주성분 분석을 이용한 문서 주제어 추출
(Document Thematic words Extraction using Principal Component Analysis)

이 창 범 † 김 민 수 †† 이 기 호 ††† 이 귀 상 †††† 박 혁 로 ††††
(Chang-Beom Lee) (Min-Soo Kim) (Ki-Ho Lee) (Guee-Sang Lee) (Hyuk-Ro Park)

요 약 본 논문에서는 문서의 내용을 대표할 수 있는 주제어를 추출하는데 있어 다변량 통계 분석 기법 중의 하나인 주성분 분석을 이용하는 모델을 제안한다. 제안한 모델은 고유값과 고유벡터를 이용하여 문서 자체내의 단어의 흐름을 파악한 후 주제어를 추출하는 방법이다.

제안한 모델을 문서 요약에 적용하여 그 성능을 평가하였다. 신문기사를 대상으로 실험한 결과 제안한 모델이 단어의 출현 빈도를 고려하는 방법, 시소러스를 이용하는 방법 모두에 비해 더 좋은 성능을 보였다. 제안한 모델은 정보검색, 정보추출, 문서요약 등에 이용될 수 있으리라 기대된다.

키워드 : 주제어 추출, 주성분 분석, 문서 요약

Abstract In this paper, We propose a document thematic words extraction by using principal component analysis(PCA) which is one of the multivariate statistical methods. The proposed PCA model understands the flow of words in the document by using an eigenvalue and an eigenvector, and extracts thematic words.

The proposed model is estimated by applying to document summarization. Experimental results using newspaper articles show that the proposed model is superior to the model using either word frequency or information retrieval thesaurus. We expect that the proposed model can be applied to information retrieval, information extraction and document summarization.

Key words : thematic word extraction, principal component analysis, document summarization

1. 서 론

색인이란 어떤 문서에 대해 그 문서의 전체적 내용을 나타내거나, 그 문서를 다른 문서들로부터 구별할 수 있도록 그 문서의 선택 단서가 되는 단어 또는 단어 구 등을 추출하는 것을 말한다[1]. 색인 방법은 크게 자동 색인과 수동 색인으로 나눌 수 있는데, 자동 색인 방법에는 크게 단어의 빈도를 계산하여 출현 빈도가 많은 순으로 색인어를 정하는 통계적인 방법과 형태소 해석, 구문 해석, 의미 해석 등의 다양한 기법을 이용하는 언어학적인 기법으로 나눌 수 있다[2]. 이러한 색인 방법 중 형태소 해석을 이용한 기법이 많이 이용되는데, 한국어에 적용이 쉽고 구현이 간단하다는 장점이 있다[3].

여기에서, 색인이란 어떤 문서의 내용을 대표하는 단어들 즉, 주제어들을 추출하는 과정이라고 볼 수 있다. 또한, 단순히 형태소 해석만을 통해서 추출된 단어를 주제어로 선택한다면, 문서의 내용과는 관련이 적은 단어들이 채택될 수 있다. 예를 들어, 어떤 문서가 과일에 대해 이야기하고 있다고 하자. 그리고 '사과', '배', '바나나', '나무 상자' 등이 그 문서에서 추출되었다고 한다면, 단순히 형태소 해석만을 이용한다면 문서의 내용과 연관 정도가 약한 '나무 상자'가 포함될 가능성이 크다. 만약, '사과', '배', '바나나' 등의 단어가 서로 연관되어 있다는 사실을 안다면, 주제어로 과일에 관련된 단어만을

· 본 연구는 한국과학재단 목적기초연구(R05-2001-000-01480-0)지원으로 수행되었음.
† 학생회원 : 전남대학교 전산학과
chblee@dal.chonnam.ac.kr
†† 비 회 원 : 전남대학교 BK사업단 post-doc
kimms@chonnam.ac.kr
††† 비 회 원 : 충북과학대학 컴퓨터정보학과 교수
kyiho_lee@yahoo.com
†††† 종신회원 : 전남대학교 전산학과 교수
gslee@chonnam.ac.kr
hyukro@chonnam.ac.kr
논문접수 : 2002년 3월 2일
심사완료 : 2002년 7월 25일

선택할 수 있게 되어 문서의 내용과 연관 정도가 약하다고 볼 수 있는 '나무 상자'라는 단어는 배제할 수 있다.

이에 본 논문에서는 통계적 분석 기법 중의 하나인 주성분 분석(Principal Component Analysis)의 공기 정보를 이용하여 주제어를 추출하는 모델을 제안한다.

제안한 모델은 문서 자체내의 단어의 흐름 또는 유사도를 아이겐시스템(eigensystem)에 정량화 할 수 있는 점을 이용한다. 이를 이용하면 '단어들이 얼마나 자주 그 문서에 나타나는가' 라는 정보와 '단어들이 얼마나 자주 같이 나타나는가' 라는 정보를 함께 정량화 할 수 있다. 이렇게 정량화 된 정보를 이용하여 해당 문서의 주제어를 추출한다. 이것은 다른 도구(시소러스, 공기사전 등)의 정보를 이용하지 않고, 해당 문서 내에서의 단어 발생 빈도와 공기 정보를 토대로 그 문서의 주제어를 추출함을 의미한다.

본 논문의 구성은 다음과 같다. 제2장에서는 주제어 추출에 사용한 다변량 통계 분석 기법 중의 하나인 주성분 분석 대해서 설명한다. 그리고, 제3장에서는 제안한 주제어 추출 방법을, 제4장에서는 실험 및 평가에 대해 기술한다. 마지막으로 제5장에서는 결론을 이야기한다.

2. 주성분 분석의 개요

$p(\geq 2)$개의 확률특징 $X_1, X_2, \ldots, X_p$를 원소로 하는 확률특징벡터 $\boldsymbol{X}$가 평균 벡터 $\bar{\boldsymbol{x}}$와 공분산 행렬 $\boldsymbol{S}$ (p× p matrix)를 갖는다고 하고, 이들을 다음의 기호로 나타내자.

$$\boldsymbol{X}=\begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_p \end{bmatrix} \quad \bar{\boldsymbol{x}}=\begin{bmatrix} \bar{x}_1 \\ \bar{x}_2 \\ \vdots \\ \bar{x}_p \end{bmatrix} \quad \boldsymbol{S}=\begin{bmatrix} s_{11} & s_{12} & \cdots & s_{1p} \\ s_{21} & s_{22} & \cdots & s_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ s_{p1} & s_{p2} & \cdots & s_{pp} \end{bmatrix} \tag{2.1}$$

단 $s_{ik}=\frac{1}{n}\sum_{j=1}^{n}(x_{ij}-\bar{x}_i)(x_{kj}-\bar{x}_k)=s_{ki}$: X_i와 X_k의 공분산

$\bar{x}_i=\frac{1}{n}\sum_{j=1}^{n}X_{ij}$: X_i의 산술 평균이다.

주성분 분석은 원래 특징벡터 $\boldsymbol{X}$를 적절히 선형 변환시켜 그것이 가지는 정보를 가능한 한 많이 보존하는 (최소의 정보손실) 소수 몇 개(m개)의 새로운 인공 특징을 생성함으로서, p 차원 변이를 m 차원으로 축소하여 전체의 특성을 요약하고, 이를 통해서 특징들 간의 다변량 구조를 밝히고자 한다.

이 변환은 $\boldsymbol{X}$의 원소들 간의 상관구조관계를 나타내는 S를 분석대상으로 하며, S는 $\bar{\boldsymbol{x}}$의 값의 변화에 의한 영향을 받지 않는다.

우선 S의 p개의 고유값(eigen value) λ_j들을 크기 순으로 배열하고 각각의 고유값에 대응되는 고유벡터(eigen vector) $\boldsymbol{e}_j$의 짝들을 $(\lambda_1, \boldsymbol{e}_1), (\lambda_2, \boldsymbol{e}_2), \ldots, (\lambda_p, \boldsymbol{e}_p)$라하고, λ_j들을 크기 순으로 배열하면,

$$\boldsymbol{S}\boldsymbol{e}_j=\lambda_j\boldsymbol{e}_j \quad , j=1,2,\ldots,p \tag{2.2}$$

$$\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p$$

와 같은 관계가 있으며, 이를 행렬(matrix) 기호를 이용하여 전체적으로 표현하면 다음과 같다.

$$\boldsymbol{SP}=\boldsymbol{P\Lambda} \quad , \quad \boldsymbol{S}=\boldsymbol{P\Lambda P'} \tag{2.3}$$

여기서 $\boldsymbol{P}$는 p개의 고유벡터 $\boldsymbol{e}_i$들로 구성된 크기 p ×p 직교행렬(orthogonal matrix)이고, $\boldsymbol{\Lambda}$는 λ_i를 i번째 대각원소, 그리고 모든 비대각 원소가 0인 크기 p× p의 대각행렬(diagonal matrix), 그리고 $\boldsymbol{P'}$는 $\boldsymbol{P}$의 전치행렬(transpose matrix)이다. 즉

$$\boldsymbol{P}=(\boldsymbol{e}_1, \boldsymbol{e}_2, \ldots, \boldsymbol{e}_p), \tag{2.4}$$

$$\boldsymbol{\Lambda}=\mathrm{Diagonal}(\lambda_1, \lambda_2, \ldots, \lambda_p),$$

이와 같은 $\boldsymbol{P}$를 이용하여 다음과 같은 $\boldsymbol{X}$의 직교변환을 생각할 때,

$$\boldsymbol{\phi}' = \boldsymbol{P'X} \tag{2.5}$$

이 변화에 의해 새로이 창조되는 벡터 $\boldsymbol{\phi}' = (\phi_1, \phi_2, \ldots, \phi_p)$를 $\boldsymbol{X}$의 주성분이라 정의한다. 이때 j번째 고유값 λ_j에 대응하는 고유벡터 $\boldsymbol{e}_j$의 원소들을 $\boldsymbol{X}$와의 선형결합 (linear combination)에서 가중 계수로 사용하고 있다. 즉, (2.5)식에서 $\boldsymbol{\phi}'$의 j 번째 원소 ϕ_j를 $\boldsymbol{X}$의 j 번째 주성분이라고 하고, 다음과 같다.

$\boldsymbol{e}_j' = (e_{1j}, e_{2j}, \ldots, e_{pj})$, $j=1,2,\ldots,p$ 일 때,

$$\phi_j=\boldsymbol{e}_j\boldsymbol{X}=e_{1j}X_1+e_{2j}X_2+\ldots+e_{pj}X_p=\sum_{i=1}^{p}e_{ij}X_j \tag{2.6}$$

위와 같이 주성분 분석이란 전체 자료의 공분산 행렬 ($\boldsymbol{S}$)의 구조를 파악하여 고유값이 큰 고유벡터들의 축으로 자료의 축을 변환하여 주성분을 구하는 분석이다.

3. 주성분 분석을 이용한 주제어 추출

3.1 주제어 추출에의 주성분 분석 응용

여러 개(p≥2)의 반응 변수에 대하여 얻어진 다변량 자료를 분석 대상으로 하는 주성분 분석은 다차원적인 변수들을 축소, 요약하는 차원의 단순화와 더불어 일반적으로 서로 상관되어 있는 반응 변수들 간의 복잡한 구조를 분석하는데 그 목적을 두고 있다. 이를 위하여 주성분 분석은 반응 변수들을 선형 변환시켜, 주성분이

라고 부르는 서로 상관되어 있지 않은, 혹은 독립적인 새로운 인공 변수들을 유도한다. 이 때 각 주성분이 보유하는 변이의 크기를 기준으로 그 중요도 순서를 생각할 수 있는데, 그들 중 첫 소수 몇 개의 주성분이 원래 자료에 내재하는 전체 변이 중 가능한 많은 부분을 보유하도록 변환시킴으로서 정보의 손실을 최소화하는 차원의 축약을 기할 수 있게 된다[4]. 결국, 주성분 분석을 이용한다면 문서의 내용을 나타내기 위하여 문서에 출현하는 모든 단어를 사용하는 대신에, 정보의 손실을 최소화하면서 소수의 몇 개 단어로 문서의 내용을 표현할 수 있다. 즉, 그 문서의 주제어를 추출할 수 있다. 표 1은 주성분 분석에서 사용하는 자료 구조를 나타낸다.

표 1 주성분 분석의 자료 구조

변수 \ 개체	X_1	X_2	…	X_p
1 2 ⋮ n	(n × p) 양적 자료이여야 한다.			

표 1과 같은 자료 구조를 구성하기 위해 변수 $X_1, X_2, \cdots, X_p$를 문서에 2번 이상 출현하는 단어로, 개체 1,2, …,n을 문서의 각 문장으로, 그리고 그에 해당하는 자료를 문장에서 출현하는 단어의 누적 빈도 수를 사용한다. 여기서 어떤 문서에 단 한 번만 출현한 단어는 그 문서의 내용을 대표할 가능성이 적다는 가정 하에 배제하였다. 그리고, 하나의 문장에서 여러 번 발생하는 단어와 여러 문장에서 한 번 발생하는 단어의 변량에 차이를 줄이기 위해 단어의 누적 빈도 수를 이용한다. 예를 들어, '정보'라는 단어가 한 문장에서만 3번 발생한 경우와 '검색'이라는 단어가 3개의 문장에서 1번씩 발생한 경우가 있다고 가정하자. 이러한 경우에 '정보' 나 '검색' 모두가 3번 발생하였지만 그 변량의 차이로 '정보'라는 단어가 더 중요하게 판명될 가능성이 높다. 그래서, '정보' 나 '검색'이라는 두 단어 모두가 중요하게 판명될 수 있도록 누적 빈도 수를 이용한다. 누적 빈도 수를 이용한다면 '정보'라는 단어의 변량은 그대로 '3'으로 유지될 수 있고, '검색'이라는 단어는 '1'에서 '3'으로 증가하게 되어 두 단어 모두가 같은 변량을 가지게 된다. 즉, '정보' 나 '검색'이라는 두 단어 모두가 해당 문서에서 중요하게 판단될 가능성을 비슷하게 만들어서 주성분 분석을 시행한다.

그림 1은 50개 개체에 대한 변수 X_1, X_2 와 주성분 분석 결과로 얻을 수 있는 주성분 Y_1, Y_2를 도식화한 것이다. 자료의 형태는 그대로 있지만 그것들은 설명할

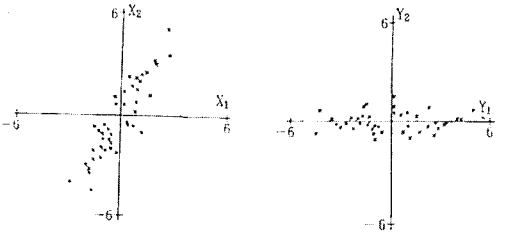

그림 1 50개 개체에 대한 변수(X_1, X_2)와 주성분 (Y_1, Y_2)의 플롯

수 있는 새로운 축이 만들어진다. 즉, X_1 축으로 설명할 수 있는 범위보다 Y_1 축으로 설명할 수 있는 범위가 더 넓다. 이러한 관점에서 어떤 문서에서 단어 X_1으로 그 문서를 설명하는 것보다 새로운 성분 Y_1 으로 그 문서를 설명하는 것이 더 효과적이다. 그렇다면 과연 몇 개의 주성분을 선택할 것이며, 새로 생성된 주성분을 어떻게 표현할 수 있을까? 그것은 다음과 같이 해결할 수 있다. 주성분 분석을 시행하여 얻은 p 개의 고유값 λ_j 들을 크기 순으로 배열하고 각각의 고유값에 대응되는 고유벡터 e_j의 짝들은 $(\lambda_1, e_1), (\lambda_2, e_2), \ldots, (\lambda_p, e_p)$이다. 단, $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p$ 의 순서이다. 그리고, 첫 m($\leq$ p)개의 주성분에 의해 설명되는 부분은 아래의 식 (3.1)이 된다.

$$(\lambda_1 + \lambda_2 + \ldots + \lambda_m)/(\lambda_1 + \lambda_2 + \ldots + \lambda_p) \qquad (3.1)$$

만약 첫 m개의 주성분들에 의해 설명되는 부분이 전체의 0.8~0.9(80~90%)를 점한다면 p보다 훨씬 작은 m개의 주성분들을 이용하더라도 정보상의 큰 손실이 없게 된다[4]. 즉, 식 (3.1)의 값이 0.8~0.9가 되는 첫 m개가 필요한 주성분의 개수가 된다.

$$e_i X = e_{1i} X_1 + e_{2i} X_2 + \ldots + e_{pi} X_p \qquad (3.2)$$

그리고, 생성된 주성분의 표현을 위해 주성분 적재 계수의 값이 0.5이상인 경우의 변수(단어)들을 사용한다. 주성분 적재 계수란 식 (3.2)에서 e_{ij} 를 말하며, 주성분에 기여하는 정도(상관도)를 의미한다. 그렇다면, 주성분과 0.5(50%)이상 상관이 있는 변수 즉, X_i 를 그 주성분의 표현으로 사용해도 무방하다고 볼 수 있다. 만약, 어떤 주성분에 대하여 주성분적재계수가 0.5이상인 것이 없으면 그 주성분에 대해서는 최대 주성분 적재 계수를 갖는 단어를 선택한다.

정리한다면, 문서의 내용을 80~90%이상 설명할 수 있는 첫 m개의 주성분들을 선택하고, 선택된 주성분에

서 그 주성분과의 상관도가 50%이상인 단어들이 주제어로 선택된다. 이는 주성분 분석 특성상 보다 많은 문장에서, 같이 출현하는 단어들을 문서의 주제어로 추출한다는 의미를 내포하고 있다.

3.2 주성분 분석을 이용한 주제어 추출의 실제

이번 절에서는 "임기 중 개헌 없다/김대통령 취임 100일 회견"라는 신문기사를 대상으로 주성분 분석을 이용한 주제어 추출 과정을 보인다. 설명력이 90%이상인 첫 m개의 주성분을 선택하였고, 선택된 주성분과의 상관도가 50%이상인 단어들을 주제어로 추출한다. 표 2는 실험 대상 문서에서 2번 이상 출현한 단어(변수) 리스트이다. 그리고 표 3은 주성분 분석에 이용한 자료구조를 보여주고 있다. 표 3의 자료 중 행의 모든 값이 '0'인 경우에는 주성분 분석을 수행하는 데 제외된다. 변수로 채택된 단어가 출현하지 않는 문장은 주성분 분석을 하는데 있어 의미가 없기 때문이다. 그리고, 표 3의 값은 각각의 문장을 개체로 보고, 변수가 발생한 누적 빈도 수를 나타내고 있다. 표 4는 표 3의 자료구조를 이용하여 주성분 분석한 결과를 보여준다.

표 2 주성분 분석에 이용한 변수 리스트의 예

변수	변수값
X1	대통령
X2	문제
X3	국가
X4	사람
X5	부정부패
X6	경제
X7	국민
X8	방법
X9	단체장

표 3 주성분 분석에 이용한 자료 구조의 예

문장\변수	X1	X2	X3	X4	X5	X6	X7	X8	X9
1	1	0	0	0	0	0	0	0	0
2	3	1	0	0	0	0	0	0	0
3	0	0	1	1	0	0	0	0	0
4	0	0	2	2	1	0	0	0	0
5	0	0	0	0	2	2	0	0	0
6	0	0	3	0	0	0	1	0	0
7	0	2	0	0	0	0	2	1	0
8	0	0	0	0	0	0	0	0	0
9	4	3	0	0	0	0	0	0	0
10	0	0	0	0	0	0	0	2	2
11	0	0	0	0	0	0	0	0	0
12	0	0	0	0	0	0	0	0	0

표 4의 결과에 기초하여, 누적 비율이 90%이상의 시점인 'PRIN4'까지의 주성분 4개를 선택한다. 그리고 'PRIN1'부터 'PRIN4'까지의 주성분 적재 계수의 값이 0.5이상인 모든 변수를 택한다. 만약, 주성분 계수의 값이 0.5이상인 경우가 없다면 주성분 적재 계수 중 최고치의 변수를 택한다. 표 5는 설명력이 90%이상인 주성분을 이용하여 "임기 중 개헌 없다/김대통령 취임 100일 회견"라는 신문기사의 주제어를 추출한 내용이다.

표 5 주성분 분석을 이용하여 추출된 주제어 예

주성분	변수	변수값
PRIN1	X1	대통령
PRIN2	X3	국가
PRIN3	X7	국민
PRIN4	X7	국민
전체(주제어)	X1, X3, X7	대통령, 국가, 국민

표 4 주성분 분석한 결과의 예

변수\주성분	PRIN1	PRIN2	PRIN3	PRIN4	PRIN5	PRIN6	PRIN7	PRIN8	PRIN9
X1	**0.76803**	0.274927	−0.15314	−0.27509	−0.12958	0.185339	0.390235	0.17887	0
X2	0.494488	−0.00972	0.307326	0.454604	0.38267	0.273102	−0.4588	−0.15085	0
X3	−0.33156	**0.66669**	0.30888	−0.05403	−0.19457	0.540753	−0.06626	0.023586	−0.10976
X4	−0.15898	0.287993	−0.05398	−0.14109	0.828807	−0.11295	0.149947	0.367089	0.109764
X5	−0.1318	−0.00888	−0.5034	0.302651	0.209538	0.395579	0.403312	−0.52392	0
X6	−0.07167	−0.13225	−0.4765	0.344139	−0.15809	0.312008	−0.22184	0.678479	0
X7	−0.04248	−0.06658	**0.45518**	**0.53051**	−0.12024	−0.06694	0.572686	0.223144	0.329293
X8	−0.05129	−0.47847	0.286247	−0.16783	0.161386	0.327185	0.25746	0.158804	−0.65859
X9	−0.05881	−0.38206	0.119133	−0.41857	0.050942	0.469607	−0.06492	−0.01002	0.658586
고유값	3.631536	1.439873	1.162859	0.808426	0.384823	0.283013	0.035279	0.004191	0
누적비율(%)	46.86%	65.44%	80.44%	**90.87%**	95.84%	99.49%	99.95%	100.00%	100.00%

4. 실험 및 평가

본 논문에서 제안한 모델을 문서 요약에 적용하여 그 성능을 실험 및 평가하였다.

문서 요약 방법은 크게 통계적 기법, 수사 구조에 기반한 방법, 지식에 기반한 방법으로 나눌 수 있다. 통계적 기법에서는 단어의 출현 빈도, 제목, 문장의 길이, 실마리 단어나 구(cue word or phrase) 등을 자질(feature)로 사용하여 각 문장이나 문단의 중요도를 계산하여 그 값이 높은 문장이나 문단을 요약으로 제시한다[5,6]. 그리고, 수사 구조에 기반한 방법은 문장들 사이의 수사 관계를 파악하여 요약을 생성한다[7]. 마지막으로, 지식에 기반한 방법은 생성하고자 하는 문서와 관련된 배경 지식을 이용하여 요약을 생성하는 방법이다[8,9,10].

본 논문에서 제안한 모델을 단어의 출현 빈도만을 고려하는 방법, 그리고 시소러스를 이용한 방법과 비교 실험을 하였다. 실험에 사용된 문서 집합은 KISTI(한국과학기술정보연구원)에서 제공하는, 두 명의 사람에 의하여 수동 요약된 신문기사 문서 집합으로 구성된 테스트 컬렉션이다. 실험은 테스트컬렉션에서 제공하는 1000건 문서 중 100건에 대해서 30% 중요문장 추출을 하였다.

실험 결과에 대한 평가는 테스트컬렉션에서 제공하는 30% 추출 요약과의 일치 문장 수를 기반으로 하였다.

4.1 실험 자료

4.1.1 시소러스(thesaurus)

실험에 사용된 정보 검색용 시소러스는 단어, 상위어, 동의어, 유의어 등의 MS-Access 테이블로 구성되어 있다. 하위어 관계는 상위어의 역으로 추정하였다. 그리고, 각 테이블의 통계적인 특성은 표 6과 같다.

표 6 시소러스의 통계적인 특성

구분	건수
단어	142,682
상위어	124,390
동의어	71
유의어	6,394

4.1.2 테스트컬렉션(test collection)

본 논문이 제안한 모델을 실험하기 위하여 K. ?I(한국과학기술정보연구원)에서 제공하는 테스트컬렉션을 사용하였다. 이 테스트컬렉션은 두 명의 사람에 의하여 수동 요약된 신문기사 문서집합(1000건)으로 구성되어 있다. 그리고, 각 문서에 대해 10%, 30% 중요문장 추출, 10% 수동요약 결과를 제공한다. 테스트컬렉션의 각 문서는 제목(#T), 본문(#S), 10% 추출 요약(#A), 30% 추출 요약(#B), 10% 수동 요약(#C)으로 나누어져 있다. 본 논문에서는 제공되는 신문기사 1000건 중 100건을 이용하여 문서 요약 실험을 하였다. 실험한 100건의 문서에 대한 통계적인 특성은 표 7과 같다.

표 7 실험대상 문서집합(100건)의 통계적인 특성

대상 영역	신문기사
문서 개수	100건
문서의 평균길이	19.51 문장
요약의 평균길이	5.83 문장
문장의 평균길이	6.79 개(명사)
평균변수개수(주성분 분석)	19.54 개(명사)

4.2 실험에 사용된 문서 요약 모형

추출된 주제어는 실험 방법에 따라 크게 네 가지로 구분할 수 있다. 첫 번째는 단순히 단어의 빈도만을 고려하여 주제어를 선택하는 경우이고, 두 번째는 시소러스를 이용하는 방법[10], 세 번째는 제안한 모델인 주성분 분석을 이용하는 것이다. 그리고, 네 번째는 시소러스와 주성분 분석을 같이 사용하여 주제어를 추출하는 방법이다.

문장 중요도를 계산할 때는 추출된 주제어가 어떤 문장에 포함되어 있으면 그 문장에 가중치를 주고, 그렇지 않으면 가중치를 주지 않는 방법으로 문장들 간의 가중치에 차등을 만들어 간다. 이러한 과정을 추출된 모든 주제어와 모든 문장에 반복 시행한다. 이렇게 계산된 문장 중요도를 그 문장의 길이 즉, 문장에 포함된 단어수로 나누어 긴 문장 선호도를 완화하였다. 이렇게 정규화 과정까지 거친 후에 각 문장들을 중요도에 따라 내림차순으로 정렬한다. 그런 후에 사용자가 원하는 비율로 문장을 추출하고, 추출된 문장은 문서에 나타난 순서로 재

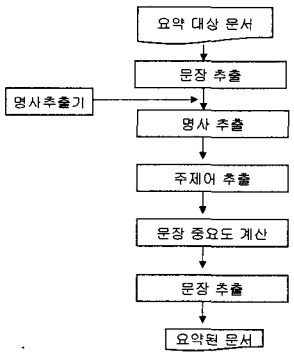

그림 2 주제어 기반 문서 요약 모형

정렬하여 요약으로 제시한다.

4.3 실험 방법 및 평가

아래의 (가)~(바)까지의 방법으로 주제어를 추출한 후 주제어 기반 문서 요약(30% 추출) 실험을 시행하였다. 그리고, 테스트컬렉션이 제시하는 30% 추출 요약과의 일치 문장 개수를 이용하여 각 방법을 평가하였다. 각 실험 방법은 다음과 같다.

- (가) 출현 빈도 2번 이상의 단어들을 주제어로 선택
- (나) 최고 점수의 사슬에 포함된 단어들을 주제어로 선택
- (다) 각각의 문장을 개체로 보고 주성분 분석을 이용하여 주제어 선택
- (라) 연이은 2개의 문장을 개체로 보고 주성분 분석을 이용하여 주제어 선택
- (마) 연이은 3개의 문장을 개체로 보고 주성분 분석을 이용하여 주제어 선택
- (바) (나)와 (다)에서 추출된 주제어를 합쳐 주제어로 선택

'임기 중 개헌 없다/김대통령 취임 100일 회견'라는 제목의 신문기사에 대해 방법(가)에 의해 추출된 주제어는 '대통령, 문제, 국가, 사람, 부정부패, 경제, 국민, 방법, 단체장등' 이며, 방법(나)는 '국가, 국민', 방법(다)는 '대통령, 국가, 국민', 방법 (라)는 '경제, 대통령, 단체장등, 사람', 방법 (마)는 '대통령, 국가, 국민', 그리고 방법

표 8 신문기사 100건에 대한 각 실험 방법의 일치 문장 수

#	요약수	(가)	(나)	(다)	(라)	(마)	(바)	#	요약수	(가)	(나)	(다)	(라)	(마)	(바)
1	9	6	5	6	5	5	6	51	8	3	1	1	2	2	1
2	4	2	2	3	1	2	3	52	4	2	3	4	4	4	4
3	12	4	3	6	6	6	6	53	6	0	2	1	0	1	1
4	9	2	2	2	2	2	2	54	3	0	0	1	1	1	1
5	6	0	0	1	3	2	1	55	7	2	5	3	3	3	3
6	4	1	0	1	1	1	1	56	3	0	1	1	1	0	1
7	6	1	4	2	3	3	2	57	3	0	0	0	1	1	0
8	3	1	1	1	1	1	1	58	8	1	2	5	3	4	5
9	3	1	0	1	2	2	2	59	11	1	5	4	5	5	4
10	6	1	1	2	2	3	2	60	5	1	3	2	1	3	2
11	6	0	0	0	0	0	0	61	15	2	3	4	4	4	3
12	6	1	2	2	2	2	2	62	4	0	0	1	3	2	1
13	4	0	0	0	0	0	0	63	5	2	2	0	2	1	0
14	4	1	0	1	1	1	1	64	9	2	2	3	3	3	2
15	5	0	0	0	0	0	0	65	9	2	3	3	2	4	3
16	6	0	0	0	0	1	0	66	8	4	1	1	1	2	1
17	9	2	0	0	0	0	0	67	7	2	1	3	4	2	3
18	6	0	0	0	0	0	0	68	7	3	1	2	2	2	2
19	5	1	1	1	1	1	1	69	5	1	1	2	2	2	2
20	2	1	1	1	0	0	1	70	3	1	0	2	2	2	2
21	7	0	0	0	0	0	0	71	4	1	2	2	3	2	1
22	3	0	0	0	0	0	0	72	3	0	1	2	2	2	1
23	6	1	0	1	1	1	1	73	4	2	3	3	3	3	3
24	3	1	0	1	1	1	1	74	4	1	0	1	0	1	1
25	3	0	0	1	0	1	1	75	4	1	2	3	2	2	3
26	3	0	1	1	1	1	1	76	6	3	2	2	2	2	2
27	6	2	2	3	3	4	3	77	3	0	1	1	1	1	1
28	9	2	3	3	3	3	3	78	6	1	3	0	0	0	0
29	4	2	2	2	2	3	2	79	5	1	1	2	1	1	2
30	4	0	2	0	0	1	0	80	7	3	4	4	4	4	4
31	5	1	0	1	1	2	1	81	8	1	2	3	3	2	3
32	3	1	1	1	1	1	1	82	16	3	5	5	4	4	5
33	5	0	1	1	1	1	1	83	5	2	4	3	3	3	3
34	15	3	5	4	3	4	5	84	9	1	4	3	2	4	3
35	5	1	2	3	1	2	1	85	7	5	4	6	5	6	6
36	9	2	4	4	3	4	3	86	5	2	2	2	3	2	2
37	3	1	1	1	2	1	1	87	3	1	1	2	1	2	2
38	8	4	4	3	3	3	3	88	5	0	2	2	3	3	2
39	6	1	0	0	0	0	0	89	2	1	1	1	1	1	1
40	4	0	1	1	0	1	1	90	2	0	1	1	1	2	1
41	4	2	3	2	2	2	2	91	7	4	0	0	1	1	0
42	5	2	1	3	1	1	2	92	1	0	0	0	0	0	0
43	4	0	0	0	0	0	0	93	1	0	1	0	0	1	0
44	6	1	2	3	3	3	3	94	4	1	1	0	1	1	0
45	4	0	2	1	2	0	1	95	5	2	1	1	1	1	1
46	3	1	1	1	1	1	1	96	4	2	3	3	3	3	3
47	4	1	1	1	0	2	1	97	6	1	4	1	1	1	1
48	2	0	1	2	2	1	2	98	2	1	1	2	2	1	2
49	7	1	1	5	5	5	5	99	1	0	0	0	0	0	0
50	6	3	3	4	4	3	4	100	2	1	1	1	1	1	1

<합계> 요약 수 : 542, (가) : 128, (나) : 159, (다) : 182, (라) : 176, (마) : 189, (바) : 176

(바)는 '대통령, 국가, 국민'이다. 이와 같이 추출된 주제어를 기반으로 신문기사 100건을 문서 요약한 결과의 평균 일치 문장 수는 그림 3과 같다. 그리고, 각 실험 방법의 일치 문장 수는 표 8과 같다. 실험 결과 단순히 단어의 출현 빈도를 이용하여 주제어를 추출하는 경우보다 본 논문에서 제안한 모델이 일치 문장이 평균 4~7개 더 많다는 사실을 확인할 수 있었다.

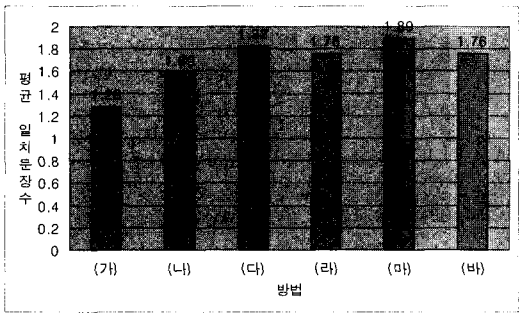

그림 3 각 실험 방법의 평균 일치 문장 수

또한, 시소러스를 이용하는 경우인 방법(나) 보다는 주성분 분석을 이용한 방법(다), 방법(라), 방법(마)가 평균 일치 문장이 많음을 알 수 있다. 그리고, 방법(바)는 시소러스만을 이용한 방법(나)의 자료 희귀성(data sparseness) 문제를 보완하는 측면에서 주성분 분석을 이용하였는데 결과는 더 좋게 나옴을 알 수 있다. 실험한 문서 100건의 평균 단어 개수는 87.52이고, 평균 어휘 사슬의 개수는 84.31이다. 이는 방법(나)는 시소러스의 동의어, 유의어, 상위어, 하위어 관계를 대부분 참조하지 못하고 있음을 단적으로 보여주고 있다. 즉, 142,682 단어의 시소러스에서는 자료 희귀성 문제가 발생될 수 있음을 암시하고 있는 것이다.

결과적으로, 실험을 통하여 알 수 있는 사실은 방법(가)보다는 시소러스를 이용한 방법(나)가, 그리고 방법(나)보다는 주성분 분석을 이용한 방법(다), 방법(라), 방법(마)가 성능 면에서 더 우수함을 알 수 있었다.

5. 결 론

본 논문에서는 통계적 분석 기법 중의 하나인 주성분 분석을 이용하여 문서의 대표 단어들 즉, 주제어들을 추출하는 모델을 제안하였다. 주성분 분석의 고유값과 고유벡터를 이용하여 문서 자체내의 단어의 흐름을 정량화 하고자 하였고, 그 정보를 이용하여 다른 도구의 도움 없이 문서 자체 내에서의 단어 발생 빈도와 공기 정보를 이용하여 주제어를 추출하고자 하였다.

제안한 모델을 문서 요약에 적용하여 실험을 하였다. 실험 결과, 단순히 단어의 출현 빈도만을 이용한 방법, 시소러스를 이용하는 방법 모두에 비해 제안한 모델이 더 좋은 성능을 보임을 알 수 있었다. 또한, 시소러스와 주성분 분석은 상호 배타적인 정보가 아니고 상호 보완될 수 있는 정보라는 사실도 확인할 수 있었다. 결국 지식베이스와 주성분 분석을 각각 이용할 수도 있지만 지식베이스의 약점이 통계적 분석 기법을 통하여 보완될 수 있고, 그 역도 가능하다는 사실을 단적으로 보여주고 있다.

제안한 모델은 문서의 키워드를 추출하는 방법이다. 이를 응용한다면 정보검색, 정보추출, 문서요약 등에 적용될 수 있을 것이라 기대된다.

참 고 문 헌

[1] Willaim B. Frakes, Richard Baeza-Yates, Information Retrieval : Data Structures & Algorithms, Prentice-Hall, 1992

[2] 황이규, 이근용, 김남수, 이용석, "구문형태소를 이용한 색인어 추출", 한글 및 한국어 정보처리, 2000

[3] 김영택, 자연언어처리, 교학사, 1994

[4] 김기영, 전명식, "다변량 통계자료분석", 자유아카데미, 1994

[5] H. P. Edmundson, "New Methods in Automatic Extracting," Journal of the Association for Computing Machinery, Vol. 16, No. 2, pp. 264-285, 1969

[6] J.Kupiec, J.Pedersen, F.Chen, "A Trainable Document Summarizer," Proc. 18th ACM-SIGIR Conf., 1995

[7] 양기주, "수사구조에 기반한 한국어 요약문 생성," 연구개발정보센터, 1997

[8] Eduard Hovy and Chin Yew Lin, "Automated Text Summarization in SUMMARIST," Proc. Association for Computational Linguistics, pp.18-24, 1997

[9] Regina Barzilay, Michael Elhadad, "Using Lexical chains for Text Summarization," proc. Association for Computational Linguistics, pp.10-17, 1997

[10] 이창범, 박혁로, "시소러스를 이용한 문서 자동 요약", 제28회 정보과학회 봄 학술발표 논문집(B), 제28권, 1호., pp.352-354, 2001

[11] Gerald Salton, Amit Singhal, "Automatic Text Theme Generation and the Analysis of Text Structure," Computer Science Department Technical Report, Cornell University 1994

[12] Jose Abracos, Gabriel Pereira Lopes, "Statistical methods for retrieving most significant paragraphs

in newspaper articles," Proc. Association for Computational Linguistics, pp.51-57, 1997
[13] 강상배, 조혁규, 권혁철, 박재득, 박동인, "한국어 문서의 통계적 정보를 이용한 문서 요약 시스템 구현", 제9회 한글 및 한국어 정보처리학술대회, pp.28-36, 1997
[14] 김계성,이현주,정영규,서연경,손기준,이상조, "단락 자동 구분을 통한 중요 문장 추출", 한글 및 한국어 정보처리 학술발표 논문집, 2000
[15] 김재훈, 김준홍, "도함유사도를 이용한 한국어 추출문서 요약", 한글 및 한국어 정보처리 학술발표 논문집, 2000
[16] 류동원, 이종혁, "단어공기정보를 이용한 자동화 문서 요약", 제27회 정보과학회 봄 학술발표 논문집(B), 제27권, 1호, pp.339-341, 2000.
[17] 류 제, 한광록, 손석원, 임기욱, "단어의 공기 관계 그래프를 이용한 문서의 핵심 문장 추출에 관한 연구", 한국정보처리학회 논문지 제7권 제11호, pp.3427-3437, 2000
[18] 박혁로, 이현민, 전남열, 최선화, 정경석, "Answer Set 구축 지원도구 개발에 관한 연구", 한국전자통신연구원 연구보고서, 2000
[19] 박혁로, 신중호, "검색/요약/필터링을 위한 텍스트 이해 모형 및 처리 기술 개발", 연구개발정보센터 연구보고서, 1999
[20] 이창기, 이근배, "WordNet을 이용한 한국어 시소러스 자동 구축", 제 11회 한글 및 한국어정보처리 학술대회
[21] 장동현,맹성현, "자동 요약 시스템", 정보과학회지 제15권 제10호, pp.42-49, 1997
[22] 한경수, 백대호, 임해창, "질의 확장을 이용한 자동 문서 요약", 제27회 정보과학회 봄 학술발표 논문집(B), 제27권, 1호, pp.339-341, 2000.
[23] 한영석, 김선섭, 나태현, 김인석, "한국어 문서 자동요약 엔진 개발", 연구개발정보센터, 1998

이 창 범

1995년 전남대학교 전산학과 졸업(학사). 2001년 전남대학교 대학원 전산학과 졸업(석사). 1995.1월 ~ 1999.5월 : 대우정보시스템(주). 2002.3월 ~ 현재 전남대학교 대학원 전산학과 박사과정. 관심분야는 정보검색, 자연어처리, 문서요약

김 민 수

1994년 전남대학교 전산학과(이학사). 1996년 전남대학교 전산통계학과(이학석사). 2000년 전남대학교 전산통계학과(이학박사). 2000년 ~ 2002년 전남대학교 BK사업단 post-doc. 관심분야는 통계적 패턴인식, 다변량 통계분석, wavelets.

이 기 호

1973년 서울문리대 해양학과(학사). 1985년 Bowling Green 주립대 전산학과(석사). 1999년 충남대 컴퓨터공학과(박사). 1991년 ~ 2000년 한국과학기술정보원 선임연구원. 2001년 ~ 현재 충북과학대 컴퓨터정보학과 객원교수. 관심분야는 정보검색, 전자도서관, 자연어처리

이 귀 상

1980년 서울대학교 전기공학과 졸업(학사). 1982년 서울대학교 대학원 전자계산기공학과 졸업(석사). 1991년 펜실바니아 주립대학 전산학과(박사). 1982년 11월 ~ 1983년 3월 금성통신연구소. 1984년 8월 ~ 현재 전남대학교 전산학과 교수. 관심분야는 멀티미디어 통신, 영상처리 및 복원, 논리합성, VLSI/CAD

박 혁 로

1987년 서울대학교 전산학과 졸업(학사). 1989년 한국과학기술원 전산학과 졸업(석사). 1997년 한국과학기술원 전산학과 졸업(박사). 1994년 ~ 1996년 연구개발정보센터 연구원. 1997년 ~ 1998년 연구개발정보센터 선임연구원. 1999년 ~ 현재 전남대학교 전산학과 조교수. 관심분야는 정보검색, 자연어처리, 데이터베이스