주성분 분석과 비정칙치 분해를 이용한 문서 요약

이 창 범[†] · 김 민 수[††] · 백 장 선[†††] · 박 혁 로[††††]

요 약

본 논문에서는 통계적 분석 기법인 주성분 분석과 비정칙치 분해를 이용한 문서 요약 방법을 제안한다. 제안한 방법은 문서내의 주제어를 추출한 후, 추출된 주제어와 문장간의 거리가 가장 짧은 문장들을 중요 문장으로 추출하여 요약으로 제시한다. 주제어를 추출하기 위해서는 주성분 분석을 이용하였으며, 이는 문서 자체내의 빈도 정보와 단어간의 연관 정보를 이용한 것이다. 그리고, 중요 문장을 추출하기 위해 비정칙치 분해를 시행하여 문장 벡터와 주제어 벡터를 획득한 후, 두 벡터간의 유클리디언 거리를 계산하였다. 신문 기사를 대상으로 실험한 결과, 제안한 방법이 출현 빈도만을 이용한 방법과 주성분 분석만을 이용한 방법보다 성능이 우수함을 알 수 있었다.

Text Summarization using PCA and SVD

Chang-Beom Lee[†] · Min-Soo Kim[††] · Jang-Sun Baek[†††] · Hyuk-Ro Park[††††]

ABSTRACT

In this paper, we propose the text summarization method using PCA (Principal Component Analysis) and SVD (Singular Value Decomposition). The proposed method presents a summary by extracting significant sentences based on the distances between thematic words and sentences. To extract thematic words, we use both word freqeuncy and co-occurence information that result from performing PCA. To extract significant sentences, we exploit Euclidean distances between thematic word vectors and sentence vectors that result from carrying out SVD. Experimental results using newspaper articles show that the proposed method is superior to the method using either word frequency or only PCA.

키워드 : 주성분 분석(Principal Component Analysis), 비정칙치 분해(Singular Value Decomposition), 문서요약(Text Summarization)

1. 서 론

정보가 기하급수적으로 증가하고 있고 그 중요성 또한 증대되고 있다. 이러한 정보과적재 상황에서 개인은 자기가 원하는 정보를 빨리 그리고 적합하게 찾고자 한다. 이에 문서 요약의 필요성이 점점 증가하고 있다. 만약, 검색 엔진의 결과로 해당 문서의 내용을 정확하게 표현하는 요약의 형태를 취한다면, 개인이 원하는 문서를 찾는 데 소비하는 시간을 줄일 수가 있다. 즉, 문서 요약 시스템은 개인이 원하는 정보를 찾는 데 걸리는 시간을 단축시킴으로써 정보과적재 문제에 대해 효과적인 해결책을 제시할 수 있다.

문서 요약은 문서의 기본적인 내용을 유지하면서 문서의 복잡도 즉 문서의 길이를 줄이는 작업이다[1]. 문서 요약의 한 방법은 문서의 내용을 대표하는 중요 문장만을 추출하여 요약으로 제시하는 것이다[2]. 본 논문에서는 중요 문장을 추출하기 위해 통계적 분석 기법 중 주성분 분석(PCA, Principal Component Analysis)과 비정칙치 분해(SVD, Singular Value Decomposition)를 이용하는 방법을 제안한다.

제안한 방법은 문서의 주제를 파악하는 단계와 중요 문장을 추출하는 단계로 나누어 볼 수 있다. 주제를 파악하는 단계에서는 문서내의 단어 중 문서의 내용을 대표하는 주제어를 추출한다. 주제어를 추출하기 위해 단순히 단어의 출현 빈도만을 이용할 경우, "몇 번 이상 발생한 단어를 선택할 것인가?", "자주 발생하는 단어가 문서를 대표할만 하지 않는다면 어떻게 할 것인가?" 등의 의구심을 표명할 수 있다. 만약, 같은 주제를 나타내는 단어들은 같이 출현하는 경향이 있다라는 것을 가정하고, 주제어를 추출하기 위해 출현 빈도뿐만 아니라 단어 관련성을 이용한다면 보다 타당한 주제어를 선택할 수 있다. 이에 본 논문에서는 주제어를 선택함에 있어 주성분 분석을 시행하여 출현 빈도와 단어 관련성 정보를 이용하고자 한다. 주성분 분석은 문장-명사 행렬의 공분산을 고유 시스템(eigen system)의 입력으

† 준 회 원 : 전남대학교 대학원 전산학과
†† 정 회 원 : 한국과학기술원
††† 정 회 원 : 전남대학교 통계학과 교수
†††† 종신회원 : 전남대학교 전산학과 교수
논문접수 : 2003년 6월 12일, 심사완료 : 2003년 10월 29일

로하여 분석을 한다. 여기서 문장-명사 행렬에는 출현 빈도 정보가 포함되며, 공분산에는 단어가 같이 출현할 정도를 수량화한 정보가 포함된다. 즉, 주제어를 선택함에 있어서 출현 빈도와 단어 관련성 정보를 동시에 이용하며, 더구나 다른 도구(정보검색용 시소러스, WordNet 등)의 도움없이 문서 자체내의 정보만을 사용한다.

중요 문장을 추출하는 단계에서는 주제어로 추출된 단어 벡터와 문장 벡터간의 거리를 이용한다. 먼저, 문장-명사 행렬에 대해 비정칙치 분해를 시행한다. 그러면, 문장-명사 행렬에서 문장-차원 행렬과 단어-차원 행렬로 분해할 수 있다. 여기서, 추출된 주제어에 해당하는 단어 벡터와 문장 벡터 사이의 유클리디언 거리를 계산하여, 계산된 거리가 최소인 문장을 중요 문장으로 추출한다.

본 논문의 구성은 다음과 같다. 2장에서는 문서 요약에 대한 기존 연구에 대해 설명하고, 3장에서는 주성분 분석과 비정칙 분해를 이용한 문서 요약 방법에 대해서, 그리고 4장에서는 실험 및 평가에 대해서 기술한다. 마지막으로 5장에서는 논문을 결론짓는다.

2. 관련 연구

단어의 출현 빈도에 기반한 방법[1, 3, 4]은 요약하고자 하는 문서에 나타난 단어의 빈도를 측정하여 대표하는 단어 집합을 설정한 후, 이를 기반으로 적당한 요약문을 생성하는 방법이다. 이 방법은 단어의 출현 빈도만을 고려하기 때문에 문서에 나타나는 문장들 사이의 관계나 문맥 구조를 제대로 표현하지 못하여, 일관된 논리를 갖지 못하는 요약을 생성할 수도 있다.

정보검색용 시소러스를 이용한 방법[5]은 정보 검색용 시소러스를 이용하여 단어의 동의어, 유의어, 상위어, 하위어 등을 파악하여 요약을 하고자 하였다. 여기서는 자료 희귀성(data sparseness) 문제가 발생할 수 있으며, 시소러스라는 다른 도구를 이용해야만 한다.

WordNet을 이용하는 방법[6]은 WordNet을 이용하여 같은 개념을 갖는 단어들을 어휘 사슬(lexical chain)로 구성하여 강력한 사슬이 있는 문장을 선정하여 요약문을 생성한다. 하지만 이 방법도 어휘 사슬을 구성하기 위해 부가적인 정보 도구 즉, WordNet을 이용해야만 한다.

주성분 분석을 이용하는 방법[7]은 부가적인 정보 도구의 도움없이 문서내의 정보의 흐름을 파악하는데 주성분 분석을 이용하고, 주성분 분석 결과로 나오는 고유값(eigenvalue)과 고유벡터(eigenvector)를 이용하여 주제어를 선정하여 중요 문장을 추출한다. 하지만 이 방법은 중요 문장을 추출하는데 있어 단순히 문장에서의 주제어 출현 여부만을 이용하였다.

제안한 방법은 정보검색용 시소러스나 WordNet 등의 다른 도구의 도움없이 해당 문서내의 단어의 출현 빈도와 연관성을 이용하고 있다. 이는, 시소러스나 WordNet 등의 정보를 구축하기 위한 비용을 절감할 수 있으며, 요약을 생성함에 있어 오직 문서 자체내의 정보만을 이용하기 때문에 요약 생성에 소요되는 시간을 줄일 수 있는 효과도 발휘할 수 있다. 또한, [7]과 같이 주제어 출현 여부만으로 중요 문장을 추출하지 않고, 주제어와 문장간의 유클리디언 거리를 이용하고 있기 때문에, 경험적인 방법보다는 보다 정형화된 방법으로 주제어와 문장간의 유사성을 계산하고 있다.

3. 주성분 분석과 비정칙치 분해를 이용한 문서 요약

"미래형 생명 존중 아파트 등장"이라는 제목의 신문 기사를 예로 하여 제안한 방법들을 기술한다. 3.1절에서는 문장-명사 행렬의 구성에 대해, 3.2절에서는 주성분 분석을 이용한 주제어 추출에 대해 설명한다. 그리고, 3.3절에서는 비정칙치 분해를 이용한 중요 문장 추출 방법에 대해 언급한다.

3.1 문장-명사 행렬의 구성

문장-명사 행렬을 구성하기 위해 해당 문서에서 2번 이상 출현한 명사만을 대상으로 한다. 이 때, 명사는 형태소 분석[8]과 태깅 과정을 수행한 후에 추출하였다. "미래형 생명 존중 아파트 등장"이라는 제목의 신문 기사에 대한 문장-명사 행렬은 <표 1>과 같다. 여기에서 행은 문서내의 문장 번호를 나타내며 열은 <표 2>에서 제시한 명사들을 의미한다. 예를 들어, 첫 번째 문장은 "공간", "System", "싱크대", "지하" 라는 명사가 1번씩 출현했다는 것을 보이고 있다. 또한, 명사 리스트에 나타난 명사가 단 한 번도 출현하지 않는 문장 즉, 문장-명사 행렬에서 행의 값이 모두 "0"인 행은 삭제한다.

〈표 2〉 명사 리스트의 예

x1	공 간	x11	회 사
x2	System	x12	컴퓨터
x3	싱크대	x13	화 재
x4	지 하	x14	센 서
x5	거 실	x15	관리실
x6	아파트	x16	외 부
x7	코리아	x17	모니터
x8	주 상	x18	벽
x9	시스템	x19	실 내
x10	주 부	x20	모 델

〈표 1〉 문장-명사 행렬 (17 × 20)의 예

문장	x1	x2	x3	x4	x5	x6	x7	x8	x9	x10	x11	x12	x13	x14	x15	x16	x17	x18	x19	x20
1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
3	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
4	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
5	0	0	0	1	0	1	1	2	0	0	0	0	0	0	0	0	0	0	0	0
6	0	0	0	0	0	1	1	0	1	0	0	0	0	0	0	0	0	0	0	0
7	0	1	0	0	0	4	0	0	1	1	1	0	0	0	0	0	0	0	0	0
8	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0
9	0	0	0	0	0	0	0	0	0	0	0	0	2	1	1	0	0	0	0	0
10	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1	0	0	0	0
11	0	0	0	0	0	0	0	0	0	0	0	1	0	0	1	0	1	0	0	0
12	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0
13	0	0	1	0	0	0	0	0	0	1	0	0	0	0	0	1	1	0	0	0
14	0	0	0	0	0	1	0	0	0	0	1	0	0	0	0	0	0	0	1	0
15	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0
16	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	1
17	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1

3.2 주성분 분석을 이용한 주제어 추출

3.2.1 주성분 분석의 개요

이 절에서는 주성분 분석에 대한 개략적인 설명을 하고자 한다[9, 10].

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

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

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 $(\lambda_1, \boldsymbol{e}_1), (\lambda_2, \boldsymbol{e}_2), \cdots, (\lambda_p, \boldsymbol{e}_p)$라하고, λ_j들을 크기 순으로 배열하면,

$$S\boldsymbol{e}_j = \lambda_j \boldsymbol{e}_j \quad , j = 1, 2, \cdots, p$$
$$\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p$$

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

$$SP = P\Lambda, \quad S = P\Lambda P'$$

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

$$\boldsymbol{P} = (\boldsymbol{e}_1, \boldsymbol{e}_2, \cdots, \boldsymbol{e}_p)$$
$$\boldsymbol{\Lambda} = \text{Diagonal}(\lambda_1, \lambda_2, \cdots, \lambda_p),$$

이와 같은 $\boldsymbol{P}$를 이용하여 다음과 같은 $\boldsymbol{X}$의 직교변환을 생

각할 때,

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

이 변화에 의해 새로이 창조되는 벡터 $\boldsymbol{\phi}' = (\phi_1, \phi_2, \cdots, \phi_p)$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, 식 (2.5)에서 $\boldsymbol{\phi}'$의 j번째 원소 ϕ_j를 $\boldsymbol{X}$의 j번째 주성분이라고 하고, 다음과 같다.

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

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

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

결국, 여러 개 ($p \geq 2$)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, 주성분이라고 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 [9]. 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.

3.2.2 주성분 분석을 이용한 주제어 추출 과정

<표 1>의 문장-명사 행렬에 대해 주성분 분석을 시행한 결과 구해진 고유값과 고유벡터는 <표 3>과 같다.

〈표 3〉 고유값 (20 × 20, 대각행렬)과 고유벡터 (20 × 20)의 예

명사\주성분	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	…	PC20
x1	-0.0518931	0.18930784	0.11237872	0.09311996	-0.5408005	-0.0885046	-0.0166791	0.08184999	-0.1350177	…	-0.4712884
x2	0.15049332	0.04580353	0.08387453	0.24016869	-0.3473801	0.10450471	-0.3452719	0.04221433	0.01432919	…	0.4248398
x3	-0.0471775	0.13762124	0.13369295	**0.485827**	-0.1457963	0.11903087	-0.1215172	0.0772375	0.04840998	…	-0.0453849
x4	0.00150351	0.29637533	-0.2993639	0.21195707	-0.2535155	0.04122765	-0.1755896	0.0937287	0.13029665	…	0.09183554
x5	-0.0883336	0.12041479	0.16490892	-0.343022	0.18923396	0.33013257	-0.461387	-0.3290976	0.28309527	…	2.36E-06
x6	**0.9042802**	-0.1140468	-0.0326891	-0.048856	0.00858351	-0.0109175	-0.0005884	0.03457682	0.10768709	…	6.40E-07
x7	0.05193401	0.20301054	-0.3715072	-0.0080272	0.11072471	-0.0074615	0.05364516	-0.1518884	-0.5035133	…	0.15672708
x8	0.04713135	**0.3617938**	-0.6443107	0.05354505	0.08508975	-0.0421324	0.13316856	-0.1649975	**0.33566**	…	-0.1242807
x9	0.2009238	-0.0475611	0.01173129	0.02018445	0.01686028	0.05581546	-0.1160373	-0.2034026	-0.6194806	…	-0.1567259
x10	0.14744023	-0.047532	**0.1719846**	0.35562655	0.09894457	0.09894767	0.01755873	-0.2330031	0.13780581	…	0.20317575
x11	0.19823998	-0.0702297	0.11829029	-0.0214565	-0.1336631	-0.0966002	0.14868438	-0.2268366	0.24600263	…	-0.4712905
x12	-0.06649	-0.1671267	-0.0473112	0.07615501	0.117816	-0.5482363	-0.2420288	-0.0328996	-0.0295768	…	0.05259826
x13	-0.1165912	-0.6329431	-0.355279	-6.396E-05	-0.3056322	0.14580257	-0.0836437	-0.0633758	0.09641463	…	-0.0525962
x14	-0.0721822	-0.2936437	-0.1360887	0.06678213	-0.0746635	**0.3797631**	0.29027963	-0.0298938	-0.0293622	…	0.12190583
x15	-0.0755241	-0.329028	-0.1483329	0.09346712	-0.0129337	-0.2643278	-0.1129049	-0.0947934	0.06711795	…	-0.0167113
x16	-0.0554225	-0.0065527	0.12691303	0.36167517	0.22106113	0.29326329	0.3957758	-0.0871277	-0.0077882	…	-0.1219039
x17	-0.0587644	-0.0419373	0.11466873	0.38836029	**0.2827909**	-0.3508275	-0.0074089	-0.1520271	0.08869196	…	-0.0358849
x18	-0.0611368	0.12651284	0.1630995	-0.2392679	-0.219019	-0.0531925	0.0964203	-0.2780063	-0.0051438	…	-1.08E-06
x19	-0.0041465	0.07327472	0.14679462	-0.1685052	-0.3270836	-0.2896094	**0.4772772**	-0.187201	0.09665573	…	0.47129166
x20	0.03702178	-0.002869	0.03642736	-0.1268952	0.14555505	-0.0224742	0.095774	**0.7154298**	0.09063222	…	1.208E-06
고유값	1.36848152	0.42019731	0.36649931	0.3033658	0.24428594	0.17819868	0.16251832	0.15668328	0.10584064	…	-3.21E-08
누적비율	0.38373927	0.50156783	0.6043388	0.68940634	0.75790717	0.8078763	0.85344846	0.8973844	**0.9270634**	…	1

주성분 분석을 시행하여 얻은 p개의 고유값 λ_j 들을 크기 순으로 배열하고 각각의 고유값에 대응되는 고유벡터 e_j의 짝들은 $(\lambda_1, e_1),(\lambda_2, e_2),\cdots,(\lambda_p, e_p)$이다. 단, $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p$의 순서이다. 그리고, 첫 m($\leq$p)개의 주성분에 의해 설명되는 부분 즉, 고유값의 누적 비율은 아래의 식이 된다.

$$(\lambda_1 + \lambda_2 + \cdots + \lambda_m)/(\lambda_1 + \lambda_2 + \cdots + \lambda_p)$$

만약 첫 m개의 주성분들에 의해 설명되는 부분이 전체의 0.8~0.9를 점한다면 p보다 훨씬 작은 m개의 주성분들을 이용하더라도 정보상의 큰 손실이 없게 된다[9]. 본 논문에서는 누적 비율이 0.9 이상인 시점의 주성분까지를 이용한다. <표 3>의 경우를 보면, 9번째 주성분까지만을 고려하여 원자료를 표현하더라도 정보상의 큰 손실은 없게 된다.

주성분의 개수를 파악했다면, 이제 각 주성분이 어떠한 정보를 나타내는지를 알 필요가 있다. 예를 들어, 영어, 국어, 물리, 화학 성적에 대해 주성분 분석을 시행한다고 보자. 그렇다면, 영어, 국어 성적과 물리, 화학 성적으로 성적이 분리될 가능성이 높다. 즉, 언어 능력과 과학 능력이라는 두 가지의 주성분으로 구분될 가능성이 높다. 여기에서 언어 능력인지 과학 능력인지와 같이 주성분의 의미를 표현함에 있어 [7]에서는 주성분 적재계수가 0.5 이상인 명사 리스트를 이용한다. 만약, 0.5 이상인 명사가 없다면 주성분 적재계수가 가장 큰 값의 명사를 이용한다. 주성분 적재계수는 주성분에 대한 명사의 기여도이다. <표 3>에서 첫 번째 주성분 (PC1)에 대한 명사 x6의 주성분 적재계수는 0.9042802이다. 결국, 첫 번째 주성분을 표현하는 명사는 x6이라고 볼 수 있으며, 또한 x6 즉, "아파트"가 첫 번째 주성분에서 추출된 주제어가 된다. 이와 같이 9번째 주성분까지 감안하여 추출된 주제어는 <표 4>와 같으며, <표 2>의 총 20개의 명사 중에서 명사의 출현 빈도와 명사간의 공기 관계를 이용하여 8개의 명사 즉, 주제어가 추출된 결과이다. 그리고, 8개의 주제어 중 "아파트"라는 명사가 설명력이 가장 높다라고 볼 수 있다. 왜냐하면, 고유값이 가장 큰 첫 번째 주성분과 관련이 있기 때문이다.

〈표 4〉 주성분 분석을 이용하여 추출된 주제어 예

주성분	명 사	주성분	명 사
PC1	x6 "아파트"	PC6	x14 "센서"
PC2	x8 "주상"	PC7	x19 "실내"
PC3	x10 "주부"	PC8	x20 "모델"
PC4	x3 "싱크대"	PC9	x8 "주상"
PC5	x17 "모니터"		

3.3 비정칙치 분해를 이용한 중요 문장 추출

3.3.1 비정칙치 분해의 개요

이 절에서는 비정칙치 분해에 대한 개략적인 설명을 한다[11-13].

행과 열의 수가 각각 n과 p인 이원표 자료행렬을

$$X = (x_{ij}) = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix},$$

$$i=1,\cdots,n \ ; \ j=1,\cdots,p$$

라 하자. 각 열을 나타내는 변수들의 평균 $\bar{x}_{\cdot j} = \sum_{i=1}^{n} x_{ij}/n$을 뺀 새로운 자료행렬을 $Y = (x_{ij} - \bar{x}_{\cdot j})$라 하자. 만약에, 자료행렬 X의 열을 구성하는 변수간에 측정단위가 다른 경우 표준화 $((x_{ij} - \bar{x}_{\cdot j})/s_j$, s_j는 변수 j의 표준편차)된 새로운 자료행렬을 사용하게 된다. 아무튼, 원래 자료행렬 X 대신에 새로운 자료행렬 Y를 사용한다 하자.

계수(rank) r인 자료행렬 Y의 비정칙치분해는 다음과 같다.

$$Y = UD_\lambda V' = \sum_{k=1}^{r} u_k \lambda_k v_k'$$

여기서, 크기가 $n\times r$과 $p\times r$행렬 $U=(u_1,\cdots,u_r)$와 $V=(v_1,\cdots,v_r)$는 직교 열 벡터 $u_k=(u_{1k},\cdots,u_{nk})'$와 $v_k=(v_{1k},\cdots,v_{pk})'$, $k=1,\cdots,r$을 갖고 있다. 그리고 V'는 V의 전치행렬(transpose matrix)이다. 또한, 대각행렬 $D_\lambda = diag(\lambda_1,\cdots,\lambda_r)$는 $\lambda_1 \geq \cdots \geq \lambda_r > 0$ 관계를 갖는 비정칙치(singular value)를 대각원소로 하고 있다.

결국, 비정칙치 분해를 이용하여 임의의 $n\times p$ 행렬 (Y)을 $n\times r$ 행렬 (U), $r\times r$ 행렬 (D_λ), 그리고 $p\times r$ 행렬 (V)로 분해를 할 수 있다. 그리고, 자료 행렬의 계수(rank)가 $r=2$이면, 2차원 공간에 좌표점으로 모두 나타나게 되며, 원래의 행렬을 2차원으로 설명할 수 있는 비율로 변환할 수 있다는 뜻이기도 하다. 즉, 원래의 $n\times p$ 행렬은 $n\times 2$, 2×2, 그리고 $(p\times 2)'=(2\times p)$ 행렬의 곱을 이용하여 변형된 $n\times p$ 행렬로 변환될 수 있다.

3.3.2 비정칙치 분해를 이용한 중요 문장 추출 과정

제안한 방법은 중요 문장을 추출하기 위해서 단순히 주제어의 출현 여부만을 이용하지 않고, 주제어 벡터와 문장 벡터간의 유클리디언 거리를 계산한다. 이를 위해서 문장-명사 행렬에 대해, 문장 벡터와 명사 벡터를 구하기 위해서 먼저 비정칙치 분해를 시행한다. 비정칙치 분해를 시행한

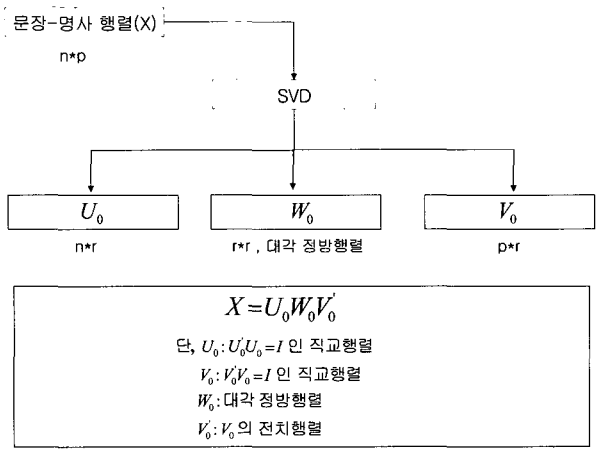
〈표 5〉 U_0 행렬 (17×20)의 예

문장/차원	1차원	2차원	3차원	4차원	5차원	6차원	7차원	8차원	9차원	10차원	11차원	…
1	-0.045686	0.047564	-0.361477	0.530133	-0.185135	0.489205	0.102892	0.437422	-0.09814	-0.003436	0.167933	…
2	-0.000015	0.000352	-0.005704	0.029066	-0.174737	-0.368363	0.198859	0.236789	0.01959	-0.149647	-0.333301	…
3	-0.000015	0.000352	-0.005704	0.029066	-0.174737	-0.368363	0.198859	0.236789	0.01959	-0.149647	-0.333301	…
4	-0.172899	-0.006093	0.037937	-0.044194	-0.001956	-0.0154	-0.006339	0.001471	-0.074276	-0.020921	0.174691	…
5	-0.23186	-0.004544	-0.880728	-0.281459	0.089424	-0.133533	-0.034329	-0.132083	0.000116	-0.171903	-0.057602	…
6	-0.226512	-0.008752	-0.098212	-0.123598	0.047288	-0.087601	0.023277	0.008199	0.265727	0.788582	0.125812	…
7	-0.827601	0.002243	0.225865	0.099543	0.030855	0.064511	0.072215	0.133159	0.237548	-0.084866	-0.204936	…
8	-0.000065	0.365809	0.012161	-0.077913	-0.0406	-0.003184	-0.237448	0.187375	0.027998	-0.026588	0.000775	…
9	-0.000133	0.843222	0.029807	-0.207866	-0.141955	0.151327	0.24202	-0.03543	0.008706	-0.015644	0.002368	…
10	-0.001386	0.183483	-0.008901	0.129179	0.125605	-0.082663	0.398959	-0.436962	-0.078464	0.080426	-0.009117	…
11	-0.001443	0.30694	-0.007898	0.150503	0.187968	-0.307056	-0.686145	0.182772	-0.010074	0.041095	-0.006113	…
12	-0.000401	0.001917	-0.024257	0.098946	-0.412912	-0.509291	0.16926	0.141426	0.005134	0.055202	0.592178	…
13	-0.036186	0.156622	-0.067659	0.647615	0.438765	-0.265616	0.097557	-0.225553	-0.02927	-0.014024	0.01383	…
14	-0.218066	-0.00613	0.087316	0.040952	-0.255575	0.032555	-0.234398	-0.42214	0.120925	-0.433591	0.31038	…
15	-0.010477	0.009743	-0.091747	0.278696	-0.626258	0.032592	-0.253651	-0.38911	-0.037835	0.278932	-0.385525	…
16	-0.358991	-0.014423	0.093715	-0.114353	-0.005568	-0.053079	-0.027573	0.007868	-0.715351	0.071586	0.152736	…
17	-0.013193	-0.002236	0.017842	-0.025965	-0.001656	-0.022278	-0.014895	0.004925	-0.566799	0.113427	-0.196646	…

〈표 6〉 V_0 행렬 (20×20)의 예

명사\차원	1차원	2차원	3차원	4차원	5차원	6차원	7차원	8차원	9차원	10차원	11차원	…
x1	-0.010574	0.020997	-0.181252	0.34793	-0.388451	0.283712	-0.089284	0.029977	-0.090407	0.215711	-0.460476	…
x2	-0.164417	0.018249	-0.054233	0.270865	-0.073861	0.301067	0.103704	0.354046	0.09269	-0.06914	-0.078308	…
x3	-0.015414	0.074812	-0.171618	0.506627	0.121425	0.12157	0.118712	0.131465	-0.084713	-0.013671	0.384653	…
x4	-0.052255	0.015762	-0.496778	0.106971	-0.045821	0.193386	0.040605	0.189463	-0.065175	-0.137289	0.233485	…
x5	-0.000081	0.00096	-0.014263	0.06757	-0.364988	-0.677486	0.335781	0.381611	0.029464	-0.191123	-0.157497	…
x6	**-0.9183**	**-0.0166**	**0.09486**	**-0.1027**	**-0.0041**	**-0.0283**	**-0.0107**	**0.00237**	**-0.1117**	**-0.0267**	**0.08255**	…
x7	-0.086299	-0.004871	-0.391495	-0.174241	0.06545	-0.120235	-0.006545	-0.07687	0.176754	0.482855	0.144348	…
x8	-0.087306	-0.00333	-0.704436	-0.242148	0.085622	-0.145209	-0.040661	-0.163915	0.000154	-0.269197	-0.2438	…
x9	-0.198461	-0.002385	0.051051	-0.010347	0.037411	-0.012554	0.056553	0.087713	0.334619	0.551004	-0.167445	…
x10	-0.162628	0.058207	0.06327	0.321402	0.224829	-0.109345	0.100544	-0.05733	0.13848	-0.07743	-0.404424	…
x11	-0.196871	-0.001424	0.125246	0.060436	-0.107584	0.052777	-0.096049	-0.179312	0.238342	-0.405947	0.223143	…
x12	-0.000284	0.246489	0.001705	0.031226	0.070552	-0.168684	-0.546978	0.229676	0.011917	0.011359	-0.011297	…
x13	-0.000062	0.751926	0.028704	-0.21235	-0.155358	0.162828	0.146038	0.072297	0.030192	-0.045316	0.011663	…
x14	-0.000286	0.376175	0.008361	-0.033848	-0.007827	0.037334	0.379606	-0.293119	-0.046381	0.050724	-0.014283	…
x15	-0.000297	0.421408	0.008762	-0.024676	0.022029	-0.084673	-0.263024	0.091426	-0.000909	0.019928	-0.007924	…
x16	-0.007074	0.124611	-0.030617	0.33415	0.27019	-0.189367	0.294051	-0.411091	-0.071631	0.051993	0.009974	…
x17	-0.007085	0.169845	-0.030216	0.343323	0.300046	-0.311374	-0.348579	-0.026545	-0.026159	0.021197	0.016332	…
x18	-0.002048	0.004272	-0.046392	0.162448	-0.497498	-0.259191	-0.049979	-0.153688	-0.021742	0.261624	0.437325	…
x19	-0.043029	0.001324	-0.001772	0.137501	-0.422174	0.035421	-0.289037	-0.503381	0.055245	-0.121096	-0.159025	…
x20	-0.070072	-0.006104	0.044613	-0.06036	-0.003458	-0.040973	-0.025151	0.007938	-0.85248	0.144863	-0.092923	…

후 구해지는 행렬은 (그림 1)과 같다 [12]. U_0 행렬은 n개의 문장과 새로운 r차원과의 관계를 나타내고, V_0 행렬은 p개의 명사와 새로운 r차원과의 관계를 나타낸다. 그리고 W_0 행렬은 차원이 설명할 수 비율이며, 첫 번째 값이 가장 크며 그 이후 값은 감소하는 경향이 있다.



(그림 1) 비정칙치 분해의 행렬

U_0 행렬의 행은 문장 벡터를 의미하고, V_0 행렬의 행은 명사 벡터를 의미한다. 이때, 모든 문장 벡터와 명사 벡터 중 주제어로 추출된 명사 벡터간의 유클리디언 거리를 계산하여 주제어와 문장간의 거리를 파악한다. 그리고, 중요 문장으로는 주제어와 거리가 가장 짧은 문장이 추출된다. 본 논문에서는 계수 (r)를 명사의 개수 p로 사용하였다. <표 1>의 문장-명사 행렬을 분해하여 생성된 U_0 행렬, V_0 행렬은 <표 5>, <표 6>과 같다.

명사 x6 "아파트"와 각 문장간의 거리는 다음 <표 7> 같으며, 거리가 가장 가까운 4번째 문장이 중요 문장으로 추출된다.

〈표 7〉 주제어와 문장간의 거리 예

문장 번호	x6과의 거리	문장 번호	x6과의 거리
문장 1	1.409127	문장 10	1.350946
문장 2	1.156406	문장 11	1.394477
문장 3	1.156406	문장 12	1.410824
문장 4	**0.778887**	문장 13	1.419945
문장 5	1.297652	문장 14	1.089496
문장 6	1.178203	문장 15	1.408876
문장 7	0.794114	문장 16	0.836959
문장 8	1.155381	문장 17	1.410556
문장 9	1.354743		

요약하면, 먼저 주성분 분석을 시행하여 요약 대상 문서의 주제어를 추출한다. 그런 후에 문장-명사 행렬에 대해 비정칙치 분해를 시행하여 문장-차원 행렬과 명사-차원 행렬을 획득한다. 이제, 명사-차원 행렬에서 주제어로 추출된 행 (명사 벡터)과 문장-차원 행렬의 모든 행 (문장 벡터)간의 유클리디언 거리를 계산한다. 그리고, 각 주제어에 대해 거리가 가장 가까운 한 문장씩을 중요 문장으로 추출하여 요약으로 제시한다.

4. 실험 및 평가

4.1 실험 자료

본 논문이 제안한 방법을 실험하기 위하여 KISTI (한국과학기술정보연구원)에서 제공하는 테스트 컬렉션을 사용하였다. 이 테스트 컬렉션은 두 명의 사람에 의하여 수동 요약된 신문기사 문서 집합 (1000건)으로 구성되어 있다. 테스트 컬렉션의 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 127건을 이용하여 문서 요약 실험을 하였다. 실험한 127건의 문서에 대한 통계적인 특성은 <표 8>과 같다.

〈표 8〉 실험 대상 문서(127건)의 통계적인 특성

대상 영역	신문기사
문서 개수	127 건
문서의 평균길이	19 문장
요약(#B)의 평균길이	5 문장
문장의 평균길이	7 개(명사)

4.2 실험 방법 및 평가

본 논문에서 제안한 방법을 출현 빈도만을 고려하는 방법과 주성분 분석만을 시행하는 방법 [7]과 비교하였다. 출현 빈도만을 고려하는 방법은 해당 문서에서 명사가 7회 이상 발생하면 주제어로 선정했다[14, 15].

다음의 (가) ~ (다)까지의 방법으로 실험을 하였다.

(가) 출현 빈도 7회 이상

(나) 주성분 분석만을 시행

(다) 제안한 방법

각 실험방법에 의해서 추출된 평균 주제어의 개수는 방법 (가)는 2개, 방법 (나)와 (다)는 7개였다. 그리고, 방법 (가)로 추출된 주제어 수는 문서에서 평균 요약의 길이인 5문장을 추출하기에는 다소 어려운 점이 있다고 볼 수 있다.

각 방법을 비교하기 위해, 각 방법으로 30% 추출한 요약과 KISTI 테스트 컬렉션에서 이미 제시된 #B(30% 추출요

약)를 비교하여 정확률과 재현율을 구한다. 정확률은 KISTI 테스트 컬렉션에서 제시된 요약 결과 문장 #B를 올바른 문장으로 간주하고 이와 일치하는 각 방법이 추출한 문장과 각 방법이 추출한 전체 요약 문장의 비를 이용하여 정확률을 구한다. 그리고, 재현율은 #B와 일치하는 각 방법이 추출한 문장과 실제로 올바른 요약문과의 비를 이용하여 구한다.

$$\text{정확률}(Precision) : \frac{\text{시스템이 제시한 올바른 요약문}}{\text{시스템이 제시한 요약문}}$$

$$\text{재현율}(Recall) : \frac{\text{시스템이 제시한 올바른 요약문}}{\text{올바른 요약문}}$$

$$F-measure : \frac{2PR}{P+R}$$

다음의 <표 9>는 각 방법의 실험 결과를 보이고 있다.

〈표 9〉 각 실험 방법의 비교

	(가)	(나)	(다)
평균 정확률	0.3553	0.3863	0.4068
평균 재현율	0.4127	0.4514	0.5001
F-measure	0.3819	0.4163	0.4487

<표 9>의 F-값에 의하면 제안한 방법 (다)가 방법 (가)보다는 약 6%, 방법 (나)보다는 약 3% 성능이 우수함을 알 수 있다.

5. 결 론

본 논문에서는 통계적 분석 기법인 주성분 분석과 비정칙치 분해를 이용한 문서 요약 방법을 제안하였다. 주성분 분석을 시행하여 주제어를 추출하고, 비정칙치 분해를 시행하여 명사 벡터와 문장 벡터를 획득한 다음, 명사 벡터 중 주제어로 추출된 벡터와 문장 벡터간의 유클리디언 거리를 계산하여, 거리가 최소인 문장을 중요 문장으로 추출하여 요약으로 제시하는 방법이다.

KISTI 테스트 컬렉션 중 127건의 문서에 대해 실험을 하였다. 그 결과 제안한 방법들이 출현 빈도 (7번 이상)만을 이용한 방법보다는 약 6%, 주성분 분석만을 이용한 방법보다는 약 3% 정도의 성능이 향상되었다.

제안한 방법은 주성분 분석과 비정칙치 분해를 이용하여 문서 자체내의 명사의 흐름을 수량화하였고, 다른 도구의 도움없이 문서 자체내에서의 명사의 발생 빈도와 명사간의 공기 정보를 이용하여 요약을 하고자 하는 방법이다.

제안한 방법처럼 통계적 정보만을 이용할 수도 있겠지만, 시소러스나 WordNet 등의 정보를 포함한다면 보다 성능이 향상될 수 있을 것으로 기대된다. 하지만, 시소러스나 WordNet 등을 사용함으로써 드는 비용과 성능 사이의 적절한 선택이 필요할 것이다.

참 고 문 헌

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

[2] 류동원, 이종혁, "단어공기정보를 이용한 자동화 문서 요약", 제27회 정보과학회 봄 학술발표논문집(B), 제27권, 제1호, pp.339-341, 2000.

[3] H. P. Edmundson, "New Methods in Automatic Extracting," Journal of the Association for Computing Machinery, Vol. 16, No.2.

[4] 강상배, 조혁규, 권혁철, 박재득, 박동인, "한국어 문서의 통계적 정보를 이용한 문서요약 시스템 구현", 제9회 한글 및 한국어정보처리학술대회, pp.28-36, 1997.

[5] 이창범, 박혁로, "시소러스를 이용한 문서 자동요약," 정보과학회 춘계학술발표논문집(B), 2001.

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

[7] 이창범, 김민수, 이기호, 이귀상, 박혁로, "주성분분석을 이용한 문서 주제어 추출," 정보과학회논문지 : 소프트웨어 및 응용, 29(9), pp747-754, 2002.

[8] 박혁로, 신중호, 이강혁, "기계 번역을 위한 정렬 코퍼스 작성 및 한글 어절 분석기 개발에 관한 연구", 연구개발정보센터 연구보고서, 1996.

[9] 김기영, 전명식, "다변량 통계자료분석", 자유아카데미, 1994.

[10] Richard A. Johnson, Dean W. Wichern, "Applied Multivariate Statistical Analysis," Prentice Hall, 1992.

[11] 최용석, "행렬도의 이해와 응용", 자유아카데미, 1999.

[12] William H. Press, Saul A. Teukolsky, et al., "Numerical Recipes in C++," Cambridge University Press, 2002.

[13] Scott Deerwester, Susan T. Dumais, Richard Harshman, "Indexing by Latent Semantic Analysis," Journal of the American Society for Information Science, 41(6), pp.381-407, 1990.

[14] 김동현, 이승우, 이근배, "중요 문장추출 휴리스틱과 MMR를 이용한 질의기반 문서요약", 제14회 한글 및 한국어 정보처리 학술발표논문집, pp.285-291, 2002.

[15] Anastasios Tombros and M. Sanderson, "Reflecting user information needs through query baised summaries," SIG IR'98, 1998.

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

[17] Jose Abracos, Gabriel Pereira Lopes, "Statistical methods for retrieving most significant paragraphs in newspaper articles," Proc. Association for Computational Linguistics, pp.51-57, 1997.
[18] 김영택 외, "자연언어처리", 생능출판사, 2001.
[19] 장동현,맹성현, "자동 요약 시스템", 정보과학회지, 제15권 제10호, pp.42-49, 1997.
[20] 우선미, 유춘식, 김용성, "용어 연관성 분석을 이용한 사용자 위주의 문서순위결정 기법", 정보과학회논문지 : 소프트웨어 및 응용, 제28권 제2호, pp.149-156, 2001.

<부 록>

● KISTI에서 제공된 원문

#T
인간 복제실험의 충격(사설)

#S
미국 조지 워싱턴 대학 메디칼 센터 연구팀이 지난 13일 사상 처음으로 인간의 배자(엠브리오)를 복제하는데 성공했다. 같은 유전인자를 가진 인간을 여럿 만들어낼 수 있는 의학기술을 개발한 것이다. 이것은 발전을 위한 인간의 연구노력이 거둔 획기적인 성과이긴 하지만 아울러 심각한 윤리문제를 야기하고 있다.
그러나 문제는 이런 실험이 앞으로 과연 어떤 결과를 초래할 것인가 하는 것이다. 복제인간이나 인간의 장기를 확보하기 위한 일란성 쌍생아의 대량 생산이라는 무서운 사태를 야기할 수 있다는 우려다. 같은 모습을 가진 사람이 무수히 출현함으로써 생겨날 혼란의 문제도 간과할 수 없다. 프랑켄슈타인과 같은 괴물의 출현에 대한 우려도 있다.
……(중략)
이같은 우려는 이미 70년대 생명공학의 초창기부터 제기된 바 있다. 그리고 이것은 인간의 과학적 탐구의 제한문제와 함께 윤리문제에 대한 논란을 일으키고 있다. 과학자 자신의 책임의식과 윤리적 태도가 선결되어야 하는 것은 물론이다. 하지만 그에 앞서 여러 규제장치도 고려되어야 할 것이다. 생명과학에 대한 관심이 높아가고 있는 우리도 이같은 문제를 미리 심각하게 검토 대비할 필요가 있다.

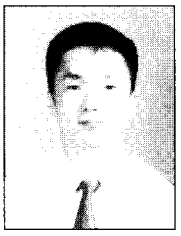
#A
미국 조지 워싱턴 대학 메디칼 센터 연구팀이 지난 13일 사상 처음으로 인간의 배자 (엠브리오)를 복제하는데 성공했다. 이것은 발전을 위한 인간의 연구노력이 거둔 획기적인 성과이긴 하지만 아울러 심각한 윤리문제를 야기하고 있다.

#B
① 미국 조지 워싱턴 대학 메디칼 센터 연구팀이 지난 13일 사상 처음으로 인간의 배자 (엠브리오)를 복제하는데 성공했다.
② 이것은 발전을 위한 인간의 연구노력이 거둔 획기적인 성과이긴 하지만 아울러 심각한 윤리문제를 야기하고 있다.
③ 이렇게 인간 복제실험에 대한 우려는 세가지 점에 초점이 맞춰지고 있다.
④ 하나는 인간이 이런 실험을 감히 할 수 있느냐는 원칙에 대한 것이며, 둘째는 실험이 인권존중의 차원에서 인류에 봉사하는 방향으로 이뤄지는가이며, 셋째는 실험결과에 대한 예측 불가능성이다.
⑤ 그러나 문제는 이런 실험이 앞으로 과연 어떤 결과를 초래할 것인가 하는 것이다.
⑥ 복제인간이나 인간의 장기를 확보하기 위한 일란성 쌍생아의 대량 생산이라는 무서운 사태를 야기할 수 있다는 우려다.

#C
미국서 사상 처음으로 인간의 배자 복제에 성공. 획기적인 성과이나 아울러 심각한 윤리문제 야기. 실험이 앞으로 초래하게 될 무서운 사태에 대한 우려가 생겨나고 있음.

● 제안한 방법으로 요약한 결과

① 미국 조지 워싱턴 대학 메디칼 센터 연구팀이 지난 13일 사상 처음으로 인간의 배자 (엠브리오)를 복제하는데 성공했다.
② 이것은 발전을 위한 인간의 연구노력이 거둔 획기적인 성과이긴 하지만 아울러 심각한 윤리문제를 야기하고 있다.
③ 일부 학자들은 이같은 실험이 과거 게르만 민족의 우수한 혈통만을 창조하려던 나치의 잔혹한 실험과 다를바 없으며, 결국 「사악한 우생학」의 폐해를 초래할 것이라고 경고한다.
④ 이렇게 인간 복제실험에 대한 우려는 세가지 점에 초점이 맞춰지고 있다.
⑤ 그러나 문제는 이런 실험이 앞으로 과연 어떤 결과를 초래할 것인가 하는 것이다.
⑥ 복제인간이나 인간의 장기를 확보하기 위한 일란성 쌍생아의 대량 생산이라는 무서운 사태를 야기할 수 있다는 우려다.



이 창 범
e-mail : chblee@empal.com
1995년 전남대학교 전산학과(학사)
1995~1999 대우정보시스템(주)
2001년 전남대학교 대학원 전산학과(이학석사)
2002년~현재 전남대학교 대학원 전산학과 박사과정
관심분야 : 정보검색, 자연어처리, 문서요약

김 민 수
e-mail : mskim@ai.kaist.ac.kr
1994년 전남대학교 전산학과(학사)
1996년 전남대학교 전산통계학과(이학석사)
2000년 전남대학교 전산통계학과(이학박사)
2000년~2002년 전남대학교 BK사업단 박사후 연구원
2003년~현재 한국과학기술원 AIPR Lab. 박사후 연구원
관심분야 : 통계적 패턴인식, 다변량 통계분석, wavelets

백 장 선

e-mail : jbaek@chonnam.chonnam.ac.kr

1981년 연세대학교 응용통계학과(학사)

1984년 연세대학교 대학원 응용통계학과 (이학석사)

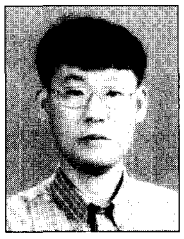
1991년 미국 Texas A&M 대학교 대학원 통계학과(이학박사)

1991년~1993년 미국 Southern Methodist University 통계학과 Postdoctoral Fellow

1993년~1995년 전남대학교 통계학과 전임강사

1995년~현재 전남대학교 통계학과 부교수

관심분야 : Nonparametric Function Estimation, Multivariate Analysis



박 혁 로

e-mail : hyukro@chonnam.ac.kr

1987년 서울대학교 전산학과(학사)

1989년 한국과학기술원 전산학과(공학석사)

1997년 한국과학기술원 전산학과(공학박사)

1994년~1996년 연구개발정보센터 연구원

1997년~1998년 연구개발정보센터 선임 연구원

1999년~2002년 전남대학교 전산학과 조교수

2003년~현재 전남대학교 전산학과 부교수

관심분야 : 정보검색, 자연어처리, 데이터베이스