주성분 보유수에 따른 중요 용어 추출의 비교

이 창 범[†] · 옥 철 영[††] · 박 혁 로[†††]

요 약

문서를 구성하는 단어들은 서로 연관이 있다는 정보를 충분히 이용할 수 있는 다변량 분석 방법 중, 주성분분석(Principal Component Analysis)을 이용하여 중요 용어를 추출하고자 한다. 본 논문에서는 주성분분석의 분석 대상을 용어 사이의 공분산행렬이 아닌 상관행렬을 이용한다. 그리고, 중요 용어를 추출하기 위해서, 보유해야 할 주성분 개수와 주성분과 용어 사이의 상관계수에 대한 최적의 임계치를 찾고자 한다. 283건의 신문기사를 대상으로, 추출된 용어에 기반한 문장 추출 실험 결과, 첫 6개까지의 주성분과 상관계수 |0.4|라는 조건에서 가장 좋은 성능을 보였다.

키워드 : 용어 추출, 주성분 분석, 상관행렬, 문장 추출

Comparison of Significant Term Extraction Based on the Number of Selected Principal Components

Lee, Changbeom[†] · Ock, Cheolyoung[††] · Park, Hyukro[†††]

ABSTRACT

In this paper, we propose a method of significant term extraction within a document. The technique used is Principal Component Analysis(PCA) which is one of the multivariate analysis methods. PCA can sufficiently use term-term relationships within a document by term-term correlations. We use a correlation matrix instead of a covariance matrix between terms for performing PCA. We also try to find out thresholds of both the number of components to be selected and correlation coefficients between selected components and terms. The experimental results on 283 Korean newspaper articles show that the condition of the first six components with correlation coefficients of |0.4| is the best for extracting sentence based on the significant selected terms.

Key Words : Term Extraction, Principal Component Analysis, Correlation Matrix, Sentence Extraction

1. 서 론

중요 용어 추출이란 문서의 내용을 충분히 대변할 수 있는 단어 특히, 명사를 추출하는 응용이라 할 수 있다. 대용량의 말뭉치를 이용한 통계적 특성을 이용하는 방법[1, 2], 정보 추출의 일환으로 규칙, 기계학습 등을 이용하는 방법[5] 등을 이용하여 용어를 추출할 수 있을 것이다. 그런데, [9]에서는 다변량 분석 기법을 이용하여 문서내의 통계 정보만을 이용하여 중요 용어를 추출하였다.

다변량 분석(Multivariate Analysis)은 서로 상관있는 변량 데이터를 다루며, 특히 주성분 분석(PCA : Principal Component Analysis)은 다차원 특징 벡터로 이루어진 데이터를 높은 차원에서의 정보를 유지하면서 낮은 차원으로 차원을 축소시키는 방법이다[6-8, 10]. 하나의 문서에 나타나는 명사들은 서로 독립적으로 출현한다고는 볼 수 없으며, 어느 정도 서로 상관이 있다. 예를 들어, 봄나물에 관한 신문기사라면, "냉이", "풋마늘대", "씀바귀", "달래" 등의 단어가 함께 나올 가능성이 높다. 또한, 문서의 각 문장들은 그 문장에 나타나는 단어들에 의해 벡터로 표현될 수 있다. 이러한 맥락에서, [9]에서는 주성분 분석을 이용하여 문서내의 명사들 중에서 중요한 명사들을 선택하는 방법을 제안하였다.

[9]에서는 고유시스템(eigensystem)[3]의 입력으로 공분산행렬을 이용하여 주성분을 획득하며, 중요 용어를 추출하는데 있어 고유값(eigenvalue)의 누적 비율과 주성분 적재 계수만을 이용하고 있다. 하지만, 본 논문에서는 다음과 같이 확장 개선한다. 첫째, 표준화된 공분산 즉, 용어간의 상관행렬로부터 구해진 주성분을 이용한다. 표준화되지 않은 공분

※ 본 연구는 정보통신부 및 정보통신연구진흥원의 대학 IT연구센터 육성·지원사업의 연구결과로 수행되었음.
† 정 회 원 : 울산대학교 컴퓨터정보통신공학부 연구교수
†† 종신회원 : 울산대학교 컴퓨터정보통신공학부 교수
††† 종신회원 : 전남대학교 전자컴퓨터정보통신공학부 교수
논문접수 : 2006년 3월 9일, 심사완료 : 2006년 4월 27일

산행렬로부터 구해진 주성분은 분산(variance)이 큰 용어에 지배적이기[6] 때문에, 표준화하여 이러한 단점을 보완한다. 둘째, 고유값의 누적 비율을 이용하여 주성분을 선택한다면, 문서에 따라 보유할 주성분의 개수가 달라질 수 있다. 이에, 주성분 보유수에 따른 성능을 비교하여 보유할 최적의 주성분 개수를 제안하고자 한다. 셋째, 중요 용어를 추출하기 위해서 주성분과 원래 변수(용어)와의 상관계수를 이용한다. 용어가 주성분에 기여하는 정도 또는 상관 정도를 확인하기 위해서 주성분 적재 계수보다는 상관계수를 이용하며, 용어를 추출하는 데 있어서 최적의 상관계수 임계치를 제안하고자 한다.

본 논문의 구성은 다음과 같다. 제 2장에서는 주성분 분석과 선택된 주성분을 이용하여 중요 용어 추출하는 방법에 대해 설명하며, 제 3장에서는 최적의 임계치를 찾기 위한 여러 가지 실험 결과에 대해 언급한다. 그리고, 제 4장에서는 간단한 결론을 맺는다.

2. 주성분에 의한 중요 용어 추출

2.1 주성분 분석의 정의 및 응용

텍스트 문서는 각 문장들의 집합이며, 각 문장들은 단어들의 모임으로 볼 수 있다. 이 때, 각 문장은 벡터로 표현될 수 있는데, 문서에서 2번 이상 출현한 명사(성분)로 구성되며, 그 성분의 값은 각 문장에서 출현한 횟수로 구성될 수 있다. 여기서 성분들은 변량(variate or variable)으로 설명될 수 있다.

따라서, $p-$변량 원래 문서에서 얻은 크기 n개의 랜덤표본벡터를 $X=(X_1,X_2,\cdots,X_p)'$라 할 때 이로부터 계산된 공분산행렬 S가 구체적인 분석 대상이 된다. 크기 $p\times p$ 행렬 S의 고유값(eigenvalue)을 ℓ, 그리고, 이 고유값에 대응되는 고유벡터(eigenvector)를 g 라 하자. S의 최대 p개의 고유값과 고유벡터의 쌍 (ℓ,g) 를 ℓ의 내림차순으로 배열한 것을 $(\ell_1,g_1),(\ell_2,g_2),\cdots,(\ell_k,g_k),\cdots,(\ell_p,g_p)$로 표기할 때, 이들은 다음 관계를 만족한다.

$$Sg_k = \ell_k g_k, \quad k=1,2,\cdots,p \tag{1}$$

$$\ell_1 \geq \ell_2 \geq \cdots \geq \ell_k \geq \cdots \geq \ell_p$$

$$g_k = (g_{1k},g_{2k},\cdots,g_{jk},\cdots,g_{pk})'$$

이 때 k 번째 표본 주성분 Y_k, $k=1,2,\cdots,p$는 S의 k번째 고유값 ℓ_k의 짝이 되는 고유벡터 g_k의 p개의 원소를 가중계수로 하는 원래 변수들과의 선형 결합으로 다음과 같이 정의된다.

$$Y_k = g_k'X = g_{1k}X_1+g_{2k}X_2+\cdots+g_{jk}X_j+\cdots+g_{pk}X_p \tag{2}$$

이런 p개의 주성분들로 이루어진 주성분 벡터 $Y=(Y_1, Y_2,\cdots,Y_k,Y_p)'$는 각각 원래 변수벡터와 S의 고유벡터와의 선형결합이다. 또한, 각 표본 주성분은 서로 독립적(무상관)이며, 각 표본 주성분의 분산은 S의 고유값이 된다.

비슷하게, 공분산행렬 S대신 상관행렬 R을 이용하여 최대 p개의 고유값과 고유벡터를 구할 수 있다. 그것들의 특성은 위의 식 (1)과 동일하지만, 공분산행렬로부터 구해진 고유값과 고유벡터와는 다르다. 여기서 상관행렬 R은 X의 표준화된 랜덤벡터 $Z=(Z_1,Z_2,\cdots,Z_p)'$에 대한 공분산행렬로부터 구할 수 있다.

공분산행렬로부터 구해진 주성분 Y_k는 분산이 최대인 변수에 의존적이지만, 상관행렬로부터 구해진 주성분은 그렇지 않다는 특성이 있다. 예를 들어, 두 개의 용어로만 구성된 문서에서 용어 X_1은 임의의 두 개 문장에서 각각 1번, 3번 발생했고, 다른 용어 X_2는 각각 1번, 9번 출현했다고 하자. 그러면, X_1과 X_2의 분산은 각각 2와 32가 되며, 결과적으로 공분산행렬로부터 얻어진 가장 중요한 첫 번째 주성분 Y_1은 X_2에 의존될 가능성이 크다. 왜냐하면, X_2의 분산이 X_1의 분산보다 크기 때문이다. 하지만, 각 용어의 분산을 동일하게 1로 만든 상관행렬로 구해진 주성분은 분산이 더 큰 용어에 의존되지 않을 가능성이 높다. 공분산행렬로부터 구해진 첫 번째 주성분은 $Y_1=0.243X_1+0.970X_2$이며, Y_1은 X_2에 크게 의존적이다. 왜냐하면, X_2의 계수가 X_1보다 상대적으로 큰 값이기 때문이다. 결국, 공분산행렬로부터 구해진 주성분은 분산이 큰 용어 즉, 많이 분포되어 있는 용어(고빈도 용어)에 지배적인 영향을 받게 될 가능성이 높다. 하지만, 상관행렬로 구해진 첫 번째 주성분은 $Y_1=0.707Z_1+0.707Z_2$와 같으며, Y_1에 대한 두 용어의 기여도가 동등하다. 이와 같이, 상관행렬로부터 구해진 주성분을 이용함으로써, 고빈도 용어의 선택 가능성을 완화함과 동시에 저빈도 용어에 대한 선택 가능성을 증가시킬 수 있다.

일반적으로 첫 m개의 주성분들 $(Y_1,Y_2,\cdots,Y_m)$, $m\leq p$에 설명되는 부분은

$$\text{누적 비율} = \begin{cases} (\ell_1+\ell_2+\cdots+\ell_m)/(\ell_1+\ell_2+\cdots+\ell_p) \\ \quad (\text{공분산 행렬 } S \text{ 사용시}) \\ (\ell_1+\ell_2+\cdots+\ell_m)/p \\ \quad (\text{상관 행렬 } R \text{ 사용시}) \end{cases} \tag{3}$$

가 된다. 따라서, 만약 첫 m개의 주성분들에 설명되는 부분이 예를 들어 80~90%를 차지한다면 p보다 작은 m개의 주성분들을 이용하더라도 변이(variability)에 대한 정보의 손실은 크지 않을 것이며, 보유할 주성분의 개수를 첫 m개로 한정할 수 있다.

보유할 주성분의 개수가 결정되었다면 각 주성분이 어떤 의미를 가지고 있는지를 알 필요가 있다. [9]에서는 식 (2)와 같은 선형결합에서 가중계수, g_{jk}가 0.5 이상이거나 최대치일 경우의 용어들로 주성분 Y_k를 설명하고자 하였다. 하지만, 본 논문에서는 주성분과 용어간의 상관계수를 이용하여 선택된 주성분을 설명하고자 한다. 다음의 식 (4)와 같이

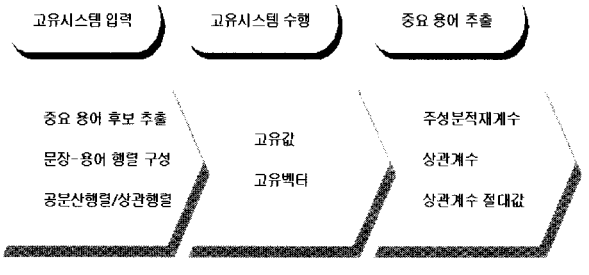
주성분 Y_k와 용어 X_j와의 상관계수는 (주성분적재계수×$\sqrt{\text{고유값}}$)/$\sqrt{\text{용어의분산}}$으로 구할 수 있다. 이때, $\sqrt{s_{jj}}$는 표준화된 변수 Z_j에 대해서는 항상 1이다.

$$\rho_{Y_k,X_j} = g_{jk}\sqrt{\ell_k} \,/\, \sqrt{s_{jj}} \qquad (4)$$

2.2 중요 용어 추출

다음의 (그림 1)은 중요 용어 추출의 각 단계를 보이고 있다. 고유 시스템의 입력은 용어간의 공분산 또는 상관행렬이며, 고유시스템의 결과로 고유값과 고유벡터를 획득할 수 있다. 고유값-고유벡터에 기반하여 주성분 적재 계수 또는 주성분과 용어와의 상관계수를 이용하여 중요 용어를 추출한다. 이번 절에서는 이러한 과정을 예제를 이용하여 기술한다.

<표 1>은 "P기업의 수뢰-축재와 탈세혐의에 대한 입장"이라는 내용의 신문 기사에서 추출한 중요 용어 후보 리스트이다. 중요 용어 후보로는 문서에서 2번 이상 출현한 명사로만 한정하였으며, 예제 문서의 경우 총 15개가 추출되었다.



(그림 1) 중요 용어 추출의 각 단계

〈표 1〉 중요 용어 후보 리스트

용어	문서 내 발생 횟수	표기	용어	문서 내 발생 횟수	표기
간부	8	X_1	부분	2	X_9
개인	2	X_2	사과문	2	X_{10}
검찰수사	2	X_3	사람	3	X_{11}
계열사	2	X_4	수뇌부	2	X_{12}
관계자	2	X_5	직원	2	X_{13}
국세청	5	X_6	P기업	24	X_{14}
A모씨	6	X_7	현직	4	X_{15}
방침	2	X_8			

그러면, 각 문장은 15차원의 행벡터로 구성할 수 있는데, 각 성분의 값은 용어가 문장에서 출현한 횟수가 된다. 예를 들어, 첫 번째 문장 벡터는 (1,0,0,…,1,0)가 될 수 있다. 이와 같이, 예제 문서는 25*15 행렬로 구성될 수 있으며, 여기서 행은 각 문장을 나타내며, 열은 <표 1>의 용어를 나타낸다. 원 문서는 30개의 문장으로 구성되어 30*15 행렬로 표현되어야 하지만, 행의 모든 값이 0인 경우에는 그 행을 제거하였기 때문에 25*15 행렬이 된 것이다.

공분산행렬 또는 상관행렬은 이러한 문장-용어 행렬을 기반으로 용어와 용어간의 공분산(covariance) 또는 상관계수(correlation coefficient)를 계산함으로써 구할 수 있다.

<표 2>는 총 15개의 주성분 중에서 첫 2개의 주성분 (Y_1, Y_2)을 나타내고 있다. 주성분 적재 계수는 식 (2)의 g_{jk}

〈표 2〉 첫 2개 주성분에 대한 주성분 적재 계수와 상관계수 (대상행렬: 공분산행렬, 상관행렬)

용어	공분산행렬						상관행렬					
	주성분적재계수		상관계수		|상관계수|		주성분적재계수		상관계수		|상관계수|	
	Y_1	Y_2	Y_1	Y_2	Y_1	Y_2		Y_1	Y_2	Y_1	Y_2	Y_1	Y_2
X_1	**0.832**	-0.181	**0.979**	-0.145	**0.979**	0.145		**0.514**	-0.072	**0.870**	-0.105	**0.870**	0.105
X_2	-0.057	-0.046	-0.167	-0.093	0.167	0.093		-0.094	-0.151	-0.159	-0.222	0.159	0.222
X_3	-0.051	0.086	-0.149	0.171	0.149	0.171		-0.155	0.368	-0.263	**0.542**	0.263	**0.542**
X_4	-0.032	**0.211**	-0.093	0.422	0.093	0.422		-0.038	0.376	-0.064	**0.553**	0.064	**0.553**
X_5	-0.053	0.119	-0.155	0.238	0.155	0.238		-0.164	**0.495**	-0.278	**0.728**	0.278	**0.728**
X_6	-0.065	0.008	-0.130	0.010	0.130	0.010		-0.182	-0.359	-0.309	-0.528	0.309	**0.528**
X_7	-0.143	-0.609	-0.267	-0.773	0.267	**0.773**		-0.172	-0.268	-0.291	-0.394	0.291	0.394
X_8	-0.070	-0.036	-0.207	-0.072	0.207	0.072		-0.256	-0.332	-0.434	-0.489	0.434	0.489
X_9	-0.052	0.061	-0.152	0.122	0.152	0.122		-0.139	-0.233	-0.235	-0.342	0.235	0.342
X_{10}	-0.070	-0.015	-0.204	-0.030	0.204	0.030		-0.269	0.118	-0.456	0.173	0.456	0.173
X_{11}	0.149	-0.296	0.365	-0.493	0.365	0.493		0.264	-0.120	0.447	-0.177	0.447	0.177
X_{12}	-0.042	0.066	-0.123	0.132	0.123	0.132		-0.023	0.059	-0.039	0.086	0.039	0.086
X_{13}	0.250	-0.003	**0.734**	-0.005	**0.734**	0.005		0.423	-0.007	**0.717**	-0.011	**0.717**	0.011
X_{14}	-0.194	-0.656	-0.346	-0.798	0.346	**0.798**		-0.189	-0.219	-0.320	-0.322	0.320	0.322
X_{15}	0.369	-0.004	**0.801**	-0.006	**0.801**	0.006		0.412	-0.066	**0.698**	-0.097	**0.698**	0.097
고유값	0.660	0.306						2.870	2.161				

을 의미하며, 원래 변수와의 선형결합에서 가중계수이다. 그리고, 고유값은 식 (1)과 같이 첫 번째 주성분(Y_1)의 고유값이 두 번째 주성분(Y_2)보다 더 크다. 한편, 상관계수는 각 주성분과 원래 변수(용어)와의 상관 정도를 나타내는데, 식 (4)에 의해 계산될 수 있다. |상관계수|는 구해진 상관계수의 절대값을 의미한다.

<표 2>에서 볼드체와 밑줄로 표시된 숫자에 대응되는 용어가 중요한 용어임을 나타낸다. 주성분 적재 계수 방법[9]은 계수가 0.5이상이거나 최고치인 용어를 선택하며, 상관계수 방법은 상관계수가 0.5이상인 경우를 선택한다. 주성분 분석의 대상 행렬과 방법에 따라 추출된 용어에 차이가 있음을 알 수 있다. 특히, 공분산행렬의 |상관계수| 방법에서, 두 번째 주성분 Y_2에 의해 X_7과 X_{14}가 선택되었는데, 이는 "P기업"의 대표 "A모씨"가 함께 추출된 경우이다. 비슷하게, 상관행렬의 |상관계수| 방법에서는 X_3,X_6등이 선택되었는데, "검찰수사"와 "국세청"은 예제 문서를 대표할 만한 용어라 볼 수 있다.

이제, 대상 행렬에 따른 용어의 가중치 변화에 대해 언급한다. 예제 문서에서 X_{13}은 2번 출현했고, X_{14}는 24번 나타났다. 그리고, 첫 번째 주성분의 고유값이 가장 크며, 고유값은 분산에 해당한다. 분산은 분포에 대한 설명력이라 볼 수 있다. 그렇다면, Y_1에 대한 가중계수 즉, 주성분적재계수를 공분산행렬과 상관행렬에 대해서 비교해도 무방하다. 주성분적재계수의 차이는 $X_{13}=0.423-0.250=0.173$이고, $X_{14}=-0.189-(-0.194)=0.005$이다. 공분산행렬 대신에 상관행렬을 이용함으로써 출현 빈도가 낮은 X_{13}은 0.173만큼 가중치가 증가 되었고, 반면에 X_{14}는 0.005만큼만 증가되었다. 다시 말하면, 출현 빈도가 낮은 X_{13}은 예제 문서에서 가장 많이 발생한 X_{14}보다 34.6배의 가중치를 가지게 된 것이다. 결국, 제안한 방법은 출현 빈도와 용어간의 상관 관계를 동시에 이용하여 문서 내의 용어를 추출할 수 있다.

3. 실험 및 평가

3.1 실험 자료 및 방법

중요 용어 추출 방법에 대한 평가를 위해, 추출된 중요 용어의 출현 여부에 기반한 문장 추출[2] 실험을 하였다. 실험에 사용된 테스트 컬렉션은 KISTI에서 제공되었던 문서(신문 기사) 집합이며, 총 1,000여건의 문서 중 제공된 30% 추출 요약의 길이가 원문의 30% 이상인 283건의 문서를 사용하였다. 여기서, 제공된 30% 요약문에 포함된 문장들을 정답 문장으로 간주한다.

평가 척도로는 다음과 같이 정확률(Precision), 재현율(Recall) 그리고 F-척도(F-measure)를 사용하였다.

$$\text{정확률} = \frac{\text{각 방법이 추출한 정확한 요약문장수}}{\text{각 방법이 추출한 요약문장수}}$$

$$\text{재현율} = \frac{\text{각 방법이 추출한 정확한 요약문장수}}{\text{제공된 30\% 수동 요약문장수}}$$

$$F-\text{척도} = \frac{2\times \text{정확률}\times \text{재현율}}{\text{정확률}+\text{재현율}}$$

[9]에서는 공분산행렬로부터 구해진 주성분들의 고유값 누적비율이 90% 이상인 시점의 첫 주성분까지 추출하였으며, 추출된 주성분들의 주성분 적재 계수가 0.5이상인 변수들을 중요 용어로 선택을 하였다. 하지만, 본 논문에서는 아래와 같은 방법으로 확장하여 실험 집단에 대한 최적의 보유 주성분 수와 상관계수를 찾고자 한다.

- 주성분 분석의 분석 대상이 다를 경우(즉, 공분산행렬과 상관행렬) 성능 비교
- 보유 주성분수를 선택함에 있어서 누적 비율의 결정 문제
- 중요 용어를 선택함에 있어서 선택된 주성분과 용어와의 상관계수 임계치 결정 문제

결과적으로, 사용된 실험 집단에 대해서는 공분산행렬보다 상관행렬로부터 구해진 주성분들이 보다 효과적임을 알 수 있었다. 또한, 첫 여섯 개의 주성분까지를 사용하는 경우와 상관계수 임계치를 |0.4|로 이용한 경우가 최적이었다.

3.2 분석 대상 행렬에 따른 비교

<표 3>과 <표 4>는 고유값의 누적비율이 0.9이상이 되는 시점의 첫 주성분까지를 선택한 경우이며, 중요 용어를 선택하기 위해 주성분 적재 계수를 이용하는 방법("계수")[9], 선택된 주성분과 변수사이의 상관계수를 이용하는 방법("상관계수") 그리고 그 상관계수의 절대값을 이용하는 방법("|상관계수|")을 나타내고 있다. 그리고, 두 번째 행은 임계치를 나타낸다. 예를 들어, |상관계수|와 0.4는 상관계수의 절대값이 0.4 이상일 경우에 중요 용어로 선택되었음을 나타낸다. 여기서, 0.4를 기준으로 선택한 이유는 상관계수의

〈표 3〉 분석 대상 행렬이 공분산행렬일 경우(단, 누적 비율 >= 0.9)

	계수	상관계수						|상관계수|				
	0.5	0.4	0.5	0.6	0.7	0.8	0.9	0.4	0.5	0.6	0.7	0.8	0.9
평균 정확률	0.320	0.355	0.321	0.317	0.352	0.368	0.439	0.503	0.467	0.380	0.322	0.322	0.380
평균 재현율	0.283	0.313	0.284	0.282	0.311	0.325	0.390	0.448	0.418	0.340	0.285	0.284	0.333
F-척도	0.300	0.333	0.302	0.298	0.330	0.345	0.413	0.474	0.441	0.359	0.302	0.302	0.355
평균 중요 용어수	5.0	9.5	7.6	5.5	3.6	1.9	0.8	13.5	12.6	10.0	6.6	3.6	1.4

절대값이 0.4 이상인 경우에는 두 변수 간에 상당한 관련이 있다고 판단될 수 있기 때문이다[4].

<표 3>에서, F-척도에 의하면, 기존의 주성분적재계수를 이용하는 방법보다 선택된 주성분과 변수 간의 상관계수를 이용하는 방법이 중요 문장 추출 성능이 우수함을 알 수 있다. 특히, 상관계수의 절대값을 이용하고 그 임계치가 0.4인 경우가 가장 우수하다. 그런데, 이 경우는 실험 집단의 평균 용어수(13.6)와 추출된 평균 중요 용어수(13.5)가 거의 같은 경우에서 발생하였다. 한편, 임계치가 0.9일 경우에, 추출된 평균 중요 용어수는 한 두개 뿐인데, 그 성능이 대체적으로 좋음을 알 수 있다. 이는 문장들 간의 점수가 동일할 경우에는 원문의 특성을 반영하여, 앞선 문장을 추출한 결과라고 판단된다.

<표 4>에서는 <표 3>에서와 거의 비슷한 양상을 보이고 있다. 하지만, 공분산행렬을 이용한 경우인 <표 3>보다 성능이 전반적으로 우수하다. 특히, 상관계수의 절대값을 이용하고 그 임계치를 0.4로 사용할 경우에 가장 높은 성능을 나타내고 있다. 여기서, 상관계수 절대값은 용어 사이의 긍정적 상관(positive correlation)과 부정적 상관(negative correlation) 모두를 이용함을 의미한다.

3.3 보유 주성분수에 따른 비교

이번 절에서는 가장 좋은 성능을 보인 방법(|상관계수| >= 0.4)에 대해, 보유 주성분수에 따른 성능을 비교하고자 한다. 보유할 주성분의 개수를 결정하는 방법은 식 (3)과 같이 고유값의 누적 비율을 이용하는 방법과 식 (5)와 같이 큰 고유값을 갖는 주성분을 보유하는 경우가 있다[6-8].

$$\begin{aligned} &\text{공분산행렬사용시} \rightarrow \ell_i > \bar{\ell}\,(Kaiser),\ \ell_i > 0.7\bar{\ell}\,(Joliffe) \\ &\text{상관행렬사용시} \rightarrow \ell_i > 1\,(Kaiser),\ \ell_i > 0.7\,(Joliffe) \end{aligned} \tag{5}$$

<표 5>에서 F-척도에 의해 상위 두개의 방법 즉, 누적비율이 0.9이상인 방법과 Joliffe 방법에서 평균 주성분 보유 개수는 약 6개임을 알 수 있다. 비슷하게, 분석 대상 행렬이 상관행렬일 경우에는 평균 주성분 보유 개수가 약 5~6개임을 <표 6>에서 나타내고 있다. 이제, 본 논문의 실험 집단에서 보유할 적당할 주성분은 6개임을 가정하고, 첫 번째 주성분만, 첫 두개의 주성분까지 등과 같이 실험한 결과를 <표 7>에서 보여주고 있다. 대상 행렬이 공분산행렬이든지 상관행렬이든지 두 경우 모두에서, 첫 3개에서 첫 6개까지

〈표 4〉 분석 대상 행렬이 상관행렬일 경우(누적비율 >= 0.9)

	계수	상관계수						|상관계수|				
	0.5	0.4	0.5	0.6	0.7	0.8	0.9	0.4	0.5	0.6	0.7	0.8	0.9
평균 정확률	0.314	0.362	0.345	0.326	0.331	0.374	0.432	0.512	0.485	0.404	0.332	0.337	0.406
평균 재현율	0.278	0.318	0.303	0.284	0.290	0.329	0.382	0.456	0.433	0.357	0.294	0.296	0.359
F-척도	0.295	0.339	0.323	0.304	0.309	0.350	0.405	0.483	0.457	0.379	0.312	0.315	0.381
평균 중요 용어수	5.6	9.7	8.0	6.0	3.7	1.9	0.7	13.6	13.0	10.5	6.6	3.4	1.1

〈표 5〉 주성분 보유 개수에 따른 성능 비교 (분석대상 : 공분산행렬, |상관계수| >= 0.4)

	누적비율						Kaiser	Joliffe
	0.4	0.5	0.6	0.7	0.8	0.9		
평균 정확률	0.324	0.350	0.401	0.447	0.483	0.503	0.481	0.504
평균 재현율	0.288	0.310	0.356	0.395	0.431	0.448	0.433	0.450
F-척도	0.305	0.328	0.377	0.419	0.455	0.474	0.456	0.475
평균 중요 용어수	8.8	10.2	11.3	12.2	13.0	9.5	12.9	13.3
평균 보유 개수	1.6	2.1	2.7	3.3	4.3	5.7	4.3	5.3

〈표 6〉 주성분 보유 개수에 따른 성능 비교(분석대상 : 상관행렬, |상관계수| >= 0.4)

	누적비율						Kaiser	Joliffe
	0.4	0.5	0.6	0.7	0.8	0.9		
평균 정확률	0.370	0.424	0.469	0.491	0.506	0.512	0.514	0.512
평균 재현율	0.328	0.375	0.417	0.436	0.450	0.456	0.458	0.456
F-척도	0.348	0.398	0.441	0.462	0.476	0.483	0.484	0.483
평균 중요 용어수	11.0	12.1	12.8	13.3	13.5	13.6	13.6	13.6
평균 보유 개수	2.1	2.6	3.2	3.9	4.8	6.2	4.9	5.8

〈표 7〉 주성분 보유 개수에 따른 성능 비교(주성분 개수 : 첫 m개, |상관계수| >= 0.4)

	공분산행렬						상관행렬					
주성분 개수	1	2	3	4	5	6	1	2	3	4	5	6
평균 정확률	0.323	0.360	0.434	0.473	0.490	0.497	0.315	0.396	0.466	0.491	0.506	0.505
평균 재현율	0.288	0.322	0.388	0.424	0.439	0.444	0.276	0.356	0.417	0.438	0.451	0.451
F-척도	0.305	0.340	0.410	0.447	0.463	0.469	0.294	0.375	0.440	0.463	0.477	0.476
평균 중요 용어수	6.6	9.5	11.1	12.0	12.5	12.9	7.7	10.6	12.0	12.6	13.1	13.3

의 주성분을 선택했을 때 그 성능이 40% 이상이 됨을 <표 7>에서 확인할 수 있다. <부록>에 첫 6개까지의 주성분과 |상관계수| >= 0.4라는 조건하에 하나의 문서에 대한 실험 결과를 원문과 함께 수록하였다.

결과적으로 본 논문에서 사용한 신문 기사와 같은 실험 집단에서는 주성분의 개수는 첫 3~6개까지를 보유함이 적당하다고 할 수 있다. 또한, 보유한 주성분과의 관련이 있는 변수 즉, 용어를 선택할 경우에는 주성분 적재 계수보다는 상관계수를, 상관계수보다는 상관계수의 절대값을 이용함이 타당하며, 그 임계치는 0.4가 가장 적당하다고 할 수 있다. 그리고, 주성분분석의 대상 행렬을 공분산행렬 보다는 상관행렬을 이용하는 경우가 전반적으로 그 성능이 우수함을 알 수 있었다.

4. 결 론

본 논문에서는 용어간의 상관관계를 이용하여 첫 몇 개의 주성분을 선택할 수 있는 주성분 분석을 응용하였다. 주성분 분석은 용어 사이에 완벽한 독립을 인정하지 않는 기법이며, 이러한 특성은 자연 언어의 특성 중의 하나를 잘 이용한다고 볼 수 있다. 하지만, 거의 모든 통계적 기법이 그러하듯이 용어의 의미적인 정보는 이용하지 못하고 있다.

본 논문에서는 문서 특히, 신문 기사를 대상으로 중요 용어를 추출할 경우, 보유해야 할 주성분의 개수와 중요 용어를 추출하기 위한 상관계수의 적당한 임계치를 제안하였다. 그 임계치는 첫 6개와 |0.4|이다. 또한, 용어의 분포에 대해 표준화 과정을 거치지 않는 경우보다, 평균을 0으로 분산을 1로 표준화하여 구해진 주성분을 이용하는 것이 보다 적당함을, 실험 결과 알 수 있었다. 이는, 문서에서 중요 용어를 추출하는 응용에서도 척도 불변(scale invariant)은 중요한 조건임을 나타낸다. 본 논문에서, 이러한 척도 불변은 고빈도 용어에 대한 선택 가능성을 낮춤과 동시에, 저빈도 용어에 대한 선택 가능성을 다소 증가시키는 효과에 응용되었다.

참 고 문 헌

[1] D. C. Manning and H. Schutze, "Foundations of Statistical Natural Language Processing," Cambridge, MA : The MIT Press, 1999.

[2] I. Mani, "Automatic Summarization," Amsterdam : John Benjamins Publishing Company, 2001.

[3] W. H. Press et al., "Numerical Recipes in C++," Second Ed., New York : Cambridge University Press, 2002.

[4] 강병서, "의사결정을 위한 현대통계학", 무역경영사, 2004.

[5] 김재훈, "정보 추출의 기술 현황", 한국정보과학회 학회지, 제22권, 제4호, pp.35-46, 2004.

[6] 김기영, 전명식, "다변량 통계자료분석", 자유아카데미, 1999.

[7] 노형진, "다변량분석 이론과 실제", 형설출판사, 2005.

[8] 손영숙, "주성분분석", http://stat.chonnam.ac.kr/~ysson/A_MULTI00-2/MULTI01-2/teaching_contents.htm

[9] 이창범, 김민수, 백장선, 박혁로, "주성분 분석과 비정칙치 분해를 이용한 문서 요약", 정보처리학회논문지, 제10-B권, 제7호, pp.725-734, 2003.

[10] 한학용, "패턴인식개론-MATLAB 실습을 통한 입체적 학습", 한빛미디어, 2005.

부 록

A. "여권 자동조회 「핸디아이」 개발/2초만에 암호 해독 첨단장치" 제목의 신문 기사

① 여권을 자동조회하는 첨단시스템이 국내에서 처음 개발됐다.

② 한국과학기술원 출신 석-박사등 고급인력 36명으로 세워진 모험기업 「핸디소프트」사는 최근 성능이 우수하고 가격이 저렴한 여권자동인식장치 「핸디아이」를 개발하고 7월 판매를 목표로 마무리작업에 한창이다.

③ 국제민간항공기구(ICAO)는 지난해 공항서비스를 개선하기위해 세계표준규격 여권을 채택토록 각국에 권고했다.

④ 이에따라 미국 일본 등 16개국이 이미 세계표준여권을 채택하고 있으며 우리나라도 올해말부터 시험적용한 다음 내년부터 확대적용할 방침이다.

⑤ 세계표준여권이 나오면 이를 자동인식하는 장치개발이 가능해진다.

.....(중략)

⑯ 한편 핸디소프트는 윈도우에서 돌아가는 강력한 한글 워드프로세서 소프트웨어인 「핸디워드 아리랑」도 개발, 6월말 한국종합전시장에서 선을 보일 예정이다.

⑰ 핸디워드 아리랑은 문서작성기능과 함께 자료관리 전

자우편 전자게시판 전자문서함 등의 기능을 한꺼번에 갖춘 강력한 소프트웨어이다.

⑱ 이 소프트웨어는 문서작성중 자동편집이나 다단편집은 물론이고 스프레드시트기능도 가졌으며, 각종 통계를 직접 데이터베이스에 전송하는 기능을 추가할 수 있다.

⑲ 아리랑은 자료정리프로그램과 전자우편함기능도 갖췄으며 다른 워드프로세서로 작성한 문서와도 폭넓은 호환성을 지닌 것이 장점이다.

⑳ 안사장은 『하반기부터 국내에 핸디워드아리랑을 9만9천원에 판매할 예정이지만 진짜 목표는 미국시장이며 유니코드로 제작하면 전세계시장도 겨냥해볼 만하다』고 야심차게말했다.

B. 제공된 30% 수동 추출 요약

① 여권을 자동조회하는 첨단시스템이 국내에서 처음 개발됐다.

② 한국과학기술원 출신 석-박사등 고급인력 36명으로 세워진 모험기업 「핸디소프트」사는 최근 성능이 우수하고 가격이 저렴한 여권자동인식장치 「핸디아이」를 개발하고 7월 판매를 목표로 마무리작업에 한창이다.

③ 국제민간항공기구(ICAO)는 지난해 공항서비스를 개선하기위해 세계표준규격 여권을 채택토록 각국에 권고했다.

④ 이에따라 미국 일본 등 16개국이 이미 세계표준여권을 채택하고 있으며 우리나라도 올해말부터 시험적용한 다음 내년부터 확대적용할 방침이다.

⑤ 세계표준여권이 나오면 이를 자동인식하는 장치개발이 가능해진다.

⑥ 일본 마쓰시타와 미국 AIT사가 핸디소프트사에 앞서 각각 여권자동인식장치를 만들어 판매중인데 대당 가격이 무려 1만8천달러와 1만2천달러나 된다.

⑦ 그러나 핸디소프트가 만든 핸드아이는 일제나 미제에 비해 성능이 우수하면서도 가격은 훨씬 저렴하다.

C. 중요 용어 추출 결과 (첫 6개의 주성분, |상관계수|≥0.4)

공분산행렬 이용시(총 20개 용어)		상관행렬 이용시(총 24개 용어)	
		공항	아이
공항	여권	국내	여권
국내	워드	모니터	워드
모니터	웨어	목표	웨어
목표	자료	문서	일본
문서	전자	미국	자료
성능	제조원가	성능	전자
세계	카메라	세계	제조원가
시스템	프로세서	세계표준	카메라
아리랑	한국	숫자	프로세서
아이	핸디	시스템	한국
		아리랑	핸디

D. 30% 자동 추출 요약(밑줄로 표시된 문장은 30% 수동 추출 요약과 일치한 문장임)

D-1. 공분산행렬을 이용한 결과

① 여권을 자동조회하는 첨단시스템이 국내에서 처음 개발됐다.

② 한국과학기술원 출신 석-박사등 고급인력 36명으로 세워진 모험기업 「핸디소프트」사는 최근 성능이 우수하고 가격이 저렴한 여권자동인식장치 「핸디아이」를 개발하고 7월 판매를 목표로 마무리작업에 한창이다.

③ 국제민간항공기구(ICAO)는 지난해 공항서비스를 개선하기위해 세계표준규격 여권을 채택토록 각국에 권고했다.

④ 그러나 핸디소프트가 만든 핸드아이는 일제나 미제에 비해 성능이 우수하면서도 가격은 훨씬 저렴하다.

⑤ 우리 법무부도 최근 이 장치를 수입하려던 계획을 바꿔 핸디소프트에 1천대를 주문했다고 한다.

⑥ 핸디아이가 채택한 영상입력기술은 이미 수년전 국내 모험기업인 크로스엔지니어링사가 기업화한 지문인식 시스템을 응용한 것이어서 인식도가 뛰어나다.

D-2. 상관행렬을 이용한 결과

① 여권을 자동조회하는 첨단시스템이 국내에서 처음 개발됐다.

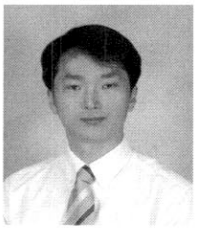
② 한국과학기술원 출신 석-박사등 고급인력 36명으로 세워진 모험기업 「핸디소프트」사는 최근 성능이 우수하고 가격이 저렴한 여권자동인식장치 「핸디아이」를 개발하고 7월 판매를 목표로 마무리작업에 한창이다.

③ 국제민간항공기구(ICAO)는 지난해 공항서비스를 개선하기위해 세계표준규격 여권을 채택토록 각국에 권고했다.

④ 이에따라 미국 일본 등 16개국이 이미 세계표준여권을 채택하고 있으며 우리나라도 올해말부터 시험적용한 다음 내년부터 확대적용할 방침이다.

⑤ 세계표준여권이 나오면 이를 자동인식하는 장치개발이 가능해진다.

⑥ 일본 마쓰시타와 미국 AIT사가 핸디소프트사에 앞서 각각 여권자동인식장치를 만들어 판매중인데 대당 가격이 무려 1만8천달러와 1만2천달러나 된다.



이 창 범

e-mail : chblee@empal.com

1995년 전남대학교 전산학과(학사)

2001년 전남대학교 전산학과(이학석사)

2005년 전남대학교 전산학과(이학박사)

1995년~1999년 대우정보시스템(주)

2005년~현재 울산대학교 컴퓨터정보통신공학부/DMITRC 연구교수

관심분야 : 정보검색, 자연어처리, 문서요약 등

옥 철 영

e-mail : okcy@ulsan.ac.kr
1982년 서울대학교 컴퓨터공학과(학사)
1984년 서울대학교 컴퓨터공학과 (공학석사)
1993년 서울대학교 컴퓨터공학과 (공학박사)
1994년~1996년 러시아 TOMSK 공과대학 교환교수
1996년~1997년 영국 GLASGOW대학교 객원교수
1984년~현재 울산대학교 컴퓨터정보통신공학부 교수
관심분야 : 한국어정보처리, 지식베이스, 기계학습, 온톨로지 등

박 혁 로

e-mail : hyukro@chonnam.ac.kr
1987년 서울대학교 전산학과(학사)
1989년 한국과학기술원 전산학과 (공학석사)
1997년 한국과학기술원 전산학과 (공학박사)
1994년~1998년 한국과학기술정보연구원 연구원
2002년~2002년 University of Maryland UMIACS Post Doc.
1999년~현재 전남대학교 전자컴퓨터정보통신공학부 교수
관심분야 : 정보검색, 자연어처리, 데이터베이스 등