The Korean Communications
in Statistics Vol. 9, No. 2, 2002
pp. 491-503

A Automatic Document Summarization Method based on Principal Component Analysis[1)]

Minsoo Kim[2)], Changbeom Lee[3)], Jangsun Baek[4)], Gueesang Lee[5)], Hyukro Park[6)]

Abstract

In this paper, we propose a automatic document summarization method based on Principal Component Analysis(PCA) which is one of the multivariate statistical methods. After extracting thematic words using PCA, we select the statements containing the respective extracted thematic words, and make the document summary with them.

Experimental results using newspaper articles show that the proposed method is superior to the method using either word frequency or information retrieval thesaurus.

Keywords : principal component analysis, document summarization, thematic word extraction

1. 서론

현재 우리들이 살고 있는 시대를 인터넷의 시대라고도 한다. 인터넷의 발달과 급속한 보급으로 쏟아지는 정보는 주체할 수 없을 정도이다. 또한 인터넷의 급속한 보급과 함께 이에 따른 정보량의 폭증으로 정보의 90% 이상을 차지하는 텍스트 정보에 대한 검색 및 정리에 시간과 비용이 급증하고 있다. 텍스트 정보에 대한 검색의 자동화 및 검색정보의 요약 및 정리 등의 정보 필터링 시스템의 개발은 정보획득의 신속성 및 정확성과 효율적 이용을 가능하게 함으로 정보화 사회에서의 경제, 산업의 경쟁력 강화에 필수적이다.

현재 정보의 대부분을 차지하며 기하급수적으로 증가하고 있음에도 불구, 텍스트 정보탐색은 여전히 단순 명사 중심의 키워드 방식의 전통적 방법의 검색기술에 의존하고 있어서 사용자가 필요로 하는 정보의 명확한 표현이 제한되어 왔고, 수집되어 있는 텍스트 형태의 정보 역시 수치 데이

1) This study was financially supported by Chonnam National University in the program, 1999
2) Lecturer, Department of Statistics, Chonnam National University, Kwangju, 500-757, Korea.
E-mail : kimms@chonnam.chonnam.ac.kr
3) Graduate Student in the doctoral program, Department of Computer and Information Science, Chonnam National University. Kwangju, 500-757, Korea.
4) Associate Professor, Department of Statistics, Chonnam National University, Kwangju, 500-757, Korea.
5) Professor, Department of Computer and Information Science / Information and Telecommunication Research Institute, Chonnam National University, Kwangju, 500-757, Korea.
6) Assistant Professor, Department of Computer and Information Science / Information and Telecommunication Research Institute, Chonnam National University, Kwangju, 500-757, Korea.

터에 전적으로 의존하는 시스템에 의해 제공되고 있으므로 담고 있는 내용 및 성격을 기준으로 한 사용자의 접근과 검색이 불가능한 상태이다.

문서요약은 정보 사용자에게 문서의 내용을 축소된 형태로 제시함으로써 사용자가 정보의 유용성을 판단하는 시간을 단축시켜 짧은 시간에 더 많은 자료를 검토할 수 있게 함으로써 정보 접근의 효율성을 향상시키는 역할을 한다. 또한 정보 검색에 있어서 문서의 주변적인 부분을 제외하고 주요 내용만 색인으로 사용함으로써 정보검색의 검색 속도를 향상시킬 수 있으며, 정보 검색 결과에 대한 사용자의 만족도를 향상시킬 수 있기 때문에 전자도서관, 정보검색, 전자상거래 등의 분야에 필요한 현대 정보 시스템의 필수적인 중요 기능으로 부상하고 있다.

문서요약은 그 생성 방법에 따라 추출요약(extract)과 생성요약(abstract)으로 구분될 수 있다(류동원, 이종혁, 2000). 문서를 요약하는 과정은 문서의 주된 내용을 파악하는 단계와 파악된 주된 내용을 요약으로 생성하는 단계을 포함하고 있다. 컴퓨터에 의한 문서자동요약 과정도 사람의 경우와 같이 주요 내용을 파악하여 요약을 생성하는 하는 것이 자연스럽고 바람직하지만, 문서의 내용을 추론 가능한 의미있는 표현으로 바꾸고, 이러한 표현을 근거하여 요약을 생성하는 것은 현재의 언어처리 기술로는 어려운 상황이다. 즉, 현재의 언어처리 기술로는 문서의 요약을 생성하기에는 어려운 실정이다. 그래서, 추출요약 즉, 문서에서 중요한 문장을 선택하여 문서에서 나타나는 순서대로 문장을 보여주는 요약방법이 주된 경향을 이루고 있다. 하지만, 컴퓨터에 의한 정보 자동 요약 과정도 사람의 경우와 같이 주요 내용 파악과 자연스러운 요약 생성의 단계를 거쳐서 수행되는 것이 바람직하나, 문서의 내용을 추론 가능한 의미적인 표현으로 변환하고, 이 표현으로부터 문장을 생성하는 것은 현재 언어처리기술로는 매우 어려운 작업이다. 따라서 현실적인 자동 요약 시스템은 문서로부터 그 내용을 대표할 수 있다고 판단되는 문장들을 추출하여 요약으로 사용하는 문장 추출 시스템이 중심을 이루고 있다(장동현, 맹성현, 1997).

문장이 문서의 내용을 대표할 수 있는가 혹은 어느정도 대표하는가를 계산하는 문제는 문서의 내용을 이해하는 것을 전제로 하고 있으나, 현재의 언어처리기술 수준으로는 문서의 내용을 컴퓨터가 이해하는 것은 매우 어려운 일이다. 따라서 대부분의 자동요약 시스템에서는 문서의 내용을 이해하지 않고, 여러 가지 단서에 의해 문장의 중요도 혹은 대표성을 계산하고 있다(박혁로, 신중호, 1999).

일반적으로 말하는 문서자동요약이란 입력된 문서에 대해 컴퓨터가 자동으로 요약을 생성하는 과정으로써 컴퓨터가 문서의 기본적인 내용을 유지하면서 문서의 복잡도 즉 문서의 길이를 줄이는 작업을 말한다(Kupec, Pedersen, Chen, 1995).

일반적인 검색엔진들은 문서의 제목과 앞부분을 약간만 보여주어 이 문제를 해결하려 하지만, 이 정도의 정보는 사용자가 검색 결과 문서의 적합성을 판단하기에 부족하다. 자동 문서요약시스템은 사용자가 원하는 정보를 찾아내는데 걸리는 시간을 단축시킴으로써 정보과적재 문제에 대해 효과적인 해결책을 제시해 줄 수 있다(Tombros, Sanderson, 1998). 정보의 효율적인 접근과 정보과적재 문제를 해결하기 위한 방안으로 문서 자동요약에 관한 연구가 활발히 진행되고 있다.

문서요약은 그 생성 방법에 따라 추출요약(extract)과 생성요약(abstract)으로 구분될 수 있다(류동원, 이종혁, 2000). 문서를 요약하는 과정은 문서의 주된 내용을 파악하는 단계와 파악된 주된 내용을 요약으로 생성하는 단계을 포함하고 있다. 컴퓨터에 의한 문서자동요약 과정도 사람의 경우와 같이 주요 내용을 파악하여 요약을 생성하는 하는 것이 자연스럽고 바람직하지만, 문서의 내용을 추론 가능한 의미있는 표현으로 바꾸고, 이러한 표현을 근거하여 요약을 생성하는 것은 현재의 언어처리 기술로는 어려운 상황이다. 즉, 현재의 언어처리 기술로는 문서의 요약을 생성하기에

는 어려운 실정이다. 그래서, 추출요약 즉, 문서에서 중요한 문장을 선택하여 문서에서 나타나는 순서대로 문장을 보여주는 요약방법이 주된 경향을 이루고 있다.

본 논문에서는 다변량 통계분석 기법중의 하나인 주성분분석(principal component anlaysis)을 이용하여 문서의 주제어를 추출하고, 추출된 주제어를 기반으로 문장을 추출하는 새로운 방법을 제안한다. 본 논문에서 제안한 방법은 주성분분석을 이용하여 주제어를 선정한 다음, 이 주제어들과 밀접한 관련이 있는 문장들을 그 관련 정도 즉, 중요도에 따라 문장을 선택하고, 선택된 문장들을 문서내의 순서에 맞게 재배치되어 요약이 생성되는 방법이다.

본 논문의 구성은 다음과 같다. 제 2장에서는 문서요약에 관련된 기존 연구를 살펴보고, 제 3장에서는 제안된 방법에 대해 설명한다. 그리고 제 4장에서는 실험 및 평가, 제5장에서는 결론을 기술한다.

2. 관련연구

2.1. 개요

기존의 문서에 대한 요약 생성에 관한 연구들은, 크게 단어의 출현빈도에 기반한 방법, 지식에 기반한 방법, 문맥구조에 기반한 방법 그리고 수사구조에 기반한 방법으로 분류할 수 있다.

단어의 출현빈도에 기반한 방법은 가장 고전적인 방법으로 요약하고자 하는 문서에 나타난 단어의 빈도를 측정하여 대표하는 단어 집합을 설정한 후, 이를 기반으로 적당한 요약문을 생성하는 방법이다(Edmundson, 1969; Kupiec, 1995; 강상배, 조혁규, 권혁철, 박재득, 박동인,1997). 이 방법은 단어의 출현빈도만을 고려하기 때문에 문서에 나타나는 문장들 사이의 관계나 문맥 구조를 제대로 표현하지 못하여, 이로 인하여 일관된 논리를 갖는 문장들을 생성하기가 어려운 단점이 있으나 요약문을 일정한 비율로 요약을 할 수 있다는 장점이 있다.

지식에 기반한 방법은 요약문을 생성하고자 하는 문서와 관련된 배경 지식을 이용하여 요약문을 생성하는 방법이다(Barzilay and Elhadad, 1997; Hovy and Lin, 1997). 이 방법은 문서를 분석하기 위해서 많은 배경지식을 필요로 하는데, 제한된 영역에서의 문서요약에는 뛰어나지만 다양한 주제를 갖는 일반적인 문서들에 대한 요약문 생성에는 어려운 점이 있다. 예를들어, Barzilay 와 Elhadad(1997)에서는 워드넷(WordNet)을 이용하여 같은 개념을 갖는 단어들을 사슬로 만들어, 즉 어휘 사슬(lexical chain)를 구성하여 강력한 사슬이 있는 문장을 선정하여 요약문을 생성한다. 하지만, 이 방법은 주제를 찾아가기 위해 사전에 수립된 정보 즉, 워드넷이라는 지식을 이용한다. 만약, 이러한 정보를 이용하기가 힘든 분야에서는 그 응용에 한계가 있다

문맥구조에 기반한 방법은 지식에 기반한 방법이 관련된 배경지식에 많은 제약을 받기 때문에 이를 해결하기 위하여 배경 지식과 무관하게 요약문을 생성하고자 하는 방법이 고안되었는데 이 방법이 문맥 구조에 기반한 방법이다. 이 방법은 문장들 사이의 문맥 관계를 파악하여 요약문을 생성하는 방법으로 지금까지 많은 연구가 이루어져 왔지만, 대부분이 언어 외적인 지식을 이용해야 한다든지, 문장들 사이의 관계를 설정하는 방법이 모호하다든지 해서 실제 상용 시스템에 적용하기 어려운 점이 있다.

기존의 문맥구조에 기반한 방법이 당면한 문제점들을 해결하기 위해 제안된 방법이 수사구조에 기반한 방법인데, 이 수사구조에 기반한 방법에서는 문서에 나타나는 수사어구(rhetorical pattern)로부터 문장들 사이의 수사관계(rhetorical relation)를 설정한 후, 이들 각각을 연결하여 하나의 수

사구조를 만들어냄으로써 요약문을 생성하는 방법이다(양기주, 1997). 이 방법은 미리 규정된 수사 관계를 바탕으로 문장들 사이의 문맥관계를 파악하여 이를 요약문 생성에 그대로 반영하므로 생성된 문장들 사이의 논리가 명확하다. 하지만, 입력 문장으로부터 수사어구를 추출하는데 있어 형태소 분석이나 구문 분석과 같은 실질적인 언어처리를 하지 않고 단순히 패턴매칭만을 사용한다. 그래서, 문장의 첫머리가 아닌, 문장의 중간이나 문장의 끝에 오는 수사어구는 처리가 불가능하다.

위와 같이 문서자동요약방법들은 장단점을 가지고 있으므로 방법들이 혼합되어 쓰여지는 경우가 많다. 그리고 최근에 개발된 방법 중 단어의 출현빈도에 기반한 방법과 지식에 기반한 방법의 혼합형태인 시소러스(thesaurus)를 이용한 방법이 있다. 이 방법은 의미기반 정보검색용 시소러스로부터 단어간의 관계 즉, 동의어, 유사어, 상위어, 하위어들에 대해 일정부분의 가중치 점수를 주어 단어 출현빈도에 추가적인 점수를 부여하는 방법이다.

본 논문에서는 다변량 통계분석 기법중의 하나인 주성분분석(principal component analysis)을 이용하여 문서를 요약하는 방법을 제안한다. 제안하는 방법은 워드넷 등과 같은 다른 도구를 이용하지 않고, 오직 문서 자체내의 단어들의 분산공분산구조로부터 얻어진 고유시스템에 기반한 방법이다.

2.2. 시소러스를 이용한 주제어 추출

자연어 처리 시스템에서는 같은 주제라도 문헌 생산자나 색인 작성자, 이용자 간에 그 표현하는 용어가 달라질 수 있어 문헌의 분석이나, 색인 작성시에 많은 어려움이 야기된다. 따라서 필요한 정보를 찾으려고 하는 이용자는 하나의 검색어만으로 해당 주제를 전부 검색할 수 없으므로 그 검색어에 관련된 개념의 상위어, 하위어, 관련어 등을 모두 검색하여야 하는 번거로움이 있다. 이에 그 해당하는 주제 분야에서 필요한 모든 개념을 수집하여 이들에 대한 개념의 대소관계나, 동의어, 동형이의어, 관련어 등을 적절히 조절하여 정보시스템과 문헌 생산자, 색인 생산자, 이용자 간에 통일적으로 사용할 수 있도록 통제하여둔 용어통제어표를 시소러스(thesaurus)라 한다. 사람이 자연어 문장을 이해하는 경우에는 각자가 가진 상식이나 지식, 단어 개념 등의 지식베이스를 이용한다. 문서의 주제어들을 추출하는데 있어 이러한 지식베이스를 이용한다면 보다 효과적일 것이다. 예를 들어, 어떤 문서에 "국민", "민족", "종족", "인민", "국가", "공화국" 등의 단어가 출현한다고 하자. 이들 단어들을 개별적으로 이용하기보다는 이들 단어들이 서로 연관이 있다는 정보를 이용한다면 그 문서의 주제를 파악하는데 더 유용하다. 사실, 이들 단어들은 사용되는 시소러스에서 유의어 관계로 설정되어 있고, "일정한 영토에 살며 독립된 통치 조직을 가지는 다수인의 사회 집단"이라는 의미를 가진다.

본 논문에서 사용되는 시소러스는 단어의 의미와 단어간의 연관 관계를 포함하고 있다. 단어의 의미 정보를 사용하기 위해서는 먼저 문장의 의미분석 단계가 필요하다. 하지만, 현실적으로 이러한 의미분석을 사용하기에는 아직까지는 미비한 상태이다. 따라서, 본 논문에서는 단어간의 연관 관계 즉, 동의어, 유의어, 하위어, 상위어 등의 관계를 이용하여 문서의 주제어를 추출한다.

시소러스를 이용하여 주제어를 추출하는 방법은 다음과 같이 네 단계로 나누어 설명할 수 있다.

첫째, 문서에서 문장을 추출하고, 각 문장들에 대해 명사를 추출한다. 문장은 문미 기호가 종결어미와 함께 나오는 경우를 말한다. 그리고, 각 문장에서 형태소 해석과 태깅(tagging) 과정을 실행하여 명사를 추출한다.

문미 기호 (5개)	'!', '.', '?', ';', ':'
종결 어미 (19개)	"다","니","요","까","오","함","음","것","말","야","당","용","유","어","엉","응","지","와","쥐"

[표 2.1] 문미 기호 및 종결 어미

둘째, 어휘 사슬을 형성한다. 어휘 사슬을 형성하기 위해 사용된 연관 관계는 단어의 반복, 동의어, 유의어, 상위어, 하위어 관계이다. 추출된 명사들간에 연관 관계가 형성된다면 그들 사이에 링크를 형성하고, 그렇지 않으면 다른 명사와 비교를 계속한다. 이러한 과정을 추출된 모든 명사를 처리할 때까지 반복한다.

셋째, 형성된 어휘 사슬에 대한 점수를 계산한다. 문서의 주제를 가장 잘 표현하는 사슬을 찾는 단계이다. 이를 위해서 연관 관계들 간에 점수에 차등을 주어서, 형성된 사슬의 전체 점수를 계산한다.

3. 제안된 방법

본 논문에서 제안된 방법은 먼저, 문장에 출현하는 명사들을 변수로 고려하고 각 문장을 관측치로하여 각 문장에서 출현한 빈도값을 산출하고 주성분분석을 시행한다. 다음으로, 고유값(eigen value)의 누적비율이 0.9이상이되는 시점까지의 주성분 m개를 선택한다. 그런 후에, 선택된 m개의 주성분에 대하여 주성분적재계수가 0.5이상인 변수들, 즉명사들을 주제어로 선택한다. 만약, 어떤 주성분에 대하여 주성분적재계수가 0.5이상인 경우가 없으면 그 주성분에 대해서는 최대 주성분적재계수를 갖는 변수를 주제어로 선택한다. 이렇게 선택된 주제어들을 기반으로 각 문장의 중요도를 계산하여, 그 중요도 순으로 문장을 추출하여 요약문을 생성한다.

어휘 사슬을 이용하는 방법은 다른 도구 즉, 워드넷을 사용하여 문서가 표현하고자 하는 내용을 파악하며, 시소러스를 이용한 방법 역시, 단어간의 연관관계를 위해 데이터베이스를 이용한다. 이에 반해 제안된 방법은 다른 도구의 힘을 빌리지 않고, 단지 문서자체에서의 정보 또는 문서자체내의 명사들의 흐름 혹은 유사도를 명사들의 분산공분산 행렬의 고유시스템에 의해 정량화하고자 하였다

3.1. 주성분분석에 의한 주제어 선정

여러 개($p \geq 2$)의 변수들에 대하여 얻어진 다변량 자료를 분석대상으로 하는 주성분 분석 기법은 다차원적인 변수들을 축소, 요약하는 차원의 단순화와 더불어 일반적으로 서로 상관되어 있는 변수들 간의 복잡한 구조 및 관계를 분석하는데 그 목적을 두고 있다. 이를 위하여 주성분 분석은 공분산행렬의 고유시스템에 의해 변수들을 선형변환하여, 주성분이라고 부르는 서로 상관되어있지 않은, 혹은 독립적인 새로운 인공변수들을 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(김기영, 전명식,1994).

본 논문에서는 문서의 정보로써 문서에 출현하는 명사들만을 이용하였다. 그래서, 이 명사들을 변수로 간주하여, 명사들의 정보손실을 최소화하는 소수의 몇 개 명사군들을 추출하는데 주성분 분석을 이용하므로써 자동문서요약을 실행하고자 하였다.

문서요약을 위한 주성분분석에서 사용되는 자료 구조는 [그림 3.1]와 같다.

변수(명사) / 문장번호	X_1	X_2	⋯	X_p
1 2 ⋮ n	(n × p) 자료			

[그림 3.1] 주성분분석의 자료 구조

여기에서 변수 $X_1, X_2, \cdots, X_p$를 문서에 2번 이상 출현하는 명사로, 개체 $1,2,\cdots,n$ 을 문서의 각 문장으로, 그리고 그에 해당하는 자료를 문장에서 출현하는 명사의 누적 빈도수로 고려한다. 그렇다면, 이러한 자료를 이용하여 주성분분석을 수행할 수 있고, 분산공분산행렬의 고유값과 고유벡터를 구해낼 수 있다. 여기에서, p 개의 고유값 λ_j들을 크기순으로 배열하고 각각의 고유값에 대응되는 고유벡터 e_j의 짝들은 $(\lambda_1, e_1), (\lambda_2, e_2), \ldots, (\lambda_p, e_p)$이다. 이 때, $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p$ 의 순서이다.

첫 $m(\leq p)$개의 주성분에 의해 설명되는 부분, 즉 누적비율 ρ_m은

$$\rho_m = (\lambda_1 + \lambda_2 + \ldots + \lambda_m)/(\lambda_1 + \lambda_2 + \ldots + \lambda_p)$$

와 같이 계산할 수 있다.

본 연구에서는 주제어를 선정하기 위하여, 먼저 고유값의 누적 비율이 0.9(90%)이상이 되는 시점에서 주성분의 개수를 결정한다. 그리고 나서, 주성분적재계수의 값이 0.5이상인 경우의 명사들을 주제어들로 선정한다. 만약, 주성분적재계수의 값이 0.5이상인 경우가 없는 경우에는 주성분 적재계수 중에서 최고치를 갖는 명사를 선택한다.

명사들의 공분산구조의 90% 이상 설명할 수 있는 첫 m개의 주성분들을 선택하고, 각각의 주성분에서 상관도가 0.5 이상인 명사를 주제어로 선택한다. 결과적으로, 주성분분석을 이용하여 보다 많은 문장에서, 함께 출현하는 명사들을 문서의 주제어로 선정한다.

3.2. 주성분분석을 이용한 주제어 선정의 예

여기서는 부록A.에 제시한 "임기중 개헌없다/김대통령 취임 100일 회견"라는 신문기사를 대상으로 주성분분석을 이용한 주제어 추출 과정을 보이고자 한다. 설명력이 90%이상인 첫 m개의 주성분을 선택하였고, 선택된 주성분과의 상관도가 50%이상인 단어들을 주제어로 추출한다. [표 3.1]은 실험 대상 문서에서 2번 이상 출현한 단어(변수) 리스트이다. 그리고 [표 3.2]은 주성분분석에 이용한 자료구조를 보여주고 있다. [표 3.2]의 값은 각각의 문장을 개체로 보고, 변수가 발생한 누적 빈도수를 나타내고 있다. [표 3.3]는 [표 3.2]의 자료구조를 이용하여 주성분분석한 결과를 보여준다.

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

[표 3.1] 주성분분석에 이용한 변수 리스트의 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 3.2] 주성분분석에 이용한 자료 구조의 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%

[표 3.3] 주성분분석 결과의 예

[표 3.3]의 결과에 기초하여, 누적비율이 90%이상의 시점인 'PRIN4'까지의 주성분 4개를 선택한다. 그리고 'PRIN1'부터 'PRIN4'까지의 주성분적재계수의 값이 0.5이상인 모든 변수를 택한다. 만약, 주성분계수의 값이 0.5이상인 경우가 없다면 주성분적재계수 중 최고치의 변수를 택한다. [표 3.4]는 설명력이 90%이상인 주성분을 이용하여 "임기중 개헌없다/김대통령 취임 100일 회견"라는 신문기사의 주제어를 추출한 내용이다.

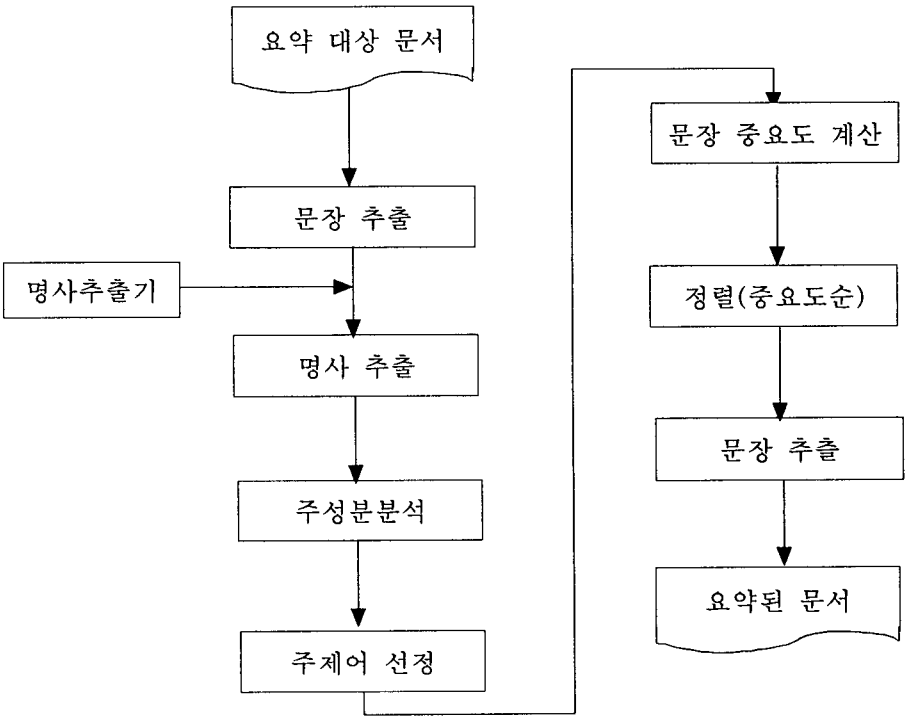
주성분	변수	변수값
PRIN1	X1	대통령
PRIN2	X3	국가
PRIN3	X7	국민
PRIN4	X7	국민
전체(주제어)	X1, X3, X7	대통령, 국가, 국민

[표 3.4] 주성분분석을 이용하여 추출된 주제어 예

3.3. 제안된 문서자동요약 방법

본 논문에서 제안된 방법의 전체적인 흐름도는 [그림 3.2]와 같다.

선정된 주제어를 기반으로 문서를 요약하기 위해서 문장의 중요도를 계산하는 단계, 문장의 중요도를 정규화하는 단계, 문장을 추출하는 단계를 수행한다. 문장의 중요도를 계산하기 위해, 선정된 주제어가 해당 문장에 포함되는 경우에는 가산점을 주었다. 선정된 모든 주제어에 대해 비교를 하여, 해당 문장의 중요도를 계산하였다. 즉, 해당 문장이 주제어를 많이 포함되면 될수록 그 문장의 중요도는 높아진다. 만약, 문장의 길이 즉, 문장내의 명사의 개수가 많다면 그 문장의 중요도는 커지는 경향이 있다. 그 결과로, 문장의 길이가 긴 문장일수록 요약문에 포함될 가능성이 커진다. 이에 이러한 긴 문장 선호도를 해소하기 위하여, 계산된 각 문장의 중요도를 각 문장의 길이 즉, 명사의 개수로 나누어주었다. 이렇게 계산된 각 문장의 중요도를 크기순으로 정렬하여, 사용자가 원하는 만큼의 문장을 추출한다. 추출된 문장을 원래 문서에 나타나는 순서대로 사용자에게 요약으로 제시한다.



[그림 3.2] 제안된 문서자동요약방법의 흐름도

4. 실험 및 평가

실험 데이터로는 KISTI(한국과학기술정보연구원)에서 제공하는 테스트컬렉션을 사용하였다. 이 테스트컬렉션은 두 명의 사람에 의하여 수동 요약된 신문기사 문서집합(1000건)으로 구성되어 있다. 테스트컬렉션의 문서집합중에서 100건의 문서에 대해 실험을 하였다. 실험한 100건의 통계적인 특성은 [표 4.1]과 같다.

대상 영역	신문기사
문서 개수	100 건
문서의 평균길이	19.51 문장
요약의 평균길이	4.83 문장
문장의 평균길이	6.79 개(명사)
평균변수 개수	19.54 개(명사)

[표 4.1] 실험대상 문서집합의 통계적인 특성

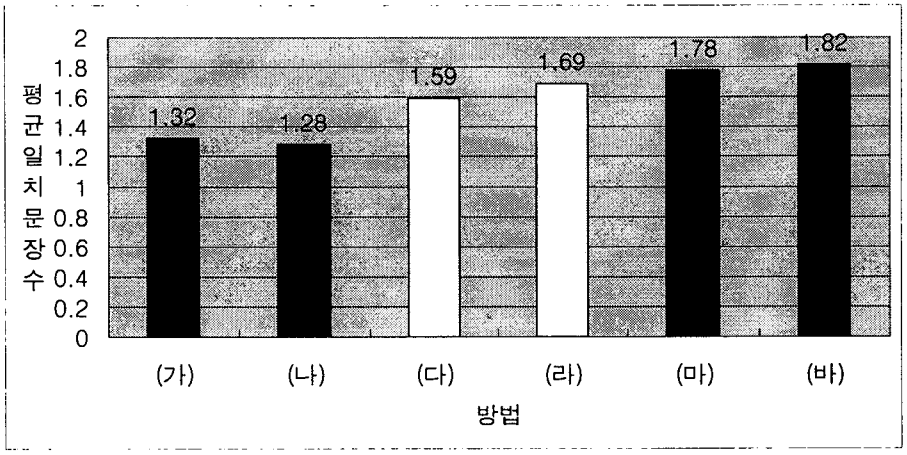
단순히 단어의 출현빈도만을 고려하여, 이를 기반으로 요약문을 생성하는 기존의 방법과 시소러스 방법을 본 논문에서 제안된 방법을 비교하기 위하여 다음과 같은 방법으로 실험을 하였다.

- (가) : 출현빈도 6번 이상의 명사를 주제어로 선택
- (나) : 출현빈도 2번 이상의 명사를 주제어로 선택
- (다) : 시소러스(가장높은점수)를 이용한 방법
- (라) : 시소러스(30점이상)를 이용한 방법
- (마) : 주성분분석을 이용한 방법(주성분적재계수 ≥ 0.3)
- (바) : 주성분분석을 이용한 방법(주성분적재계수 ≥ 0.5)

각 실험 방법의 성능을 비교하기 위해서 테스트컬렉션이 제시하는 30%추출요약과 각각의 실험방법으로 생성된 30% 요약을 비교하여, 실험으로 생성된 요약 문장이 테스트컬렉션의 요약에 포함되는 개수를 파악하였다.

[그림 4.1]은 각 실험 방법의 평균 일치 문장수를 나타낸다.

분산분석결과 유의수준 $\alpha = 0.05$에서 각 방법 간의 평균일치 문장수는 유의한 차이를 보였다. 또한 [그림 4.1]과 [표 4.2]에서와 같이 본 논문이 제시한 방법과 시소러스방법이 명사 출현빈도방법에 비해 우수함을 알 수 있다. 그런데, 시소러스방법은 데이터베이스를 이용하는 방법이므로 수행시간이 너무 많이 소요되어 실제적으로 사용되기는 어려움이 따른다는 것을 고려하면 주성분분석을 이용하는 방법이 현실적으로 기존의 방법들에 비해 우수하다고 여겨진다.



[그림 4.1] 각 실험 방법의 평균 일치 문장수

Duncan Grouping	방법	평균
A	(바)	1.82
A	(마)	1.78
A	(라)	1.69
A	(다)	1.59
B	(가)	1.32
B	(나)	1.28

[표 4.2] 실험방법에 대한 던컨(Duncan)의 다중비교결과

5. 결론

본 논문에서는 다변량 통계분석 기법중의 하나인 주성분분석을 이용한 문서자동요약 방법을 제안하였다. 제안된 방법은 문서에서 주제어를 추출한 후, 그 주제어를 이용하여 중요한 문장을 추출하여 요약으로 제시하였다. 주제어를 추출하기 위해서 주성분분석의 결과인 고유값과 고유벡터를 이용하였다. 또한, 워드넷과 같은 다른 도구를 이용하지 않고, 오직 문서 자체내의 정보만을 이용하여 주제어를 선정하였다.

제안된 방법에서는 긴 문장 선호도를 해소하기 위해서 문장의 길이를 감안하였고, 사용자가 원하는 비율만큼 요약을 생성할 수 있다. 그리고, 신문기사를 대상으로 실험한 결과 제안한 모형이 단어의 출현빈도만을 고려하는 기존의 방법보다 더 좋은 성능을 보였다.

제안된 방법에서는 다만 명사만을 고려하였지만, 동사나 형용사 같은 용언까지 고려한다면 더 좋은 성능을 보일 것이라 예상된다.

참고문헌

[1] Barzilay, R. and Elhadad, M.(1997). *Using Lexical chains for Text Summarization*, proc. Association for Computational Linguistics, 10-17.

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

[3] Hovy, E. and Lin, C. Y.(1997). Automated Text Summarization in SUMMARIST, *Proc. Association for Computational Linguistics*, 18-24.

[4] Kupiec, J. Pedersen, J. and Chen, F.(1995). A Trainable Document Summarizer, *Proc. 18th ACM-SIGIR.*

[5] Tombros, A. and Sanderson, M.(1998). Advantages of Query Biased Summaries in Information Retrieval, *Proceedings of ACM-SIGIR'98*, 2-10.

[6] 강상배, 조혁규, 권혁철, 박재득, 박동인(1997), 한국어 문서의 통계적 정보를 이용한 문서요약 시스템 구현, 제 9회 「한글 및 한국어 정보처리 학술대회」, 28-36.

[7] 김기영, 전명식(1994),「다변량 통계자료분석」,자유아카데미

[8] 류동원, 이종혁(2000), 단어공기정보를 이용한 자동화 문서요약, 제27회「정보과학회 봄 학술발표논문집(B)」, 제27권, 1호, 339-341.

[9] 박혁로, 신중호(1999), 검색/요약/필터링을 위한 텍스트 이해 모형 및 처리 기술 개발, 「연구개발정보센터 연구보고서」.

[10] 양기주(1997), 수사구조에 기반한 한국어 요약문 생성, 「연구개발정보센터」.

[11] 장동현, 맹성현(1997), "자동 요약 시스템", 「정보과학회지」 제15권 제10호, 42-49.

[2001년 10월 접수, 2002년 4월 채택]

부 록 A

"임기중 개헌 없다/김대통령 취임 100일 회견"라는 제목의 신문기사

◎경제 살리는게 최우선 과제/15대땐 개혁인물 공천… 정계개편 안해/5·16은 쿠데타… 역사가 심판

김영삼대통령은 3일 『임기중에는 어떤 이유로도 헌법개정을 하지않겠다』고 말했다.

김대통령은 이날 청와대 춘추관에서 가진 취임1백일에 즈음한 내외신 기자회견에서 『대통령과 국회의원의 선거시기가 달라 문제가 있는것은 사실이나 임기중에는 개헌을 하는일이 없을 것이며, 5년동안 깨끗하게 최선을 다한 대통령이 되도록 할 것』이라고 강조했다.

〈관련기사 3-4면〉

김대통령은 정계개편 가능성을 묻는 질문에 『지금은 그럴 필요도 없을 뿐만 아니라 그럴 시기도 아니다』며 『다만 15대 총선 공천과정에서 국가에 책임을 질수 있고 깨끗하고 도덕적이며 개혁에 알맞는 사람이 많이 나올수 있도록 배려할 수는 있을 것』이라고 말했다.

김대통령은 TV와 라디오로 생중계된 이날 회견에서 『평화적 시위는 새로운 시위문화 정착을 위해 허락할 것』라고 전제, 『그러나 최근 학생들이 폭력적 시위를 하며 경찰을 무장해제시키고 친북계학생단체가 공개적으로 인공기를 달아놓고 북한과 통화한 것은 실정법 위반』이라며 『누구든 국가의 기강을 해치고 법을 지키지 않는 사람은 부정부패 척결 차원에서 결코 용납하지 않을 것』이라고 강조했다.

김대통령은 또 『우리의 당면과제중 가장 중요한 것이 경제를 살리는 것이지만 부정부패 척결없이는 경제가 살아날 길이 없다』며 『경제계가 열심히 일할 수 있도록 지원을 아끼지 않을 것』이라고 거듭 강조했다.

김대통령은 『민주주의 국가에서 재벌기업 해체는 있을수 없는 일 』이라며 『다만 대기업은 중소기업의 영역을 침범하지 말고 중소기업과 보완적 관계가 돼야 하며 주식도 가능하면 국민과 근로자에게 분배하는 것이 좋겠다』고 말했다.

금융실명제 실시문제에 대해 김대통령은 『대통령선거때 국민에게 약속한 대로 반드시 실시할 것이지만 현 시점에서 그 시기와 방법을 말하는 것은 옳지 않다』고 말했다.

김대통령은 『5·16은 분명히 쿠데타라고 생각하며 우리 역사를 크게 후퇴시킨 큰 시작이었다고 생각한다』며 『그러나 정치보복은 없어야 한다고 약속했듯이 이 모든 문제들은 역사의 심판에 맡기는 것이 옳다고 생각하며 전직대통령들에 대해서도 후일 역사의심판에 맡겨두자』고 말했다.

이어 김대통령은 개각문제에 언급, 『대통령이 되기 전부터 장관들을 자주 바꾸는것은 잘못된 것이라고 생각했으며, 현재 개각은 일체 고려하지 않고 있다』고 말했다.

단체장선거등 향후 정치일정에 대해 김대통령은 『자치단체장 선거는 반드시 해야하나 행정력면에서 이것을 따로따로 할 수가 없는 만큼 전산화등을 통해 몇개 선거를 묶어서 동시선거를 할 수 있는 방법을 연구할 필요가 있다』고 말했다.

이에앞서 김대통령은 회견연설에서 『벌써부터 작은 아픔을 못이겨 개혁을 이제 그만 덮어두자는 목소리가 나오고 있으나 개혁은 결코 중단될 수 없으며 우리들의 의식과 생활속에 뿌리내릴 때까지 지속될 것』이라고 말했다. 〈이혁주기자〉