주제어구 추출과 질의어 기반 요약을 이용한 문서 요약

(Document Summarization using Topic Phrase Extraction and Query-based Summarization)

한 광 록 † 오 삼 권 †† 임 기 욱 †††
(Kwang-Rok Han) (Sam-Kweon Oh) (Kee-Wook Rim)

요 약 본 논문에서는 추출 요약 방식과 질의어 기반의 요약 방식을 혼합한 문서 요약 방법에 관해서 기술한다. 학습문서를 이용해 주제어구 추출을 위한 학습 모델을 만든다. 학습 알고리즘은 Naive Bayesian, 결정트리, Supported Vector Machine을 이용한다. 구축된 모델을 이용하여 입력 문서로부터 주제어구 리스트를 자동으로 추출한다. 추출된 주제어구들을 질의어로 하여 이들의 국부적 유사도에 의한 기여도를 계산함으로써 요약문을 추출한다. 본 논문에서는 주제어구가 원문 요약에 미치는 영향과, 몇 개의 주제어구 추출이 문서 요약에 적당한지를 실험하였다. 추출된 요약문과 수동으로 추출한 요약문을 비교하여 결과를 평가하였으며, 객관적인 성능 평가를 위하여 MS-Word에 포함된 문서 요약 기능과 실험 결과를 비교하였다.

키워드 : 추출요약, 질의어 기반 요약, 문서요약, 국부적 유사도

Abstract This paper describes the hybrid document summarization using the indicative summarization and the query-based summarization. The learning models are built from learning documents in order to extract topic phrases. We use Naive Bayesian, Decision Tree and Supported Vector Machine as the machine learning algorithm. The system extracts topic phrases automatically from new document based on these models and outputs the summary of the document using query-based summarization which considers the extracted topic phrases as queries and calculates the locality-based similarity of each topic phrase. We examine how the topic phrases affect the summarization and how many phrases are proper to summarization. Then, we evaluate the extracted summary by comparing with manual summary, and we also compare our summarization system with summarization method from MS-Word.

Key words : indicative summarization, query-based summarization, document summarization, locality-based similarity

1. 서 론

개인이나 기업 또는 국가가 발전하기 위하여 정보가 미치는 영향을 무시할 수 없으며 역사의 변천에 따라 수많은 자료들이 기하급수적으로 불어나고 있다. 또한 컴퓨터와 인터넷의 발달로 날마다 증가하는 정보의 홍수 속에서 살아가고 있다. 오늘날 우리가 접하는 인터넷상의 대부분의 정보는 체계적으로 분류되고 정리되어 있기보다는 정보 제공자들의 관심도와 편의에 의하여 여러 가지 형태로 제공되기 때문에 우리가 원하는 정보를 얻기까지 많은 시간을 소모하게 되는 경우도 있다. 여러 곳에 분산되어 있고 정돈되어 있지 않은 정보원으로부터 필요한 정보를 신속하고 정확하게 검색해 내기 위해서는 모든 자료들을 자세히 읽고 검색하기보다는 가장 핵심이 되는 문서의 내용만을 추출함으로써 빠른 시간 내에 문서 전체의 내용을 효과적으로 이해할 수 있는 문서 요약이 요구된다.

문서 요약이란 특정한 사용자가 작업에 적절하게 문서를 축약하여 문서의 복잡도나 길이를 줄이면서 필요

· 본 논문은 과학기술부 지정 호서대학교 반도체 제조장비 국산화 연구센터의 지원에 의해 이루어졌음
† 종신회원 : 호서대학교 컴퓨터공학부 교수
krhan@office.hoseo.ac.kr
†† 종신회원 : 호서대학교 컴퓨터공학부 교수
ohsk@office.hoseo.ac.kr
††† 종신회원 : 선문대학교 지식정보산업공학과 교수
rim@sunmoon.ac.kr
논문접수 : 2002년 12월 18일
심사완료 : 2004년 1월 6일

한 정보를 유지한 상태로 사용자에게 내용을 전달하는 것을 말한다. 즉, 원래의 자료의 크기에 비해 상당히 축소된 문서를 만들어내야 하며 요약된 내용이 원문의 내용과 의미상 맥락을 같이 해야 한다. 그러나 같은 자료가 주어진다고 하더라도 독자의 관심도에 따라 얻는 정보의 중요성이 달라지기 때문에 기계적으로 독자의 의도를 정확히 파악하여 자료를 요약하기란 쉬운 일이 아니다.

저자의 관점에 따라 의미가 달라지는 문서가 아니고 객관적인 입장에서 쓰여진 이론을 기술하는 문서의 경우에는 원래의 자료를 추상화하거나 단순화시켜 문서를 요약한다. 이 경우에는 문서에 접근할 때 연역적인 시각으로 문서의 논리를 명제화하고 이로부터 연역적으로 추출한 결과를 요약문으로 사용할 수 있지만, 이러한 요약 방법은 개발 및 응용에 많은 비용을 수반하기 때문에 원문의 내용을 부분적으로 삭제하는 방식을 주로 사용한다.

그러나 많은 양의 문서를 요약하기 위해서는 요약 방법에 대한 정확성과 효율성을 동시에 요구된다. 본 논문에서는 문서의 내용을 요약하기 위하여 문서의 주제를 대표하는 중요한 단어나 구절들로 구성된 주제어구 후보들을 수동으로 추출한 후에 이들을 학습하여 모델을 구축하고 이 모델을 기반으로 새로운 입력 문서의 주제어구들을 추출한다. 주제어구들이 추출되면 이 주제어구들을 질의어로 하고 이들의 국부적 유사성을 계산하여 기여도가 높은 문단을 추출함으로서 문서를 요약한다. 즉, 추출 요약(indicative summarization)과 질의어 기반의 요약(query-based summarization)을 이용하는 혼합적 접근 방법(hybrid approach)으로 문서를 요약한다. 주제어구를 추출하기 위한 도구로서 KEA-2.0[1]을 이용하고, 질의어 기반의 요약은 seft[2]를 이용한다. 또한 효과적으로 주제어구 리스트를 추출하기 위하여 여러 가지 기계학습 모델들을 적용하고 그 결과를 검토했으며, 주제어구가 요약에 미치는 영향과 정확성을 평가한다.

2. 관련 연구

문서 요약에는 사용자의 관심과 작업 분량, 특성 등에 따라 여러 가지 방법들이 연구되고 있다. 단일 문서만을 요약 대상으로 하는 단일 문서 요약과 많은 문서를 하나의 문서로 요약하는 다중 문서 요약이 있다. 보통 인터넷 문서에 대해서는 단일 문서 요약을 이용하고, 회사나 전체적인 흐름을 중요시하는 작업에서는 여러 가지의 문서에서 중요한 공통 내용만을 추출하는 다중 문서 요약을 이용한다[3]. 문서의 사용 여부에 따라서 원문이 어떤 것인지 만을 제시하는 즉, 문서가 원하는 정보에 적합한지 만을 판단하는 추출 요약(indicative summarization)이 있고[4], 문서의 중요한 내용을 내포하는 요약문만을 추출하여 정보를 전달하는 정보적 요약(informative summarization)도 있다[5]. 이와 같은 요약 방법들은 사용자가 원하는 지식이 무엇인지, 어떠한 정보를 얻기 위한 작업인지의 여부에 상관없이 문서의 내용 자체를 요약하기 때문에 이러한 요약을 포괄적인 요약(generic summary)이라고도 한다[4]. 이와 달리, 사용자의 의도에 맞는 요약 결과를 얻기 위하여 문서 내의 내용을 포괄적으로 요약하기보다는 사용자에게 중심이 되는 단어, 즉 사용자가 관심을 갖는 정보에 근거하여 요약을 하는 질의어 기반 요약 방법도 있다[6]. 질의어 기반 요약의 경우 Mark Sanderson이 Inquery 검색 시스템의 국부적 문맥 분석을 이용하여 요약문을 생성하는 방법을 연구하였으나[7], 의사 적합성 피드백과, 제목, 문서의 첫 문장 등을 질의로서 확장하여 이용한 질의확장 기반의 요약에 비해 성능 향상을 보이지 못하였다[8].

그리고 원문의 중요한 문장을 그대로 추출하여 요약문을 만드는 추출 요약이 있고[9-11], 원문 내에서 여러 문장이나 절들을 압축하여 새로이 문장을 만들어 내는 생성요약(abstract summary)도 있다. 이 생성요약은 여러 가지 지식을 필요로 한다. 예를 들어, 자연어 처리 기술이나 문법적으로 언어에 대한 기본적인 지식이 프로그램 내에 구현되어야 한다. 그러나 사실상 이것을 구현하기란 쉽지 않다. 이와 같이 문서들로부터 요약 정보를 생성하거나 추출하기 위하여 여러 가지 방법들이 연구되고 있다.

3. 요약 시스템 설계 및 구현

3.1 시스템 개요

본 논문에서는 질의어 기반 요약 방법과 추출 요약 방법을 혼합하여 문서의 요약문을 추출해 내도록 한다. 이를 위하여 먼저 학습문서와 각 문서에 대하여 수동으로 추출한 주제어구 리스트를 대상으로 여러 가지 학습 알고리즘을 적용하여 학습모델을 만든다. 생성된 모델을 기반으로 새로운 입력 문서로부터 주제어구들을 추출해 낸다. 이와 같이 추출된 주제어구들을 질의어로 하여 국부적 유사도 계산을 이용한 질의어 기반의 요약 방식으로 문서의 요약문을 추출해 낸다.

(1) 요약 시스템

전체 문서 요약 과정은 그림 1과 같이 3단계로 구성된다. 첫 번째 단계는 학습문서와 해당 문서의 주제어구 리스트를 이용하여 학습 모델을 생성하는 과정이다. 두 번째 단계는 만들어진 학습 모델을 이용하여 입력 문서에 대한 주제어구들을 추출하는 과정이다. 세 번째 단계는 입력 문서와 추출된 주제어구들을 질의어로 하여 국

부적 유사도(locality-based similarity)를 이용한 기여도를 계산함으로써 문서의 요약문을 추출한다.

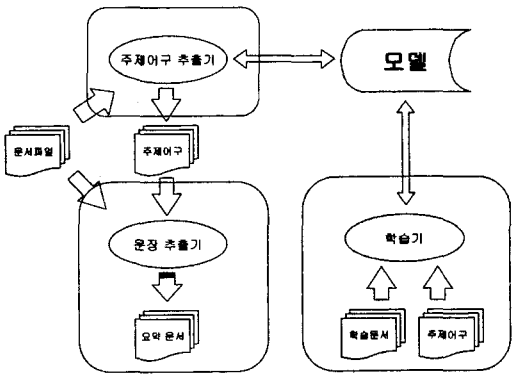


그림 1 요약 시스템의 구성도

(2) 전처리 과정

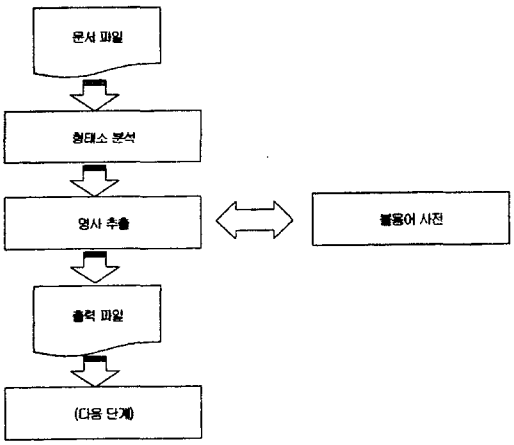


그림 2 전처리 과정

그림 1의 각 과정을 진행하기 위해서는 그림 2와 같은 전처리 과정을 거친다. 그림 2는 입력 문서의 각 문장에 대하여 형태소 분석을 하고, 불용어 사전을 참조하여 문장 중에서 중요도를 계산하는데 도움이 되지 않는 단어들을 제거하는 과정을 거친다. 일반적으로 불용어란, 문서내에 많이 출현하지만 의미를 부여하기 힘든 단어들을 일컫는데, 한국어의 경우, 대명사, 조사, 부사 등의 단어들이 이에 해당한다. 추출된 단어들은 주제어구 추출을 위한 후보 단어가 된다. 한글 문장의 형태소 분석에는 HAM 라이브러리를 이용하였다[12]. 전처리 과정은 학습과 주제어구 추출, 요약문장 추출의 세 단계에 모두 적용된다.

3.2 주제어구 리스트 추출

주제어구 리스트 추출은 먼저 기계 학습에 의해 모델을 구축하고, 이 모델을 기반으로 입력된 문서에 대한 주제어구를 추출한다. 기계 학습에서는 Naive Bayesian, 결정트리, Supported Vector Machine 알고리즘을 이용한다. 학습을 위한 입력 데이타는 학습문서와 각 문서에 대하여 수동으로 추출한 주제어 리스트가 사용된다.

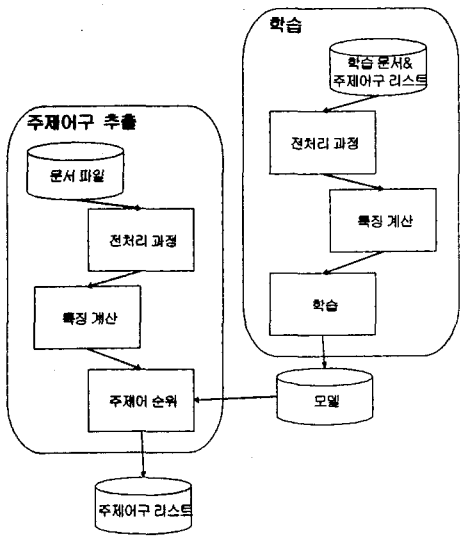


그림 3 학습과 주제어 추출과정

그림 3의 오른쪽은 학습하는 단계로 입력 데이타를 전처리 과정을 거친 후 주제어구 리스트와 관련된 학습문서의 특징을 추출하고, 해당 특징값을 계산하여 학습모델을 구축하는 과정을 나타낸다.

(1) 특징 추출

본 논문에서는 주제어구를 추출하기 위하여 전처리 과정을 거쳤는데, 전처를 과정을 거친 문서내의 단어들은 모두 주제어구의 후보단어가 되며, 각각의 단어는 *TF*IDF*, 최초 출현 거리를 계산한 값들을 특징값으로써 가지게 된다. 즉 각각의 후보 단어는 (단어 | TF*IDF 값 | 최초출현거리)의 쌍을 가지게 되고, (2)항에서 설명할 알고리즘의 입력이 된다.

① TF*IDF

TF*IDF값은 해당 주제어구가 일반적으로 사용되는 빈도와 특별한 문서 내에서 사용되는 빈도를 척도로 사용되며 다음과 같은 수식으로 표현된다[1].

$$TF*IDF = \frac{freq(phrase, doc)}{size(doc)} \times -\log_2 \frac{df(phrase)}{N}$$

freq(*phrase, doc*) : 문서 *doc*에서 주제어구 *phrase*

의 출현빈도

size(doc) : 문서 *doc*에 나타난 단어의 수

df(phrase) : 전체 문서에서 주제어구 *phrase*를 포함하는 문서들의 수

N : 전체 문서의 수

② 최초 출현 거리

최초 출현 거리는 문서의 시작으로부터 해당 주제어구까지의 거리를 나타낸다[1]. 이 값은 주제어구 이전에 나타난 단어들의 수를 문서 전체의 단어수로 나눈 값으로 0과 1사이의 값이 된다.

(2) 학습모델 구축

문서의 특징을 이용하여 얻은 특징 값을 기반으로 여러 가지 학습 알고리즘을 이용하여 학습모델을 구축한다.

① Naive Bayesian 학습

Naive Bayesian 학습은 개체의 종류가 정해진 경우 각 특성들 사이의 독립을 가정하여 학습하는 방법으로 각각의 인스턴스 결과 값을 독립적인 확률로 구하여 규칙을 만들고 임의의 데이타에 대한 예상 결과값을 추측해내는 방법이다. 이 알고리즘은 확률을 구할 때 데이타의 양이 많을수록 결과의 적중 확률이 높아진다. 다음은 Naive Bayesian 학습의 공식이다[13].

$$P[phrase|tif, dis] = \frac{P[tif|phrase] \times P[dis|phrase] \times p[phrase]}{P[tif, dis]}$$

여기서 $P[tif|phrase]$는 하나의 주제어구 후보가 $TF*IDF$ 값 tif를 갖는 확률이고, 이들은 전부 주제어구 후보가 주제어구가 될 조건부 확률을 나타낸다. $P[dis|phrase]$는 최초 발생 거리 dis를 갖는 확률이다. $P[phrase]$는 사전확률이다. 즉 위의 수식은 하나의 주제어구 후보가 주제어구 가 될 조건부 확률들의 곱이라고 할 수 있다.

② 결정트리 학습

결정트리는 정보이론을 기반으로 하여 귀납적 유도 학습 방법으로 가장 많이 사용되어지는 것 중 하나로서 1949년 Shannon과 Weaver에 의해 처음으로 소개되었다. 잡음이 있는 데이타에 유리하며, 표현을 구별하는 학습 능력이 뛰어난 점이 결정 트리의 특성이며, 이 특성을 이용하여 문서의 범주화에 많이 사용된다. 루트에서 시작하여 결과값이 하나로 함축될 때까지 가능한 모든 값들의 개수만큼 가지를 확장해 가는 방법이다. 그러나 이 알고리즘은 예상할 수 있는 결과값이 단 하나일 때만 가능하다는 단점이 있다. 그러나 이산값을 가지는 함수의 추정을 이용하여 규칙 집합을 구축하기가 쉽고 학습을 통해 생성된 결정 트리를 규칙 집합으로 이해하는 것이 가능하다. 결정 트리는 음성 인식이나 문서 분류, 문서 요약, 중심어구 찾기 등의 모호성을 해소할 때 주로 사용되며, 구문 분석 시 문장 내 단어의 품사를 결정하거나 접속사의 접속 범위를 결정하는 등 구문 분석에도 자주 사용된다. 가장 잘 알려진 결정 트리 알고리즘으로 ID3가 있으며, 그 후속으로 C4.5와 C5 등이 있다. 본 논문에서는 C4.5 알고리즘을 이용한다[14].

③ Supported Vector Machine 학습

Supported Vector Machine은 특별한 형태의 선형 모델로서 각 인스턴스 사이의 최적의 경계(최적 분리면)를 발견하는 알고리즘이다[15]. 최적 분리면은 그림 4와 같이 클래스들 사이를 가장 명확하게 분리하는 것이고, 이 최적 분리면에 가장 가까운 인스턴스들이 최적 분리면과 최단 거리를 갖는 support vector들이다.

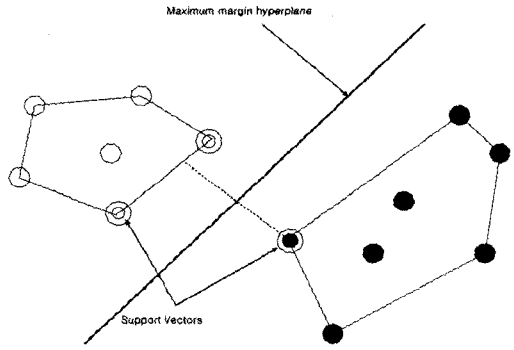


그림 4 Supported Vector 모형

그림 4에서 두 개의 클래스를 분리하는 분리면은 다음과 같이 나타낸다.

$$\chi = \omega_1 a_1 + \omega_2 a_2$$

a_i : 속성값(*attribute value*)
ω_i : 학습되는 가중치

또한 support vector로 표현되는 최적 분리면은 다음과 같이 표현된다.

$$\chi = b + \sum \alpha_i \gamma_i \mathbf{a}(i) \cdot \mathbf{a}$$

γi : Supported Vector $\boldsymbol{a}(i)$의 클래스 값
b, α_i : 파라미터의 개수
$\boldsymbol{a}$: 테스트 인스턴스 벡터
$\boldsymbol{a}(i)$: Supported vector

(3) 주제어구 추출

학습과정에서 생성된 모델을 이용하여 주제어구 리스트를 추출하는 과정으로 그림 3의 왼쪽에 그 모형을 나타낸다. 주제어구 리스트는 학습된 모델을 기반으로 추출한다. 그림 3과 같이 구축된 학습 모델을 기반으로 입력 문서로부터 계산된 자질값에 따라 상위 20개의 주제어구 리스트를 추출한다. 추출된 각 주제어구는 여러 개의 단어로 구성되며 적절한 주제어구들을 선택하기 위해

추출된 리스트에 대하여 다음과 같은 규칙을 적용한다.

① 주제어구를 구성하는 단어가 다른 주제어구의 단어와 중복이 되는 경우, 중복된 단어 하나만으로 구성된 주제어구를 삭제한다.

② 주제어구를 구성하는 단어의 길이가 다를 경우 구성된 단어의 수가 큰 것을 우선순위로 한다.

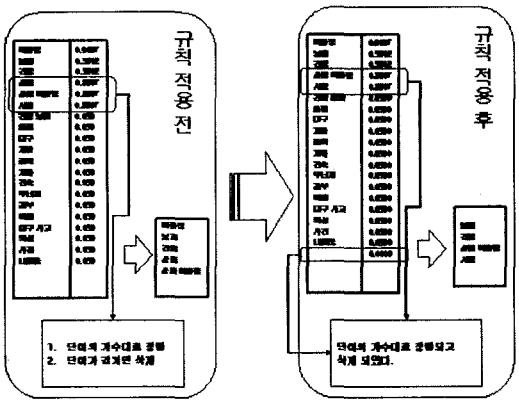


그림 5 규칙에 따른 주제어구 추출 예

그림 5는 이러한 규칙이 적용되는 예를 나타낸다. 규칙 적용 전의 주제어구 리스트에서는 하나의 주제어구인 '삼풍'과 또 다른 주제어구인 '백화점'은 각각 '삼풍백화점'이라는 주제어구와 단어가 중복이 된다. 따라서 규칙①을 적용하면 '삼풍'과 '백화점'이라는 주제어구는 삭제되고 그 다음 높은 자질값을 가지는 '서울'이라는 주제어구가 선택된다. 또한 추출된 주제어구 중에서 용언이 포함된 경우 용언은 주제어구 추출에서 제외한다. 이는 용언이 주제어구로 추출된 경우가 거의 없었으며, 용언이 추출된 주제어구를 이용하여 요약문을 선택할 경우 중요하지 않은 문장들이 선택될 확률이 너무 높기 때문이다.

주제어구는 요약문장을 추출할 때에 가장 중요한 역할을 한다. 너무 많은 주제어구를 포함시키면 중요하지 않은 구나 절들이 요약문으로 선택 될 수 있기 때문에 주제어구의 개수를 정하는 문제도 고려해야 한다.

3.3 요약문 추출

그림 6은 요약문이 추출되는 과정을 나타낸다. 입력 문서와 3.2절의 과정에 의하여 추출된 주제어구들을 질의어로 하여 국부적 유사도를 기반으로 하여 기여도를 계산하고, 그 값을 이용하여 요약문을 추출해 낸다.

(1) 국부적 유사도 계산

국부적 유사도 계산에서는 검색 대상을 일련의 문서들의 연속이라고 간주하기보다 단어들의 연속으로 간주한다[2]. 또한 문서 내에서 각 질의어 항의 출현은 가까

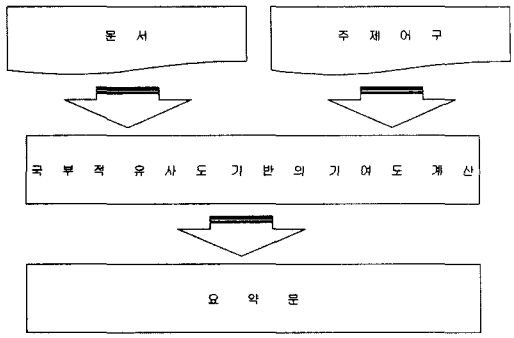


그림 6 요약문 추출 과정

이 근접한 다른 단어들에 영향을 미친다고 가정한다. 그래서 분리된 질의어 항으로부터의 영향은 누적이 되고 각 질의어 항의 출현에 따른 기여도는 합산되어 문서 내의 어떤 특정한 위치에 대하여 유사도 값에 도달한다. 높은 밀도의 질의어 항들을 포함하는 문서 영역이나 다소 덜 밀집되었지만 중요도가 높은 질의어 항의 문서 영역들은 그 문서에서 보다 높은 결합 영향력을 갖게 되고, 잠정적으로 높은 관심도를 갖는 문서 내의 지역성을 나타낸다. 일단 관심도가 가장 높은 영역을 식별하게 되면 그 영역에 집중된 문서 창이 평가된 유사도 값에 따라 사용자에게 제시된다. 특정 위치(x)에서 각각의 질의어가 가지는 기여도를 나타내는 기여도 함수 c_t는 다음과 같이 정의한다[2].

$$c_t(x, l) = h_t \cdot \sqrt{1-(d/s_t)^2}$$

$$h_t = f_{q,t} \cdot \log_e(\frac{N}{f_t})$$

$$s_t = \frac{n}{f_t}$$

여기서 $d=|x-l|$은 질문에 있는 질의어 항과 그 영향력이 평가되는 위치 사이에 있는 단어들의 거리를 나타내고, N은 문서 집합에 있는 모든 항들의 개수이다. f_t는 항 t가 나타낸 빈도수이고, $f_{q,t}$는 항 t의 질의 내부 빈도이다. 또한 n은 문서 집합내의 고유한 단어들의 수이다. 문서 파일 집합에 있는 한 단어의 위치 x에서 질의어 집합 Q에 관한 값인 $C_Q(x)$를 다음과 같이 정의한다[2].

$$C_Q(x) = \sum_{t \in Q} \sum_{l \in I_t} c_t(x, l)$$

여기서 Q는 질의어 항들의 리스트이고, I_t는 항 t가 나타난 문서 집합에서의 단어 위치들의 집합이다. 즉 $C_Q(x)$는 각각의 주제어구들이 특정 위치 x에서 가지는 기여도 합이라고 간주 할 수 있다.

(2) 요약문 추출

그림 7은 입력 문서이며 총 36개의 문장으로 이루어져 있다. 이 문서를 3.2절의 방법에 의해 그림 8과 같이 주제어구들을 추출하고, 이 주제어들을 질의어로 하여

며칠 전 낌새가 있었다. 벌어진 틈들이 거미줄처럼 벽을 그렸던 것이다. 삼풍 백화점은 서울의 부유촌인 서초구에 있는데, 이 5층짜리 건물 벽체의 천정 틈에서 물이 쏟아지고 있었다. 지난 목요일 아침 9시 30분에 꼭대기층의 식당 천장의 상당 부분이 내려앉았다. 그리고 12시에 에어컨의 기능이 중지됐다. 오후 이곳을 쇼핑하고 있었던 39세의 주부 장인석은 "마치 내부는 사우나탕과 같았습니다. 더 이상 머물 수가 없었습니다"라고 말했다. 장씨는 운이 좋았다. 오후 6시 직전 지하 슈퍼에서 저녁 시간대에 세일을 하기 때문에 백화점은 손님들로 붐볐는데, 바로 이때 벽체의 벽이 안으로 무너졌다. 백화점 바로 앞에서 교통 신호를 기다리고 있던 택시 운전사 박민수는 "백화점 빌딩이 약 10초 만에 무너지는 것을 목격했습니다. 마치 TV에서 보는 폭파반들이 폭파하는 모습이었습니다"라고 말했다. 모든 것이 지하로 무너져 내려, 곧 거대한 무덤을 이루었다. 주말까지 130명의 사망자가 확인되었고, 약900명이 다쳤다. 첫 보도는 가스 누출이 사고의 원인일 가능성을 비춰, 한국 남부 지방 대구에서 발생한 4월의 비극을 떠올렸다. 대구 사고는 지하철 건설 공사장에서 101명이 사망한 사건이었다. 그러나 지난 주의 비극은 대구 사고보다 더 살벌한 이유가 있었다. 6년 전에 지은 삼풍 백화점은 안전하지도 못했을 뿐 아니라, 소유자와 경영자들이 그에 못지 않은 의심을 받았다. 건물이 내려앉은 오전에 안전 전문가를 불러들였는데, 전문가들은 건물을 점검한 결과, 건물이 곧 붕괴될 위험이 있다고 밝혔던 것이다. 그러나 간부들은 대피 명령을 내리지 않은 채, 5층 벽의 구멍 난 틈을 메우려했고, 종업원들에게는 상층 2개 층에 있던 상품을 지하 저장고로 옮기라고 지시했다. 건물 붕괴 후 이준 회장과 간부들은 형사상 직무 유기 혐의로 구속되었다. 그날의 회의록도 압수되었다. 뉴스 보도에 의하면 간부들은 자신들의 안전을 먼저 생각한 나머지, 붕괴 30분 전에 회의를 마치고 건물을 빠져나갔다고 한다. 경찰은 백화점 붕괴의 원인으로 시멘트 배합이 잘못되었거나 건축할 때 사용된 강철 빔이 충분하지 못했을 것이라고 생각한다. 경찰은 건축과 건물 안전을 책임지고 있는 4명의 구청 직원을 조사하기 위해 찾고 있다고 했다. 주말에 구조반원들은 자원 봉사자들과 부근의 용산 미군 부대의 지원을 받아 생존자를 구출했다. 이 지역에 살고 있는 수백명의 사람들이 헌혈 요청에 응했다. 건물이 붕괴된 지 52시간이 지난 토요일, 24명의 청소원들이 구조됐다. 이들은 교대 근무를 마치고 옷을 갈아 입고 있었을 때 이 비극적인 일이 발생했던 것이다. 삼풍 건물 붕괴는 최근 몇 개월 동안 한국을 강타한 세 번째 참사였다. 대구 폭발 사건 전에는 서울 성수대교의 가운데 부분이 지난 10월에 붕괴돼 32명이 사망했었다. 이러한 사건들은 한국 건축 기술의 질적 수준과 지방 관청의 검열 관행을 의심하게 했다. 영향력 있는 조선일보는 "이런 일은 이제 도처에서 발생하고 있어, 우리 생활의 일부가 돼버린 느낌이다'라고 적었다. 한국인들은 이러한 비극적 사건들에 대하여 정부에 책임을 돌리는 경향이 있어, 김 대통령으로서는 이중 부담이 되고 있다. 지난 주초에 민주 자유당 소속의 후보들이 몇 십년 만에 열린 지방 선거에서 참패했었다. 민자당은 지방 선거에서 겨우 삼분의 일밖에 승리하지 못했다. 정지 주도권을 쥐기 위해 김 대통령은 건물 붕괴 사고의 즉각적인 수사를 지시했다.

그림 7 입력 문서

국부적 유사도 기반의 기여도 계산을 하여 그림 9와 같이 요약문을 추출한다. 하나의 요약문은 다섯 문장으로 구성되도록 조절하였으며, 요약문은 국부적 유사도의 기여도 값에 따라 세 개의 후보를 추출한다.

붕괴
건물
백화점
건물 붕괴
대구 사고

그림 8 주제어구 리스트

각각의 출력된 요약문의 헤더에는 요약 정보를 기록한다. 첫째 열은 요약 순위를 나타내고, 둘째 열은 해당 문서의 파일명을, 마지막 열은 그 문서에서 기여도가 가장 큰 문장 번호를 나타낸다. 그림 9의 요약문의 헤더인 '== 1 (100%) == decisiontree\Time039.txt:23'에서 '1 (100%)'는 100% 기여도로 정규화된 1 순위 후보임을 나타내고, 'decisiontree\Time039.txt'는 문서파일의 이름이다. 그리고 '23'은 기여도가 가장 큰 문장의 번호이다.

4. 실험 및 평가

본 논문에서 추출한 요약문을 평가하기 위하여 여러 가지 시도를 하였다. 실험을 위해 인문, 경제, 정치 분야의 문서 중 길이가 약 30-120문장을 전후하는 문서들을 각각 50개씩 총 150문서를 추출하여 실험에 이용하였다. 또한 10명의 4년제 대학에 재학 중인 학생을 선별하여, 수작업으로 주제어구를 추출하여 주제어구 추출을 위한 학습 데이타로 활용하였다. 본 실험에서는 이렇게 수작업으로 추출된 주제어구를 학습하여, 문서 내의 주제어구를 자동으로 추출할 수 있도록 하였으며, 5개로 주제어구를 제한하여 추출한 경우 주제어구가 수작업에 의해 추출된 주제어구와 일치된 확률이 75-80% 정도를 보여주고 있다.

요약문의 평가는 추출된 요약문 중에 사람이 수동으로 발췌한 요약문의 문장을 얼마나 포함하고 있는 지를 계산하여 평가한다. 그림 9와 같이 요약문으로 추출된

== 1 (100%) == decisiontree\Time039.txt:23 그날의 회의록도 압수되었다. 뉴스 보도에 의하면 간부들은 자신들의 안전을 먼저 생각한 나머지, 붕괴 30분 전에 회의를 마치고 건물을 빠져나갔다고 한다. 경찰은 백화점 붕괴의 원인으로 시멘트 배합이 잘못되었거나 건축할 때 사용된 강철 빔이 충분하지 못했을 것이라고 생각한다. 경찰은 건축과 건물 안전을 책임지고 있는 4명의 구청 직원을 조사하기 위해 찾고 있다고 했다. 주말에 구조반원들은 자원 봉사자들과 부근의 용산 미군 부대의 지원을 받아 생존자를 구출했다.
== 2 (97%) == decisiontree\Time039.txt:18 그러나 지난 주의 비극은 대구 사고보다 더 살벌한 이유가 있었다. 6년 전에 지은 삼풍 백화점은 안전하지도 못했을 뿐 아니라, 소유자와 경영자들이 그에 못지 않은 의심을 받았다. 건물이 내려앉은 오전에 안전 전문가를 불러들였는데, 전문가들은 건물을 점검한 결과, 건물이 곧 붕괴될 위험이 있다고 밝혔던 것이다. 그러나 간부들은 대피 명령은 내리지 않은 채, 5층 벽의 구멍 난 틈을 메우려했고, 종업원들에게는 상층 2개 층에 있던 상품을 지하 저장고로 옮기라고 지시했다. 건물 붕괴 후 이준 회장과 간부들은 형사상 직무 유기 혐의로 구속되었다.
== 3 (68%) == decisiontree\Time039.txt:15 주말까지 130명의 사망자가 확인되었고, 약900명이 다쳤다. 첫 보도는 가스 누출이 사고의 원인일 가능성을 비춰, 한국 남부 지방 대구에서 발생한 4월의 비극을 떠올렸다. 대구 사고는 지하철 건설 공사장에서 101명이 사망한 사건이었다. 그러나 지난 주의 비극은 대구 사고보다 더 살벌한 이유가 있었다. 6년 전에 지은 삼풍 백화점은 안전하지도 못했을 뿐 아니라, 소유자와 경영자들이 그에 못지 않은 의심을 받았다.

그림 9 추출된 요약문의 예

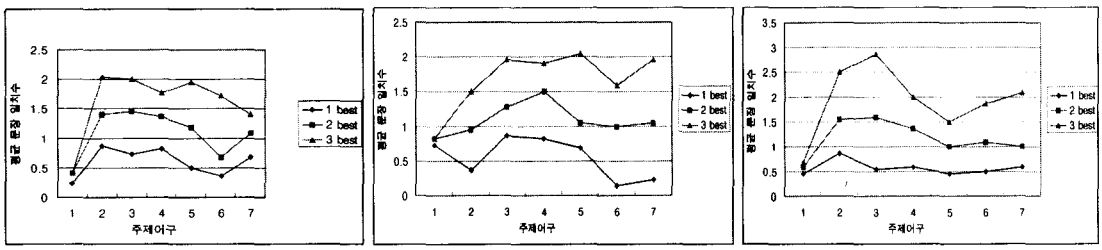
순위를 정한 3개의 후보들에 대하여 1 best, 2 best, 3 best 형태로 평가하였다. 1 best는 1 순위 요약문만을 평가하였고, 2 best는 1 순위와 2 순위 후보를 포함시켰으며, 3 best에서는 1, 2, 3순위 후보를 모두 포함시켜 평가했다.

또한 문단 단위와 문장 단위의 평가방법도 시도하였다. 문단 단위 평가에서는 각 문서에서 수동으로 추출한 요약문의 한 문단과 질의어 기반의 국부적 유사도에 의한 기여도를 계산하여 추출한 문단을 비교하여 얼마만큼의 정확성이 있는지, 즉 몇 개의 문장이 서로 일치하는 지를 판단하는 실험을 하였다. 문장 단위 평가에서는 모든 문서의 요점이 한 곳에 집중되어 있지 않고 산재되어 있을 경우를 가정하여 중요하다고 판단되는 문장 다섯 개를 부분 추출하여 요약문의 결과와 비교하는 실험을 하였다. 마지막으로 문서를 구성하는 문장 수에 따른 정확도를 평가하였다.

4.1 문단 단위의 평가

각 문서에서 수동으로 추출한 요약문의 문단과 본 논문의 방법으로 추출한 요약 문단을 비교하여 몇 개의 문장이 서로 일치하는 지를 실험한다. 그림 10은 문단 단위의 평가 결과를 나타낸다.

그림 10(a)의 Naive Bayesian 실험 결과는 1 best에서 주제어구 수가 2, 3, 4개일 때 평균 1개의 문장이 일치하고, 2 best에서는 평균 1.5개 문장이, 그리고 3 best에서는 평균 2개의 문장이 일치하였다. 그림 10(b)의 Supported Vector에서는 1 best, 2 best, 3 best일때 각각 평균 1개, 1.5개, 2개의 문장이 일치하였다. 그림 10(c)의 결정 트리 결과는 1 best에서 전반적으로 1개의 문장이 일치하고, 2 best에서는 주제어구 수가 2, 3개일 때 평균 1.5개의 문장이 일치하였다. 그리고 3 best에서는 주제어구 수가 3개일 때 가장 높은 평균 3개의 문장이 일치했다.

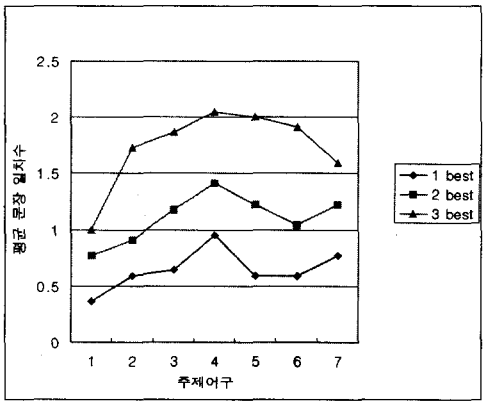


(a) Naive Bayesian (b) Supported Vector Machine (c) Decision Tree

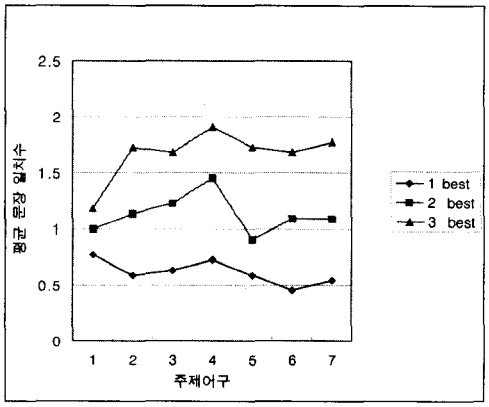
그림 10 문단 단위 평가 결과

4.2 문장 단위 실험

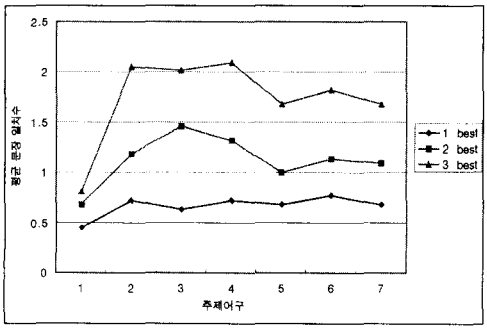
또 다른 평가를 위하여 문서에서 중요하다고 판단되는 산재되어 있는 개별 문장 다섯 개를 수작업을 통해 부분적으로 추출하여 본 논문에서 추출한 요약 문단과 비교하여 몇 개의 문장이 서로 일치하는지를 실험하였다. 그림 11은 각 학습 알고리즘 별로 실험한 문장 단위의 평가 결과를 나타낸다. 그림 11에서 모든 알고리즘이



(a) Naive Bayesian



(b) Supported Vector Machine



(c) Decision Tree

그림 11 문장 단위의 평가 결과

거의 비슷한 결과를 나타내고 있으며 3-4개의 주제어구에 대하여 평균적으로 두 개의 문장이 일치함을 보인다.

결과적으로 문단별 평가에서 주제어구 수가 평균 3-4개일 때 가장 높은 결과를 나타내었고, 여러 개의 주제어구들이 리스트에 포함될지라도 문서내의 출현빈도가 낮은 주제어구들을 질의어로 하여 계산할 경우에는 오히려 요약문 추출에 있어서 성능 저하의 원인이 될 수 있음을 알았다. 또한 학습 알고리즘은 결정 트리가 가장 좋은 것으로 평가되었다.

마지막으로 80개 이하의 문장들로 구성된 문서들과 80개 이상의 문장들로 구성된 문서들에 대하여 결정트리로 학습한 후 그 결과를 비교하였으며, 그 결과 80 문장 이하의 문서가 80 문장 이상의 문서보다 전반적으로 정확도가 높았다.

4.3 정확성 비교

본 논문에서는 실험의 보다 객관적인 평가를 위하여 MS-Word에 포함된 요약 기능과의 비교평가를 수행하였다. 평가를 위해 인문, 경제, 정치 분야의 문서 중 길이가 약 80문장을 전후하는 문서들을 이용하였으며, 요약문의 길이는 문서의 20%로 하였으며, 길이의 단위는 문장으로 하였다. 평가를 위해 선택된 문서들은 우선 사람에 의해 수동으로 요약문을 추출하고, 추출된 문장들을 본 논문에서 제안한 방법과 MS-Word의 요약기능을 이용한 결과와 비교를 하였으며, Optimistic, Pessimistic, Intersection, Union의 네 가지 경우로 각각 평가하였다[10,16]. 또한 수작업으로 요약문을 추출하기 위하여 4년제 대학에 재학 중인 학생 10명을 무작위로 선택하여 실험에 참가시켰다.

Optimistic: 각각의 수작업에 의한 결과와 구현된 시스템의 결과를 비교하여 일치도가 높은 것을 선택한다.

Pessimistic: 각각의 수작업에 의한 결과와 구현된 시스템의 결과를 비교하여 일치도가 낮을 것을 선택한다.

Union: 각각의 수작업에 의해 추출된 핵심 문장들 모두를 시스템의 결과와 비교한다.

Intersection: 각각의 수작업에 의해 추출된 핵심 문장들 중 일치하는 것만을 추출하고 이를 시스템의 결과와 비교한다.

아래의 표 1은 평가 결과를 보여 주고 있다. 우선 주목할 점은 수작업에 의해 추출된 요약문이 동일 문서에 대해 평균 47% 정도의 일치율을 보여 주고 있다는 점이다. 즉 비슷한 수준의 교육을 받은 사람일지라도 보는 관점에 따라 문서 내에서 중요 문장이라고 판단되는 부분이 상당부분 다를 수 있다는 것이다. 이것은 기존의 연구에서도 이미 언급된 사실이다[10,16]. 중요한 것은 여러 사람이 중요하다고 인식 하는 문장들 즉 Intersection과의 일치율이다. 서로 다른 관점을 가진 사람이

표 1 비교 평가 결과

사람에 의해 추출된 요약문의 일치율 : 47%				
방법	Optimistic (%)	Pessimistic (%)	Intersection (%)	Union (%)
제안된 방법	47	30	50	55
MS-Word의 요약	43	26	38	52

동일하게 중요하다고 인식한 만큼 중요문장으로서의 가치가 높다고 판단되며 이러한 문장들과의 일치율이 본 실험에서 중요한 요소로 간주되었다. 표 1에서 나타나듯이 본 논문에서 제안된 방법이 MS-Word에서 제공하는 요약문에 비해 정확도가 높음을 알 수 있다.

5. 결 론

본 논문에서는 추출 요약 방식과 질의어 기반의 요약 방식을 혼합한 문서 요약 방법에 관해서 기술했다. 학습 문서를 이용해 주제어구 추출을 위한 학습 모델을 구축하였으며 학습 알고리즘으로는 Naive Bayesian, 결정트리, Supported Vector Machine을 적용하였다. 구축된 학습 모델을 기반으로 입력 문서로부터 주제어구들을 자동으로 추출한다. 추출된 주제어구들을 질의어로 하여 질의어들의 국부적 유사도에 의한 기여도를 계산하는 질의어 기반 요약방식으로 요약문을 추출한다. 원문을 요약할 때 주제어구가 미치는 영향에 대해 실험하고 각각의 추출된 요약문과 사람이 수동으로 추출한 요약문과의 일치성을 비교하였다.

평가 결과는 문단 단위 요약에서 80문장 이하로 구성된 문서들에 대하여 가장 좋은 요약 결과를 나타냈으며, 학습 알고리즘으로는 결정 트리가 가장 좋았다. 또한 주제어구의 개수는 3~4개가 가장 적절한 것으로 평가되었으며, 상용화된 프로그램인 MS-Word에서 제공하는 요약 기능보다는 나은 성능을 보이는 것으로 평가된다.

참 고 문 헌

[1] Witten I. H., Paynter G W., Frank E., Gutwin C., Nevill-Manning C. G., "KEA : Practical Automatic Keyphrase Extracting," ACM Digital Library, pp. 254-255, 1999.

[2] Owen de Kretser, Moffat. "Needles and Haystacks: A Search Engine for Personal Information Collections," Proceedings of the 23rd Australasian Computer Science Conference, Canberra, pp. 58-65, 2000.

[3] Francine R. Chen, Dan S. Bloomberg, "Extraction of Inductive Summary Sentences from Imaged Documents," Proceedings of the International Conference on Document Analysis and Recognition, pp. 227-232, 1997.

[4] Min-Yen Kan, Kathleen R. McKeown, Judith L. Klavans "Applying Natural Language Generation to Indicative Summarization," Proceedings of the 8th European Workshop on Natural Language Generation, Toulouse, France, pp. 92-100, 2001.

[5] Min-Yen Kan, Kathleen R. McKeown, Judith L. Klavans "Domain-specific Informative and Indicative Summarization for Information Retrieval," Proceedings of the Document Understanding Workshop, New Orleans, USA: September, 2001.

[6] 한경수, "질의 기반을 이용한 적합성 피드백 기반 자동문서 요약", 고려대학교 컴퓨터학과 석사 논문, 2000.

[7] Mark Sanderson, "Accurate User Directed Summarization from Existing Tools," 1999.

[8] Jade Goldstein and Chin-Yew Lin, "Summarizing Text Documents: Sentence Selection and Evaluation Metrics," 1999.

[9] Je Ryu, Kwang-Rok Han, etal,, "Automatic Extraction of Core Sentences from Document," Proceedings of the International Conference on Electronics, Information & Communications, 2000.

[10] 류제, 한광록, 손석원, 임기욱, "단어의 공기 관계 그래프를 이용한 문서의 핵심 문장 추출에 관한 연구", 정보처리논문지, 제7권 제11호, pp. 3427-3437, 2000.

[11] 류제, 한광록, "단어의 공기 관계 그래프를 이용한 인터넷 문서의 키워드 추출", HCI2000 학술대회발표논문집 9권 1호, pp. 894-899, 2000.

[12] 강승식, "한글 형태소 분석 HAM 라이브러리", http://nlp.kookmin.ac.kr

[13] Frank, E., Paynter, G.W., Witten, I.H., Gutwin, C., Nevill-Manning, C.G., "Domain-Specific Keyphrase Extraction," Proceedings of the 16th International Joint Conference on Artificial Intelligence, Morgan Kaufmann Publishers, San Francisco, CA, pp. 668-673, 1999.

[14] Ting, K.M., Witten, I.H., "Issues in stacked generalization", Journal of Artificial Intelligence Research, 10, pp. 271-289, 1999.

[15] Witten I.H, Frank E, "Data Mining Practical Machine Learning Tools and Techniques with Java Implementations," Morgan Kaufmann Publishers, pp. 188-193. 1999.

[16] G. Salton, A. Singhal, C. Buckley, M. Mitra, "Automatic Text Structuring and Summarization," Information Processing & Management, 1997.

한 광 록

1984년 인하대학교 전자공학과 졸업(공학사). 1986년 인하대학교 대학원 정보공학 전공(공학석사). 1989년 인하대학교 대학원 정보공학 전공(공학박사). 1989년~1991년 한국체육과학원 선임 연구원 1991년~현재 호서대학교 컴퓨터공학부 교수. 1999년~현재 호서대학교 벤처전문대학원 교수. 2001년~2002년 ISI University of South California 방문연구원. 관심분야는 정보검색, 자연어처리, 기계번역, HCI, 지능형 에이전트

오 삼 권

1980년~1984년 삼성전자 통신 연구소 1994년 Queen's Uinversity, Canada(컴퓨터 과학 박사). 1994년~1995년 한국전자통신연구원. 1995년~현재 호서대학교 컴퓨터공학부 부교수. 관심분야는 Distributed system, Computer Game, Communication protocol, Fault tolerance

임 기 욱

1977년 인하대학교 공과대학 전자공학과 졸업. 1987년 한양대학교 전자계산학 석사. 1994년 인하대학교 전자계산학 박사 1977년~1983년 한국전자기술연구소 선임연구원. 1983년~1988년 한국전자통신연구소 시스템소프트웨어 연구실장. 1988년~1989년 미 캘리포니아 주립대학(Irvine) 방문연구원. 1989년~1996년 한국전자통신연구원 시스템연구부장 주전산기(타이컴) III, IV개발 사업책임자. 1997년~1999년 정보통신연구진흥원 정보기술전문위원. 2001년~2003년 한국전자통신연구원 컴퓨터소프트웨어 연구소장. 2000년~현재 선문대학교 지식정보산업공학과 교수. 관심분야는 실시간데이타베이스시스템, 운영체제, 시스템구조