(19)대한민국특허청(KR)
(12) 공개특허공보(A)

(51) 。Int. Cl.
G06F 17/30 (2006.01)

(11) 공개번호 10-2007-0006367
(43) 공개일자 2007년01월11일

(21) 출원번호 10-2005-0061647
(22) 출원일자 2005년07월08일
심사청구일자 2005년07월08일

(71) 출원인 울산대학교 산학협력단
울산 남구 무거동 산29

(72) 발명자 이창범
울산 남구 무거2동 산 29 울산대학교7호관 324-1호
최호섭
울산 남구 무거2동 산 29 울산대학교 7호관 324-1호
옥철영
울산 남구 무거2동 산 29 울산대학교 7호관 324호
박혁로
광주 북구 용봉동 300번지 전남대학교 공동실험실습관508호

(74) 대리인 장한특허법인

전체 청구항 수 : 총 10 항

(54) 문서 중요 단어 추출 장치 및 방법

(57) 요약

본 발명은 문서 중요 단어 추출 장치 및 방법에 관한 것으로서, 주성분 분석과 비정칙치 분해를 이용하여 문서의 중요 단어를 추출하는, 문서 중요 단어 추출 장치 및 방법에 관한 것임.

본 발명은 문서 내의 중요 단어를 추출하기 위한 장치로서, 문서 내의 각 단어를 좌표 축으로 하는 다차원 공간을 생성하고, 그에 상응하는 문장-단어 행렬을 생성하기 위한 행렬 생성 수단, 상기 행렬 생성 수단에서 생성한 문장-단어 행렬을 주성분 분석하고, 그 결과로 선택된 각 주성분을 상응하는 고유벡터의 계수로 표현한 후, 고유값으로 정량화하기 위한 주성분 분석 수단 및 상기 주성분 분석 수단에서 정량화한 각 주성분에 대응하는 고유값의 누적비율 및 주성분적재계수를 고려하여 주성분 가운데 중요 단어를 추출하기 위한 중요 단어 추출 수단을 포함함.

대표도

도 1

특허청구의 범위

청구항 1.

문서 내의 중요 단어를 추출하기 위한 장치로서,

문서 내의 각 단어를 좌표 축으로 하는 다차원 공간을 생성하고, 그에 상응하는 문장-단어 행렬을 생성하기 위한 행렬 생성 수단;

상기 행렬 생성 수단에서 생성한 문장-단어 행렬을 주성분 분석하고, 그 결과로 선택된 각 주성분을 상응하는 고유벡터의 계수로 표현한 후, 고유값으로 정량화하기 위한 주성분 분석 수단; 및

상기 주성분 분석 수단에서 정량화한 각 주성분에 대응하는 고유값의 누적비율 및 주성분적재계수를 고려하여 주성분 가운데 중요 단어를 추출하기 위한 중요 단어 추출 수단

을 포함하는 문서 중요 단어 추출 장치.

청구항 2.

문서 내의 중요 단어를 추출하기 위한 장치로서,

문서 내의 각 단어를 좌표 축으로 하는 다차원 공간을 생성하고, 그에 상응하는 문장-단어 행렬을 생성하기 위한 행렬 생성 수단;

상기 행렬 생성 수단에서 생성한 문장-단어 행렬을 비정칙치 분해하여 노이즈를 제거하기 위한 비정칙치 분해 수단;

상기 비정칙치 분해 수단에서 노이즈를 제거한 문장-단어 행렬을 주성분 분석하고, 그 결과로 선택된 각 주성분을 상응하는 고유벡터의 계수로 표현한 후, 고유값으로 정량화하기 위한 주성분 분석 수단; 및

상기 주성분 분석 수단에서 정량화한 각 주성분에 대응하는 고유값의 누적비율 및 주성분적재계수를 고려하여 주성분 가운데 중요 단어를 추출하기 위한 중요 단어 추출 수단

을 포함하는 문서 중요 단어 추출 장치.

청구항 3.

제 1 항 또는 제 2 항에 있어서,

상기 주성분 분석 수단은,

$$\rho_k = \frac{\sum_{i=1}^{k} \lambda_k}{\lambda_1 + \lambda_2 + \ldots + \lambda_p} \quad , where \quad k = 1, 2, \ldots, p$$

의 식을 이용하여 주성분을 선택하는

문서 중요 단어 추출 장치.

청구항 4.

제 1 항 또는 제 2 항에 있어서,

상기 중요 단어 추출 수단은,

상기 주성분 분석 수단에서 정량화한 각 주성분에 대응하는 고유값의 누적비율이 제 1 기설정치 이상인 주성분을 선택한 후, 그 주성분의 적재계수가 제 2 기설정치 이상인 단어들을 중요 단어로 선택하는

문서 중요 단어 추출 장치.

청구항 5.

제 4 항에 있어서,

상기 제 1 기설정치는 90% 이고, 상기 제 2 기설정치는 0.5 인

문서 중요 단어 추출 장치.

청구항 6.

문서 내의 중요 단어를 추출하기 위한 방법으로서,

문서 내의 각 단어를 좌표 축으로 하는 다차원 공간을 생성하고, 그에 상응하는 문장-단어 행렬을 생성하는 행렬 생성 단계;

상기 생성한 문장-단어 행렬을 주성분 분석하는 주성분 분석 단계;

상기 주성분 분석 결과 선택된 각 주성분을 상응하는 고유벡터의 계수로 표현하는 고유벡터 표현 단계;

각 주성분의 고유벡터를 상응하는 고유값으로 정량화화는 정량화 단계; 및

상기 정량화한 각 주성분에 대응하는 고유값의 누적비율 및 주성분적재계수를 고려하여 주성분 가운데 중요 단어를 추출하는 중요 단어 추출 단계

를 포함하는 문서 중요 단어 추출 방법.

청구항 7.

문서 내의 중요 단어를 추출하기 위한 방법으로서,

문서 내의 각 단어를 좌표 축으로 하는 다차원 공간을 생성하고, 그에 상응하는 문장-단어 행렬을 생성하는 행렬 생성 단계;

상기 생성한 문장-단어 행렬을 문장-단어 행렬을 비정칙치 분해하여 노이즈를 제거하는 노이즈 제거 단계;

상기 노이즈를 제거한 문장-단어 행렬을 주성분 분석하는 주성분 분석 단계;

상기 주성분 분석 결과 선택된 각 주성분을 상응하는 고유벡터의 계수로 표현하는 고유벡터 표현 단계;

각 주성분의 고유벡터를 상응하는 고유값으로 정량화화는 정량화 단계; 및

상기 정량화한 각 주성분에 대응하는 고유값의 누적비율 및 주성분적재계수를 고려하여 주성분 가운데 중요 단어를 추출하는 추출 단계

를 포함하는 문서 중요 단어 추출 방법.

청구항 8.

제 6 항 또는 제 7 항에 있어서,

상기 주성분 분석 단계는,

$$\rho_k = \frac{\sum_{i=1}^{k} \lambda_k}{\lambda_1 + \lambda_2 + \ldots + \lambda_p} \quad , where \quad k = 1, 2, \ldots, p$$

의 식을 이용하여 주성분을 선택하는

문서 중요 단어 추출 방법.

청구항 9.

제 6 항 또는 제 7 항에 있어서,

상기 중요 단어 추출 단계는,

상기 정량화한 각 주성분에 대응하는 고유값의 누적비율이 제 1 기설정치 이상인 주성분을 선택한 후, 그 주성분의 적재계수가 제 2 기설정치 이상인 단어들을 중요 단어로 선택하는

문서 중요 단어 추출 방법.

청구항 10.

제 9 항에 있어서,

상기 제 1 기설정치는 90% 이고, 상기 제 2 기설정치는 0.5 인

문서 중요 단어 추출 방법.

명세서

발명의 상세한 설명

발명의 목적

발명이 속하는 기술 및 그 분야의 종래기술

본 발명은 문서 중요 단어 추출 장치 및 방법에 관한 것으로서, 주성분 분석과 비정칙치 분해를 이용하여 문서의 중요 단어를 추출하는, 문서 중요 단어 추출 장치 및 방법에 관한 것이다.

컴퓨터와 인터넷 기술이 발전함에 따라, 개인이 접하는 문서의 양은 방대하게 증가하여, 개인이 스스로 처리할 수 있는 수준을 넘어서고 있다. 따라서, 문서로부터 중요 단어를 추출하고, 검색할 수 있는 방안이 요구되고 있다.

종래에는 텍스트 문서의 내용을 대표하는 단어를 추출하는데 있어서, 대용량 말뭉치, 시소러스와 같은 정보 도구를 이용하였다. 그런데, 이러한 종래 방법은 일반적인 목적을 위하여 구축된 정보 도구를 이용하기 때문에 특정한 도메인에는 적합하지 않을 수도 있으며, 외부적인 리소스가 요구되는 문제점이 있었다.

따라서, 대용량 말뭉치, 시소러스와 같은 정보 도구를 이용하지 않고서도 문서 내의 중요 단어를 추출할 수 있는 방안이 절실히 요구된다.

발명이 이루고자 하는 기술적 과제

본 발명은 상기와 같은 요구에 부응하기 위하여 창안된 것으로서, 대용량 말뭉치 또는 시소러스와 같은 정보 도구를 이용하지 않고, 주성분 분석과 비정칙치 분해를 이용하여 문서의 중요 단어를 추출하는, 문서 중요 단어 추출 장치 및 방법을 제공하는데 그 목적이 있다.

본 발명의 다른 목적 및 장점들은 하기에 설명될 것이며, 본 발명의 실시예에 의해 알게 될 것이다. 또한, 본 발명의 목적 및 장점들은 특허청구범위에 나타낸 수단 및 조합에 의해 실현될 수 있다.

발명의 구성

상기와 같은 목적을 달성하기 위한 본 발명은 문서 내의 중요 단어를 추출하기 위한 장치로서, 문서 내의 각 단어를 좌표축으로 하는 다차원 공간을 생성하고, 그에 상응하는 문장-단어 행렬을 생성하기 위한 행렬 생성 수단, 상기 행렬 생성 수단에서 생성한 문장-단어 행렬을 주성분 분석하고, 그 결과로 선택된 각 주성분을 상응하는 고유벡터의 계수로 표현한 후, 고유값으로 정량화하기 위한 주성분 분석 수단 및 상기 주성분 분석 수단에서 정량화한 각 주성분에 대응하는 고유값의 누적비율 및 주성분적재계수를 고려하여 주성분 가운데 중요 단어를 추출하기 위한 중요 단어 추출 수단을 포함한다.

또한, 본 발명은 문서 내의 중요 단어를 추출하기 위한 장치로서, 문서 내의 각 단어를 좌표 축으로 하는 다차원 공간을 생성하고, 그에 상응하는 문장-단어 행렬을 생성하기 위한 행렬 생성 수단, 상기 행렬 생성 수단에서 생성한 문장-단어 행렬을 비정칙치 분해하여 노이즈를 제거하기 위한 비정칙치 분해 수단, 상기 비정칙치 분해 수단에서 노이즈를 제거한 문장-단어 행렬을 주성분 분석하고, 그 결과로 선택된 각 주성분을 상응하는 고유벡터의 계수로 표현한 후, 고유값으로 정량화하기 위한 주성분 분석 수단 및 상기 주성분 분석 수단에서 정량화한 각 주성분에 대응하는 고유값의 누적비율 및 주성분적재계수를 고려하여 주성분 가운데 중요 단어를 추출하기 위한 중요 단어 추출 수단을 포함한다.

한편, 본 발명은 문서 내의 중요 단어를 추출하기 위한 방법으로서, 문서 내의 각 단어를 좌표 축으로 하는 다차원 공간을 생성하고, 그에 상응하는 문장-단어 행렬을 생성하는 행렬 생성 단계, 상기 생성한 문장-단어 행렬을 주성분 분석하는 주성분 분석 단계, 상기 주성분 분석 결과 선택된 각 주성분을 상응하는 고유벡터의 계수로 표현하는 고유벡터 표현 단계, 각 주성분의 고유벡터를 상응하는 고유값으로 정량화화는 정량화 단계 및 상기 정량화한 각 주성분에 대응하는 고유값의 누적비율 및 주성분적재계수를 고려하여 주성분 가운데 중요 단어를 추출하는 중요 단어 추출 단계를 포함한다.

또한, 본 발명은 문서 내의 중요 단어를 추출하기 위한 방법으로서, 문서 내의 각 단어를 좌표 축으로 하는 다차원 공간을 생성하고, 그에 상응하는 문장-단어 행렬을 생성하는 행렬 생성 단계, 상기 생성한 문장-단어 행렬을 문장-단어 행렬을 비정칙치 분해하여 노이즈를 제거하는 노이즈 제거 단계, 상기 노이즈를 제거한 문장-단어 행렬을 주성분 분석하는 주성분 분석 단계, 상기 주성분 분석 결과 선택된 각 주성분을 상응하는 고유벡터의 계수로 표현하는 고유벡터 표현 단계, 각 주성분의 고유벡터를 상응하는 고유값으로 정량화화는 정량화 단계 및 상기 정량화한 각 주성분에 대응하는 고유값의 누적비율 및 주성분적재계수를 고려하여 주성분 가운데 중요 단어를 추출하는 추출 단계를 포함한다.

이하 첨부된 도면을 참조로 본 발명의 바람직한 실시예를 상세히 설명하기로 한다. 이에 앞서, 본 명세서 및 청구범위에 사용된 용어나 단어는 통상적이거나 사전적인 의미로 한정해서 해석되어서는 아니 되며, 발명자는 그 자신의 발명을 가장 최선의 방법으로 설명하기 위해 용어의 개념을 적절하게 정의할 수 있다는 원칙에 입각하여 본 발명의 기술적 사상에 부합하는 의미와 개념으로 해석되어야만 한다.

따라서, 본 명세서에 기재된 실시 예와 도면에 도시된 구성은 본 발명의 가장 바람직한 일 실시 예에 불과할 뿐이고 본 발명의 기술적 사상을 모두 대변하는 것은 아니므로, 본 출원시점에 있어서 이들을 대체할 수 있는 다양한 균등물과 변형예들이 있을 수 있음을 이해하여야 한다.

도 1은 본 발명의 일실시예에 따른 문서 중요 단어 추출 장치의 구성도이다.

전통적인 정보 검색 시스템은 일반적으로 그 자체로 인덱스 의미를 가지는 명사 또는 문서의 논리적 관점을 나타내는 대표 단어를 이용한다. 그러나, 문서안의 모든 명사가 항상 대표어(representatives)가 될 수는 없다. 예를 들어, “apple”, “banana”, “pear”, “strawberry”, “fruit” 그리고 “basket”과 같은 명사를 포함하는 과일과 관련된 문서를 고려해보기로 한다. 그리고, “apple”, “banana”, “pear”, “basket”의 네 단어가 대표어(representatives)로 선택되었다고 가정해본다.

“basket”이란 단어는 “a fruit basket”이라는 구로서 나타날 수도 있을 것이다. “basket”을 선택 리스트에서 제외하고, “strawberry” 또는 “fruit”를 추가하면, 교정된 단어 리스트는 좀 더 많은 정보를 제공하고, 예시(example)를 묘사하기에 효과적일 수 있다.

본 발명에서는 위에서 살펴본 예와 같이 어법(term usage, 언어용법, 관용법)의 노이즈를 효과적으로 제거하고, 문서 내의 대표어(representatives) 또는 중요 단어를 추출하기 위해서, 고유치-고유벡터 쌍에 의하여 데이터 차원을 줄이는 주성분 분석(PCA)을 이용한다.

주성분 분석(PCA : Principal Component Analysis)은 원래 변수들의 선형결합으로 표시되는 새로운 주성분(principal components)을 찾아서, 이를 통하여 자료의 요약과 용이한 해석을 제공한다.

주성분 분석(PCA)은 몇몇의 선형결합을 통하여 분산-공분산 행렬 구조를 설명하는 것과 관계가 있다. 대수적인 측면에서 모집단 주성분은 모집단의 공분산 행렬이 알려져 있다는 전제하에서 어떤 판정기준이 최적화되도록 원래의 변수들을 선형변환시킨 것이다. 따라서, 고려되는 판정기준들과 이에 연관된 주성분의 최적성은 통계적으로 중요한 의미를 함축하게 된다. 한편 기하학적으로는 이 선형결합들은 원래의 좌표체계를 직교회전시켜 얻어지는 새로운 좌표축의 선택과 관련되어 있다.

한편, 고유벡터(eigenvector)와 그에 상응하는 고유값(eigenvalue)은 주성분 분석(PCA)을 공분산 행렬에 적용함으로써 얻을 수 있다. 바꿔 말하면, 공분산 행렬은 일련의 고유값과 고유벡터의 쌍으로 분해될 수 있음을 뜻한다. 여기서, k번째 주성분까지의 고유값 누적 비율(Principal Componet, ρ_k)은 하기의 [수학식1]과 같이 표현된다.

$$\rho_k = \frac{\sum_{i=1}^{k} \lambda_k}{\lambda_1 + \lambda_2 + \ldots + \lambda_p} \quad , where \quad k = 1, 2, \ldots, p$$

주성분 분석(PCA)은 분산-공분산 구조를 이용하기 때문에 문장-단어 행렬의 통계적 공기(cooccurrence) 패턴으로부터 중요 단어를 추출할 수 있도록 한다.

이를 위하여, 본 발명의 일실시예에 따른 문서 중요 단어 추출 장치(100)는, 도 1에 도시된 바와 같이 행렬 생성부(11), 주성분 분석부(12) 및 중요 단어 선택부(13)를 포함한다.

행렬 생성부(11)는 문서 내의 각 단어를 좌표 축으로 하는 다차원 공간을 생성하고, 그에 상응하는 문장-단어 행렬을 생성한다.

주성분 분석부(12)는 행렬 생성부(11)에서 생성한 문장-단어 행렬을 주성분 분석하여 주성분을 선택하고, 선택한 각 주성분을 상응하는 고유벡터의 계수로 표현한 후, 하기의 [표 1]을 이용하여 고유값으로 정량화한다. 여기서, 주성분 분석부(12)가 주성분을 추출할 때는 상기의 [수학식1]을 이용한다.

[표 1]

term \ PC	PC_1	PC_2	PC_3	PC_4	PC_5	PC_6	PC_7	PC_8	PC_9	PC_{10}
X_1	-0.67	-0.39	0.29	0.07	0.25	-0.15	-0.48	0.00	0.00	0.00
X_2	0.00	-0.28	-0.09	0.45	0.15	0.54	0.16	-0.04	0.26	-0.55
X_3	-0.36	0.10	0.26	0.28	0.03	-0.37	0.76	0.00	0.00	0.00
X_4	0.06	-0.24	0.44	0.00	-0.03	0.58	0.19	0.04	-0.26	0.55
X_5	0.16	-0.37	0.06	0.21	-0.54	-0.21	-0.06	0.66	0.11	-0.01
X_6	**0.44**	-0.32	-0.18	0.37	0.58	-0.32	-0.01	0.02	-0.13	0.27
X_7	-0.16	0.06	-0.18	-0.44	0.48	0.16	0.19	0.66	0.11	-0.01
X_8	0.02	**0.61**	0.24	**0.46**	0.12	0.10	-0.28	0.22	0.38	0.24
X_9	0.19	0.23	0.41	0.05	0.14	-0.03	-0.12	0.22	-0.64	-0.50
X_{10}	0.37	-0.16	**0.59**	-0.36	0.15	-0.15	0.04	-0.14	0.52	-0.16
eigenvalue(λ_i)	1.39	0.95	0.62	0.47	0.26	0.10	0.01	0.00	0.00	0.00
cumulative ratio(ρ_k)	0.36	0.62	0.78	**0.90**	0.97	1.00	1.00	1.00	1.00	1.00

중요 단어 선택부(13)는 주성분 분석부(12)에서 정량화한 각 주성분에 대응하는 고유값의 누적비율 및 주성분적재계수를 고려하여 주성분 가운데 중요 단어를 추출한다. 즉, 중요 단어 선택부(13)는 주성분 분석부(12)에서 정량화한 각 주성분에 대응하는 고유값의 누적비율을 이용하여 기설정치(90%) 이상이 되는 주성분을 선택한 후, 그 주성분의 적재계수가 기설정치(0.5) 이상 또는 최고치인 단어들을 중요 단어로 선택한다.

이처럼, 주성분 분석을 이용하면 문서로부터 중요 단어를 추출할 수 있으나, 어법(term usage) 상에 존재하는 노이즈를 효과적으로 제거하기는 어렵다. 즉, 앞에서 설명한 본 발명의 일실시예에서는 다양한 어법(term usage, 언어용법, 관용법)에 의한 노이즈가 포함되어 있을 원 자료 행렬(행 : 문장, 열 : 단어)에 대하여 주성분 분석(PCA)을 수행한다. 결과적으로, 주성분 분석(PCA)만 수행하게 되면 필수적이지 않은 단어가 중요 단어로 선택될 수 있는 가능성이 있다.

이하에서는 원 자료 행렬(행 : 문장, 열 : 단어)에 포함되어 있을 수 있는 노이즈 데이터를 제거하는 본 발명의 다른 실시예를 설명한다. 원 자료 행렬(행 : 문장, 열 : 단어)을 비정칙치 분해(SVD)하면 주성분의 비율에 따라 노이즈 데이터를 제거할 수 있다. 결과적으로 새로운 문장-단어 행렬 또는 교정된 문장-단어 행렬의 행수 및 열수는 원 자료 행렬의 행수와 열수와 정확히 일치한다. 그러나, 이 행렬은 원 자료 행렬 보다 적은 양의 노이즈를 가질 수 있다.

도 2를 참조하여 본 발명의 다른 실시예를 살펴보기로 한다.

도 2는 본 발명의 다른 실시예에 따른 문서 중요 단어 추출 장치의 구성도이다.

본 발명의 다른 실시예에 따른 문서 중요 단어 추출 장치(200)는 비정칙치 분해(SVD)를 이용하여 다양한 어법(term usage, 언어용법, 관용법) 내의 노이즈를 제거한다.

A는 문장을 행으로 하고 단어를 열로 하는 s×t 행렬이다. 행렬 A는 s×r 열-직교 행렬(U_0), 양수 또는 0을 엘리먼트로 가지는 r×r 대각 행렬(W_0), 그리고 t×r 직교 행렬(V_0)의 전치 행렬의 곱으로 쓰여질 수 있을 것이다. 여기서, r은 행렬 A의 열($r \leq \min(s,t)$)일 때, 비정칙치 분해(SVD)는 하기의 [수학식 2]를 이용한다.

$$A = U_0 \cdot W_0 \cdot V_0^T$$

여기서, $U_0^T U_0 = I$, $V_0^T V_0 = I$ 이고, W_0는 주성분의 대각 행렬로서, 비정칙치를 나타낸다.

실제로, 차원을 줄이기 위해서는 전체 열 대신에 하기의 [수학식3]에 의하여 계산된 열, r을 이용할 수 있다. σ_k가 90% 이상일 때, k는 차원을 줄이기 위하여 문장과 단어의 조합에서 가장 중요한 기초를 이루는 구조를 획득할 수 있을만큼 충분히 큰 수 가운데 선택될 수 있다.

$$\sigma_k = \frac{\sum_{i=1}^{k} w_i}{w_1 + w_2 + \ldots + w_r}, \quad k = 1, 2, \ldots, r$$

한편, 문장-단어 행렬 A는 [수학식4]와 같이 정보의 손실 없이 A와 차원이 같은 교정된 행렬 $\hat{\mathbf{A}}$ (s×t)로 다시 쓰여질 수 있다.

$$A \approx \hat{A} = U \cdot W \cdot V^T$$

여기서, U가 s×k 행렬일 때, W는 k×k 행렬이고, V^T는 k×t 행렬이다. 따라서, 원 행렬의 중요한 구조는 지키면서 원 문장-단어 행렬의 빈도수 분포에 존재하는 노이즈를 제거하기 위하여 원 문장-단어 행렬 A 대신에 교정된 문장-단어 행렬 $\hat{\mathbf{A}}$를 이용할 수 있다.

이하에서는 예제 지문을 이용하여 교정된 문장-단어 행렬($\hat{\mathbf{A}}$)을 생성하고, 생성한 문장-단어 행렬을 이용하여 문서 내에서 다양한 어법(term usage)의 노이즈를 제거하는 과정을 살펴보기로 한다. 하기의 [표2]는 8개 문장과 61개의 단일한 단어로 구성된 한국 신문 기사(환경 보호 상 수여와 관련된 기사)에서 추출한 단어 리스트를 보여주고 있다. 본 발명에서는 문서 내에서 두번 이상 쓰인 명사를 중요 단어의 후보로 간주하기 때문에, [표2]에는 10개 단어가 포함되어 있다. [표2]에서 X4 단어는 한국어 형태소 분석기의 기능 불량으로 인한 올바르지 못한 명사이다.

[표 2]

Term (Noun)	Notation	Remark
hwan-gyeong (Environment)	X_1	
UNEP (United Nations Environment Program)	X_2	
se-gye (world)	X_3	
i-ran (called or named)	X_4	extraction error
sang (prize)	X_5	
cheong-so-nyeon (young people)	X_6	
ta-im (TIME magazine)	X_7	
go-ril-ra (gorilla)	X_8	
ja-yeon (nature)	X_9	
il-bon (Japan)	X_{10}	

아래의 행렬 A는 각 문장 내에서 각 단어의 빈도수를 가지는 원 문장-단어 행렬을 보여준다. 예를 들어, 첫번째 열을 살펴보면 X1 단어가 첫번째 문장에서 2번 발생하였고, 두번째 문장에서 1번, 다섯번째 문장에서 2번 발생했음을 보여주고 있다.

$$\mathrm{A} = \begin{pmatrix} 2 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 1 & 2 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 2 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 2 & 1 & 0 \\ 0 & 0 & 0 & 1 & 1 & 1 & 0 & 0 & 1 & 2 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

행렬 $\hat{\mathbf{A}}$는 행렬 A를 [수학식 4]를 이용하여 비정칙치 분해(SVD)한 결과이다. 아래에서 볼 수 있듯이 행렬 $\hat{\mathbf{A}}$는 소수점 둘째자리에서 반올림된 값이다. 실제로, U_0는 8*10 행렬이고, W_0는 10*10 행렬이고, V_0는 10*10 행렬이다. 그러나, [수학식3]과 같이 σ_k 를 이용하면 문장-단어 행렬의 빈도수 분포 정보의 많은 손실 없이 차원을 줄일수 있다. 앞에서 설명한 기사를 예로 들면, 하기의 [표3]에서 볼 수 있듯이 0.9 보다 큰 주성분(singular value)의 누적율이 여섯번째에 나타나므로, k는 6이다. 이것은 누적율을 이용하여 약 10% 가량의 노이즈를 제거할 수 있음을 말해준다.

[표 3]

singular value	4.01	3.07	2.43	1.92	1.46	1.06	0.76	0.24	0.00	0.00
cumulative ratio	0.27	0.47	0.64	0.76	0.86	**0.93**	0.98	1.00	1.00	1.00

$$\hat{\mathbf{A}} = \begin{pmatrix} 2.08 & 0.88 & 1.08 & 0.87 & 0.93 & 0.10 & -0.16 & -0.05 & 0.00 & 0.05 \\ 1.00 & 0.09 & 0.82 & 0.09 & 1.09 & -0.10 & 0.13 & 0.09 & 0.02 & -0.06 \\ -0.02 & 1.03 & -0.02 & 0.03 & 1.02 & 1.97 & 0.04 & 0.01 & 0.00 & -0.01 \\ 0.17 & -0.22 & 0.08 & -0.25 & -0.12 & 0.19 & 0.71 & -0.06 & 0.02 & 0.09 \\ 1.90 & 0.10 & 1.00 & 0.11 & 0.04 & -0.08 & 1.13 & 0.01 & -0.01 & -0.04 \\ 0.01 & -0.02 & 1.02 & -0.02 & -0.01 & 0.02 & -0.02 & 1.99 & 1.00 & 0.01 \\ -0.01 & 0.02 & -0.02 & 1.02 & 1.01 & 0.98 & 0.03 & 0.01 & 1.00 & 1.99 \\ -0.01 & -0.07 & 0.18 & -0.08 & 0.92 & 0.09 & -0.12 & -0.09 & -0.02 & 0.05 \end{pmatrix}$$

$$\mathbf{U} = \begin{pmatrix} -0.65 & 0.20 & -0.17 & 0.00 & 0.41 & 0.43 \\ -0.37 & 0.14 & -0.02 & -0.04 & 0.25 & -0.59 \\ -0.28 & -0.49 & -0.27 & -0.72 & -0.28 & 0.06 \\ -0.03 & 0.05 & -0.01 & 0.06 & -0.52 & -0.26 \\ -0.45 & 0.44 & -0.04 & 0.16 & -0.59 & -0.03 \\ -0.18 & 0.02 & 0.93 & -0.31 & 0.00 & 0.04 \\ -0.34 & -0.70 & 0.15 & 0.59 & -0.09 & 0.02 \\ -0.11 & -0.10 & -0.06 & -0.06 & 0.26 & -0.63 \end{pmatrix}$$

$$\mathbf{W} = \begin{pmatrix} 4.01 & & & & & \\ & 3.07 & & & & \\ & & 2.43 & & & \\ & & & 1.92 & & \\ & & & & 1.46 & \\ & & & & & 1.06 \end{pmatrix}$$

$$\mathbf{V^T} = \begin{pmatrix} -0.64 & -0.23 & -0.41 & -0.25 & -0.44 & -0.22 & -0.12 & -0.09 & -0.13 & -0.17 \\ 0.47 & -0.09 & 0.26 & -0.16 & -0.31 & -0.55 & 0.16 & 0.01 & -0.22 & -0.46 \\ -0.18 & -0.18 & 0.29 & -0.01 & -0.16 & -0.16 & -0.02 & 0.77 & 0.44 & 0.12 \\ 0.14 & -0.38 & -0.10 & 0.31 & -0.12 & -0.44 & 0.11 & -0.32 & 0.15 & 0.62 \\ -0.07 & 0.09 & 0.05 & 0.22 & 0.38 & -0.44 & -0.76 & 0.01 & -0.06 & -0.12 \\ 0.19 & 0.46 & -0.15 & 0.42 & -0.67 & 0.14 & -0.27 & 0.07 & 0.05 & 0.04 \end{pmatrix}$$

한편, $\hat{\mathbf{A}}$ 행렬의 고유벡터와 그에 상응하는 고유값은 하기의 [표4]와 같다.

[표 4]

term \ PC	PC_1	PC_2	PC_3	PC_4	PC_5	PC_6	PC_7	PC_8	PC_9	PC_{10}
X_1	-0.66	-0.38	0.28	0.06	0.23	-0.04	-0.46	-0.17	0.00	-0.20
X_2	-0.01	-0.30	-0.07	**0.49**	0.21	0.28	-0.02	0.28	-0.22	0.64
X_3	-0.35	0.11	0.23	0.26	0.00	-0.31	0.48	0.60	0.17	-0.17
X_4	0.05	-0.26	0.48	0.06	0.04	0.33	0.60	-0.46	0.07	-0.02
X_5	0.15	-0.38	0.08	0.23	-0.52	-0.64	-0.02	-0.21	-0.08	0.17
X_6	**0.45**	-0.30	-0.22	0.32	0.54	-0.20	0.01	-0.01	0.16	-0.44
X_7	-0.17	0.04	-0.14	-0.40	0.52	-0.47	0.28	-0.20	-0.12	0.40
X_8	0.02	**0.61**	0.24	0.48	0.15	-0.14	-0.20	-0.34	0.31	0.20
X_9	0.19	0.23	0.41	0.06	0.14	-0.10	-0.07	0.06	-0.81	-0.20
X_{10}	0.37	-0.15	**0.58**	-0.36	0.13	-0.06	-0.27	0.33	0.34	0.22
eigenvalue(λ_i)	1.39	0.95	0.62	0.47	0.26	0.04	0.00	0.00	0.00	0.00
cumulative ratio(ρ_k)	0.37	0.63	0.79	**0.92**	0.99	1.00	1.00	1.00	1.00	1.00

$\hat{\mathbf{A}}$와 A를 비교해보면 서로 다르다. 예를 들어, $\hat{\mathbf{A}}$의 첫번째 열, 첫번째 값은 2.28이고, A보다 0.28 증가한 것을 알 수 있고, 첫번째 열, 다섯번째 값은 1.90이고, A보다 0.1 감소한 것을 알 수 있다. 결과적으로, 행렬 $\hat{\mathbf{A}}$는 비정칙치 분해함으로써 원 행렬 A에서 어법(term usage) 상의 노이즈가 제거된 것이라고 볼 수 있다. 따라서, 교정된 행렬 $\hat{\mathbf{A}}$에 의하여 추출된 중요 단어는 보다 합리적일 수 있다.

[표1]과 [표4]에서 행렬 A와 $\hat{\mathbf{A}}$의 고유벡터와 그에 상응하는 고유값을 각각 보여주고 있다. [표1]과 [표4]에서 첫번째 4개의 주성분은 ρ_k 0.90, 0.92에 의하여 각각 선택된다. 이것은 A와 $\hat{\mathbf{A}}$의 분산이 이 4개의 주성분으로 매우 잘 요약될 수 있음을 의미한다.

이처럼, 어법(term usage) 상의 노이즈를 제거하면서 중요 단어를 추출하기 위하여 본 발명의 다른 실시예에 따른 문서 중요 단어 추출 장치(200)는 도 2에 도시된 바와 같이, 행렬 생성부(21), 비정칙치 분해부(22), 주성분 분석부(23) 및 중요 단어 선택부(24)를 포함한다.

행렬 생성부(21)는 문서 내의 각 단어를 좌표 축으로 하는 다차원 공간을 생성하고, 그에 상응하는 문장-단어 행렬을 생성한다.

비정칙치 분해부(22)는 행렬 생성부(21)에서 생성한 문장-단어 행렬을 비정칙치 분해하여 노이즈를 제거한다. 즉, 비정칙치 분해부(22)는 행렬 생성부(21)에서 생성한 문장-단어 행렬을 세개의 행렬, $U_0 \cdot W_0 \cdot V_0^T$ 로 분해한다. 여기서, W_0는 대각 행렬이며 비정칙치를 나타낸다. W_0의 모든 원소를 이용하여 $U_0 \cdot W_0 \cdot V_0^T$를 계산한다면 원래의 문장-단어 행렬 A와 정확히 일치한다. 그러나, 비정칙치의 누적 비율을 계산하여 90% 이상인 원소 만을 이용하여 세 개의 행렬에 대한 곱셈을 할 수 있다. 그러면, 원래의 문장-단어 행렬과 행과 열의 수가 같은 행렬을 구할 수 있다. 그러나, 원래의 행렬에 내재되어 있던 노이즈(자료 분포 형태에 도움이 되지 않는 단어)는 어느 정도 제거된다.

주성분 분석부(23)는 비정칙치 분해부(22)에서 노이즈를 제거한 문장-단어 행렬을 주성분 분석하여 주성분을 선택하고, 선택한 각 주성분을 상응하는 고유벡터의 계수로 표현한 후, 상기의 [표 1]을 이용하여 고유값으로 정량화한다. 여기서, 주성분 분석부(23)가 주성분을 추출할 때는 상기의 [수학식1]을 이용한다.

중요 단어 선택부(24)는 주성분 분석부(23)에서 정량화한 각 주성분에 대응하는 고유값의 누적비율 및 주성분적재계수를 고려하여 주성분 가운데 중요 단어를 추출한다. 즉, 중요 단어 선택부(24)는 주성분 분석부(23)에서 정량화한 각 주성분에 대응하는 고유값의 누적비율을 이용하여 기설정치(90%) 이상이 되는 주성분을 선택한 후, 그 주성분의 적재계수가 기설정치(0.5) 이상인 단어들을 중요 단어로 선택한다.

이하에서는 본 발명의 실시예에 따라 문서로부터 중요 단어를 추출하는 방법을 다시 한번 살펴보기로 한다.

도 3은 본 발명의 일실시예에 따른 문서 중요 단어 추출 방법에 대한 흐름도이다.

먼저, 문서 내의 각 단어를 좌표 축으로 하는 다차원 공간을 생성하고, 그에 상응하는 문장-단어 행렬(행 : 문장, 열 : 단어)을 생성한다(S301).

그리고, 생성한 문장-단어 행렬을 [수학식 1]을 이용하여 주성분 분석한다(S302). 그리고, 주성분 분석 결과 선택된 각 주성분을 상응하는 고유벡터의 계수로 표현하고(S303), 각 주성분의 고유벡터를 상응하는 고유값으로 정량화한다(S304).

이후, 정량화한 각 주성분에 대응하는 고유값의 누적비율 및 주성분적재계수를 고려하여 주성분 가운데 중요 단어를 추출한다(S305). 즉, 정량화한 각 주성분에 대응하는 고유값의 누적비율이 기설정치(90%) 이상인 주성분을 선택한 후, 그 주성분의 적재계수가 기설정치(0.5) 이상인 단어들을 중요 단어로 선택한다.

도 4는 본 발명의 다른 일실시예에 따른 문서 중요 단어 추출 방법에 대한 흐름도이다.

먼저, 문서 내의 각 단어를 좌표 축으로 하는 다차원 공간을 생성하고, 그에 상응하는 문장-단어 행렬(행 : 문장, 열 : 단어)을 생성한다(S401).

그리고, 생성한 문장-단어 행렬을 비정칙치 분해하여 노이즈를 제거한다(S402).

그리고, 노이즈를 제거한 문장-단어 행렬을 [수학식 1]을 이용하여 주성분 분석한다(S403), 그리고, 주성분 분석 결과 선택된 각 주성분을 상응하는 고유벡터의 계수로 표현하고(S404), 각 주성분의 고유벡터를 상응하는 고유값으로 정량화한다(S405).

이후, 정량화한 각 주성분에 대응하는 고유값의 누적비율 및 주성분적재계수를 고려하여 주성분 가운데 중요 단어를 추출한다(S406). 즉, 정량화한 각 주성분에 대응하는 고유값의 누적비율이 기설정치(90%) 이상인 주성분을 선택한 후, 그 주성분의 적재계수가 기설정치(0.5) 이상인 단어들을 중요 단어로 선택한다.

이상과 같이, 본 발명은 비록 한정된 실시예와 도면에 의해 설명되었으나, 본 발명은 이것에 의해 한정되지 않으며 본 발명이 속하는 기술 분야에서 통상의 지식을 가진 자에 의해 본 발명의 기술 사상과 아래에 기재될 특허 청구범위의 균등 범위 내에서 다양한 수정 및 변형이 가능함은 물론이다.

발명의 효과

상술한 바와 같이 본 발명은, 통계적 도구인 주성분 분석과 비정칙치 분해를 이용하여 문서로부터 중요 단어를 추출함으로써, 대용량 말뭉치, 시소러스와 같은 정보 도구를 이용하지 않고서도 문서 내의 중요 단어를 추출할 수 있는 효과가 있다.

도면의 간단한 설명

도 1은 본 발명의 일실시예에 따른 문서 중요 단어 추출 장치의 구성도,

도 2는 본 발명의 다른 실시예에 따른 문서 중요 단어 추출 장치의 구성도,

도 3은 본 발명의 일실시예에 따른 문서 중요 단어 추출 장법 중 비정칙치 분해 과정에 대한 흐름도,

도 4는 본 발명의 일실시예에 따른 문서 중요 단어 추출 방법 중 주성분 분석 과정에 대한 흐름도이다.

< 도면의 주요부분에 대한 부호의 설명 >

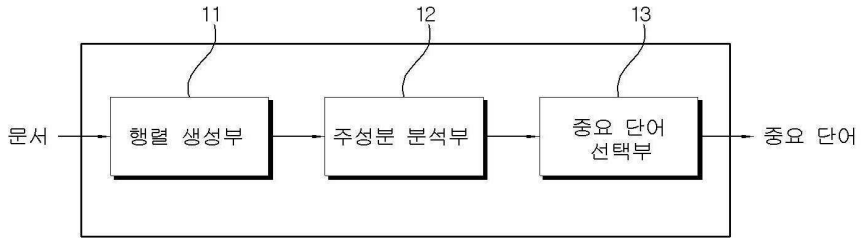
100,200 : 문서 중요 단어 추출 장치 11,21 : 행렬 생성부

12,23 : 주성분 분석부 13,24 : 중요 단어 선택부

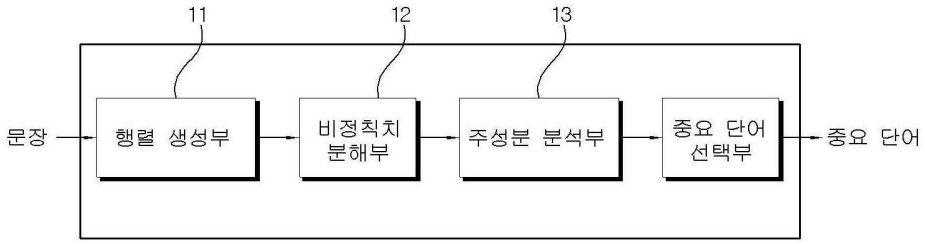
22 : 비정칙치 분해부

도면

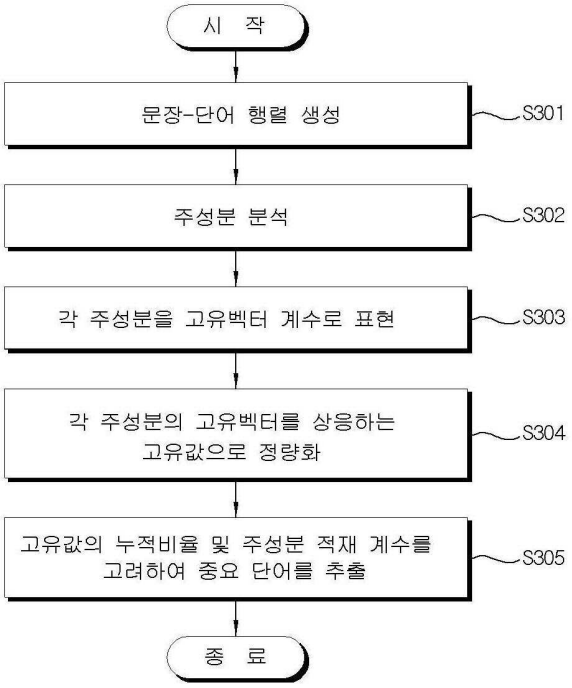
도면1



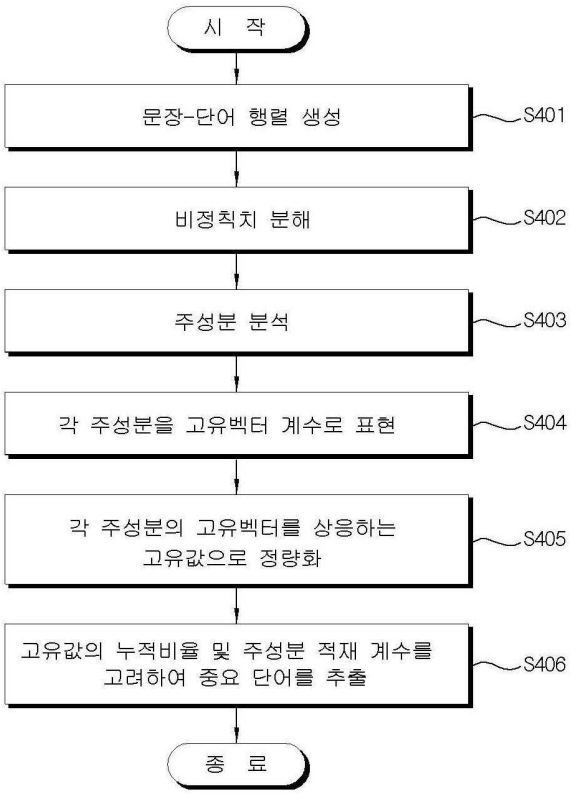
도면2



도면3



도면4


시 작
문장-단어 행렬 생성
S401
비정칙치 분해
S402
주성분 분석
S403
각 주성분을 고유벡터 계수로 표현
S404
각 주성분의 고유벡터를 상응하는 고유값으로 정량화
S405
고유값의 누적비율 및 주성분 적재 계수를 고려하여 중요 단어를 추출
S406
종 료