_{ak}$$

<101>

<102> 상기 [수학식 9]에서, Δg_a는 a행의 전체 원소에 대한 평균 가중치의 변화량이고, $\Delta g_a^{\,i}$는 i번째 문서의 a번째 용어 가중치의 변화량, n은 전체 문서의 개수이다.

수학식 10

$$g_a^{\,new} = g_a^{\,old} + \Delta g_a$$

<103>

<104> 상기 [수학식 10]에서, $g_a^{\,old}$의 초기 값은 1로 설정된다.

<105> 그리고, $H_{ki}^{\,new}$는 다음의 [수학식 11] 내지 [수학식 15]와 같이 구할 수 있다.

수학식 11

$$\widetilde{H}_{ji}^{\,old} = \frac{H_{ji}^{\,old}}{\sum_{k=1}^{r} H_{ki}^{\,old}}$$

<106>

<107> 상기 [수학식 11]에서, $\widetilde{H}_{ji}^{\,old}$는 정규화된 $H_{ji}^{\,old}$이다.

수학식 12

$$b_j^{\,c} = \frac{1}{f^c}\sum_{k=1}^{f^c}\widetilde{H}_{ji_k}^{\,old}$$

$$\sum_{j=1}^{r} b_j^{\,c} = 1$$

<108>

<109> 상기 [수학식 12]에서, c는 군집의 일련번호로서 c=1,2,...,e이고, e는 군집의 개수를 나타낸다. 예컨대, D_c를 c번째 군집이라 하면, c번째 군집에 포함된 문서의 개수는 $|D_c| = f^c$이며 $d_{i_k} \in D_c$로서 k=1,2,...,f^c이다.

수학식 13

$$l_{ji_k^c} = \left| b_j^c - \widetilde{H}_{ji_k^c}^{old} \right|$$

<110>

<111> 상기 [수학식 13]에서, $l_{ji_k^c}$ 은 c번째 군집 내에서 의미특징의 중요도의 차이를 나타낸다.

수학식 14

$$\Delta \overline{H}_{ji_k} = \begin{cases} if, b_j^c > \overline{H}_{ji_k^c}^{old}, \Delta \overline{H}_{ji_k^c} = +\zeta \cdot l_{ji_k^c} \\ if, b_j^c < \overline{H}_{ji_k^c}^{old}, \Delta \overline{H}_{ji_k^c} = -\zeta \cdot l_{ji_k^c} \end{cases}$$

$$\overline{H}_{ji_k^c}^{new} = \overline{H}_{ji_k^c}^{old} + \Delta \overline{H}_{ji_k^c}$$

<112>

<113> 상기 [수학식 14]에서, ξ는 조절 상수로서 0.5를 갖는 것이 바람직하다.

수학식 15

$$H_{ji_k^c}^{new} = \approx\!\approx \widetilde{H}_{ji_k^c}^{new} \times \sum_{p=1}^{r} H_{pi_k^c}^{old}$$

$$\approx\!\approx \widetilde{H}_{ji_k^c}^{new} = \frac{\widetilde{H}_{ji_k^c}^{new}}{\sum_{p=1}^{r} \widetilde{H}_{pi_k^c}^{new}}$$

<114>

<115> 다음의 [수학식 16]은 상기 [수학식 9]에 의하여 재계산된 용어의 가중치 행렬이다.

수학식 16

$$G = \begin{pmatrix} g_1 & 0 & \cdots & 0 \\ 0 & g_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & g_n \end{pmatrix}$$

<116>

<117> 상기 [수학식 16]에서, 가중치 행렬 G의 원소값은 g_a^{new}의 a번째 용어와 일치하는 A_{a*}의 용어가 존재하는 경우

g_a^{new}의 원소값을 가지며, 그렇지 않으면 1의 값을 가진다.

<118> 마지막으로, 상기 문서군집 생성수단(340)은 계산된 용어 가중치를 문서 집합에 적용함으로써 문서를 군집하는 기능을 수행한다.

<119> 행렬 A에 대한 추정 가중치는 상기 [수학식 16]을 이용하여 다음의 [수학식 17]과 같은 행렬 $\tilde{A}$로 적용한다.

수학식 17

<120> $$\tilde{A} = GA$$

<121> 그 다음, 재계산된 가중치가 적용된 행렬 $\tilde{A}$에 비음수 행렬 인수분해를 이용한 방법 또는 K-평균(K-means) 방법을 이용하여 군집한다.

<122> 그리고, 상기 출력부(400)는 상기 문서군집부(300)를 통해서 생성된 문서 군집을 사용자가 확인할 수 있도록 출력하는 기능을 수행하는 바, 본 발명의 바람직한 실시예에서 모니터 또는 프린터 등을 포함하는 것으로 설정하였으나, 본 발명이 이에 한정되지 아니한다.

<123> 이하, 본 발명의 바람직한 실시예에 따른 용어 가중치 재계산을 이용한 문서 군집 방법에 관하여 도 2 를 참조하여 설명한다.

<124> 도 2 는 본 발명의 바람직한 실시예에 따른 용어 가중치 재계산을 이용한 문서 군집 방법에 관한 전체 흐름도이다.

<125> 상기 도 2 에 도시된 바와 같이, 먼저 군집 대상이 되는 문서 집합으로부터 사용자 피드백에 의하여 문서를 추출하고 군집을 결정한다(S10).

<126> 다시 말하면, 사용자에 의하여 문서 집합으로부터 용어의 가중치 재계산을 위한 문서를 입력받고, 군집 레이블(Label)에 일치하도록 문서를 분류한다.

<127> 본 실시예에서 의미하는 사용자 피드백은 일반적인 질의를 확장하여 용어를 재작성하는 연관 피드백과는 달리, 정확한 군집 결과가 도출될 수 있도록 용어에 대한 가중치를 재계산하기 위하여 사용자가 원하는 결과의 군집에 포함되는 문서를 추출하는 피드백을 의미한다.

<128> 다음으로, 추출된 문서 집합을 전처리한다(S20).

<129> 상기 S20 단계는 상기 S10 단계에서 추출된 문서에 대한 불용어 제거 및 어근 추출 등으로 수행된다.

<130> 다음으로, 전처리한 문서로부터 벡터 모델을 생성함으로써 용어-문장행렬 A를 생성한다(S30).

<131> 상기 S20 단계의 전처리 후, 총 m개의 용어와 n개의 문장으로 이루어진 m×n 행렬 A는 $A=[A_{*1},A_{*2},\ldots,A_{*n}]$로 나타낼 수 있으며, 각 행 벡터 A_{*i}는 i번째 문서의 용어-빈도벡터이다.

<132> 다음으로, 비음수 행렬 인수분해를 이용하여 용어-문장행렬 A를 비음수 의미특징 행렬 및 비음수 의미변수 행렬로 분해한다(S40).

<133> 상기 S40 단계에 의하여, 추출된 문서 집합 내의 문장 벡터들은 의미특징 벡터에 가중치인 의미변수를 곱한 값의 선형 합으로 표시된다. 의미특징 벡터는 문장의 내부 특징을 나타내며, 의미변수는 문장 내에서 의미특징의 중요도를 나타낸다.

<134> 다음으로, 비음수 의미변수 행렬에 가중치 재계산 방법을 적용하여 용어의 가중치를 재계산한다(S50).

<135> 본 발명은 비음수 행렬 인수분해에 기반한 용어의 가중치를 지도 학습 방법으로 추정하고, 추정된 용어의 가중치를 군집 대상 집합에 적용하여 문서를 군집하는 문서 군집 방법을 제안하는 바, 본 발명의 바람직한 실시예에 따른 가중치 재계산 방법은 상기 단수값 분해 또는 비음수 행렬 인수분해 등과 같이 원본 데이터를 분해하여 군집하는 모든 방법에 적용할 수 있다.

<136> 마지막으로, 계산된 용어 가중치를 문서 집합 전체에 적용한 후 문서 군집 방법을 이용하여 문서를 군집한다(S60).

<137> 마지막으로, 상기 S60 단계의 군집 결과를 사용자에게 제공하여 사용자가 군집 결과에 만족하는 경우 절차를 종료한다(S70).

<138> 상기 S70 단계의 판단 결과, 사용자가 군집 결과에 만족하지 않는 경우 상기 S50 단계로 절차를 이행한다.

<139> 이하, 본 발명의 바람직한 실시예에 따른 용어 가중치 재계산을 이용한 문서 군집 방법에 관한 실험 결과에 대하여 설명한다.

<140> 상기 실험은 20,000개의 문서를 포함하는 20개의 뉴스 그룹 중 일부를 무작위로 추출한 자료를 사용하여 수행되었다.

<141> 뉴스 그룹은 컴퓨터 그래픽, 운영체제 윈도우, 컴퓨터 하드웨어, 종교, 의학, 정치 등 20개의 다양한 주제로 구성되며, 각 주제에 포함된 기사의 수는 서로 같다.

<142> 실험에 사용된 평가 자료에 대하여 다음의 [표 1]에 나타내었다.

표 1

<143>

문서 집합의 속성	20 뉴스 그룹
총 문서 개수	20,000
사용 문서 개수	5,400
클러스터 개수	20
사용 클러스터 개수	10
최대 클러스터의 문서 개수	1,000
최소 클러스터의 문서 개수	100
중간 클러스터의 문서 개수	500
평균 클러스터의 문서 개수	540

<144> 성능 평가의 척도로는 다음의 [수학식 18]에 의한 NMI(Normalize Mutual Information)를 사용하였다.

<145> 두 개의 문서 군집 C 및 C'이 주어질 때 이들 간의 상호정보 MI(C,C')는 다음의 [수학식 18]과 같이 정의된다.

수학식 18

$$MI(C,C') = \sum_{c_i \in C, c_j' \in C'} p(c_i, c_j') \cdot \log_2 \frac{p(c_i, c_j')}{p(c_i) \cdot p(c_j')}$$

$$NMI(C,C') = \frac{MI(C,C')}{\max(H(C), H(C'))}$$

<146>

<147> 상기 [수학식 18]에서, $p(c_i)$ 및 $p(c'_j)$는 각각 군집 c_i 및 c'_j에 문서 집합의 문서가 포함될 확률이고, $p(c_i, c'_j)$는 문서 집합의 문서가 동시에 군집 c_i 및 c'_j에 포함될 확률이다. H(C) 및 H(C')는 각각 C 및 C'의 엔트로피이다.

<148> 다음의 [표 2]와 같이 서로 다른 두 가지 군집 방법의 NMI를 군집의 개수를 2에서 10까지 증가시키면서 비교하였다.

표 2

<149>

k	K-means	가중치 K-means	비음수 행렬 인수분해	가중치 비음수 행렬 인수분해
2	0.6545	0.6648	0.6874	0.7235
3	0.6898	0.6987	0.6998	0.7432
4	0.6211	0.6541	0.7021	0.7534
5	0.6455	0.6675	0.7212	0.7512
6	0.6635	0.6784	0.6985	0.7224
7	0.6422	0.6544	0.7201	0.7403
8	0.6665	0.6855	0.7032	0.7644
9	0.6633	0.6913	0.6991	0.7411
10	0.6420	0.6702	0.7045	0.7321
평균	0.6543	0.6739	0.7034	0.7413

<150> 상기 [표 2]에서, K-means 방법은 표준 K-means 방법을 의미하고, 비음수 행렬 인수분해(NMF)는 쉬(Xu)가 제안한 군집 방법이다. 또한 가중치 K-means 방법 및 가중치 비음수 행렬 인수분해는 K-means 및 비음수 행렬 인수분해 방법에 가중치를 재계산하여 적용한 방법이다.

<151> 평가는 두 가지 군집 방법에 재계산된 가중치를 적용한 것과 적용하지 않고 군집한 결과에 [수학식 18]을 이용하여 수행되었다.

<152> 상기 [표 2]에 나타낸 바와 같이, 가중치를 재계산한 방법이 가중치를 계산하지 않은 방법에 비해서 더 좋은 성능을 나타냄을 알 수 있다. 가중치가 부여된 K-means 방법은 가중치가 부여되지 않은 K-means 방법에 비하여 평균 NMI가 2.9% 더 높으며, 가중치가 부여된 비음수 행렬 인수분해 군집 방법은 가중치가 부여되지 않은 비음수 행렬 인수분해 군집 방법에 비하여 평균 NMI가 5.2% 더 높다.

<153> 가중치를 재계산하지 않은 K-means 방법이 최하의 성능을 나타낸다.

<154> 비음수 행렬 인수분해 군집 방법이 K-means 방법보다 성능이 좋은 이유는 K-means의 단순한 유사도를 이용한 군집보다 비음수 행렬 인수분해를 이용하여 자료의 내부 구조를 반영하여 군집하는 것이 더욱 정확도를 기할 수 있기 때문이다. 또한, 가중치가 부여된 비음수 행렬 인수분해 군집 방법은 자료의 내부 구조를 반영하면서, 사용자가 부여한 학습을 통하여 용어의 가중치를 재계산함으로써 사용자가 원하는 군집 결과로 유도하므로, 가장 우수한 성능을 나타낸다.

<155> 이상으로 본 발명의 기술적 사상을 예시하기 위한 바람직한 실시예와 관련하여 설명하고 도시하였지만, 본 발명은 이와 같이 도시되고 설명된 그대로의 구성 및 작용에만 국한되는 것이 아니며, 기술적 사상의 범주를 일탈함이 없이 본 발명에 대해 다수의 변경 및 수정이 가능함을 당업자들은 잘 이해할 수 있을 것이다. 따라서 그러한 모든 적절한 변경 및 수정과 균등물들도 본 발명의 범위에 속하는 것으로 간주되어야 할 것이다.

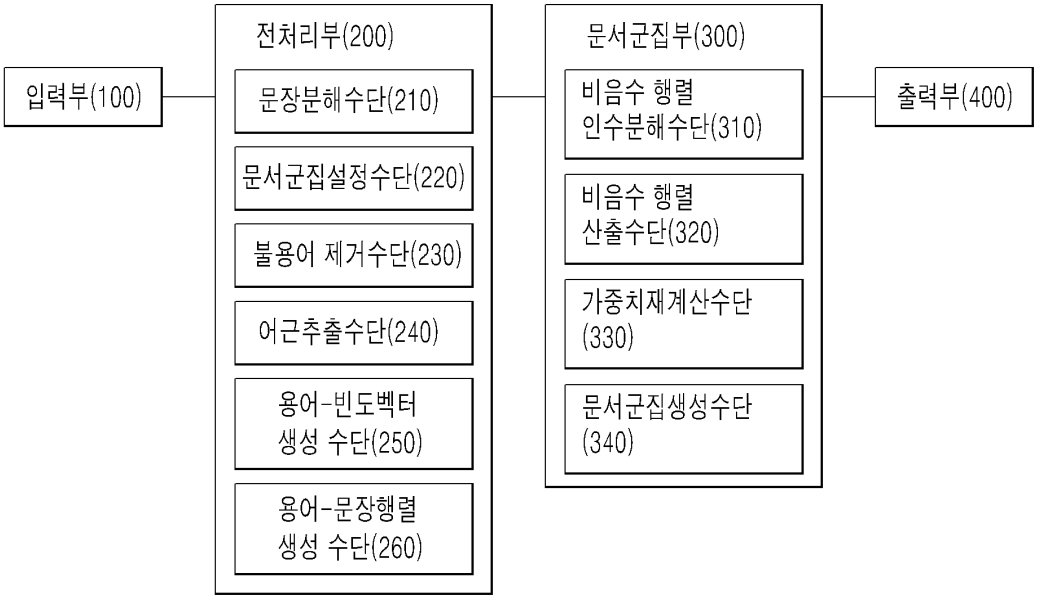
도면의 간단한 설명

<156> 도 1 은 본 발명의 바람직한 실시예에 따른 용어 가중치 재계산을 이용한 문서 군집 장치의 전체 구성도.

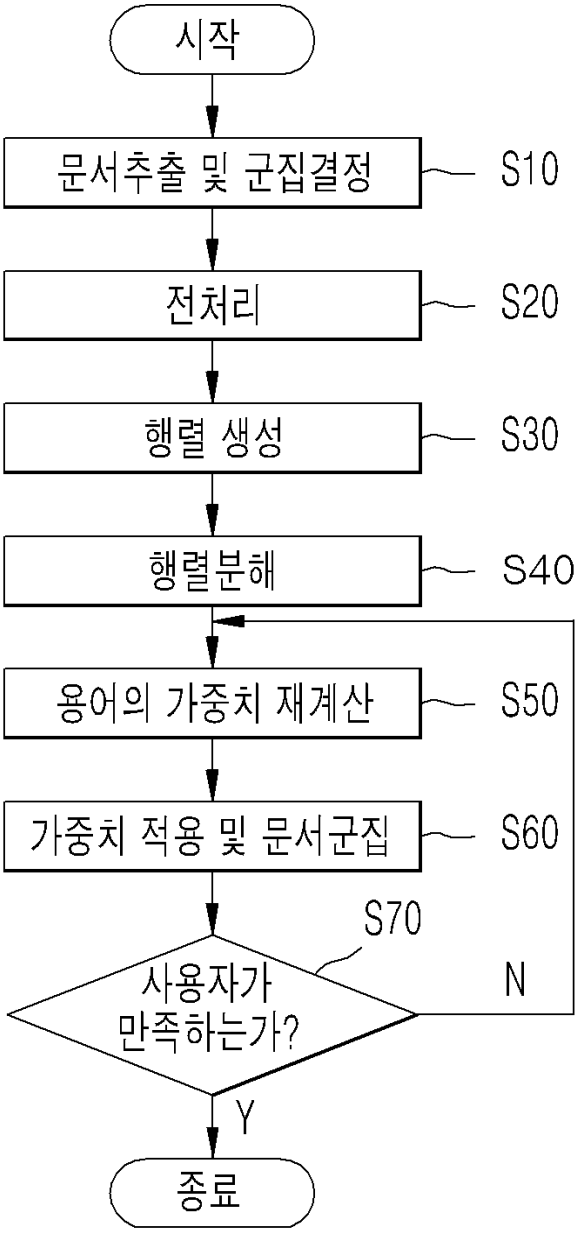
<157> 도 2 는 본 발명의 바람직한 실시예에 따른 용어 가중치 재계산을 이용한 문서 군집 방법의 전체 흐름도.

도면

도면1


입력부(100)
전처리부(200)
문장분해수단(210)
문서군집설정수단(220)
불용어 제거수단(230)
어근추출수단(240)
용어-빈도벡터 생성 수단(250)
용어-문장행렬 생성 수단(260)
문서군집부(300)
비음수 행렬 인수분해수단(310)
비음수 행렬 산출수단(320)
가중치재계산수단 (330)
문서군집생성수단 (340)
출력부(400)

도면2


시작
문서추출 및 군집결정
S10
전처리
S20
행렬 생성
S30
행렬분해
S40
용어의 가중치 재계산
S50
가중치 적용 및 문서군집
S60
S70
사용자가
만족하는가?
N
Y
종료