# 한국어 문서의 통계적 정보를 이용한 문서 요약 시스템 구현

강 상 배*, 조 혁 규**, 권 혁 철*, 박 재 득***, 박동인***
* 부산대학교 전자계산학과, ** 성심외국어전문대학 경영정보과,
*** 시스템공학연구소 자연어정보처리연구부

# Implementation of the Text Abstraction System using the Statistical Information of Korean Documents

Sang-Bae Kang*, Hyuk-Kyu Cho**, Hyuk-Chul Kwon*,
Jae-Deuk Park***, Dong-In Park***
* Pusan National University, Department of Computer Science
** Sungsim Junior College of Foreign Languages, Dept. of Management Information
*** SERI, Dept. of NL Information Processing

**요 약**

이 논문에서는 문장 유사도 측정 기법과 말뭉치 정보를 이용한 문서요약 시스템을 구현하였다. 문서 요약은 문서에서 문장 단위로 단어를 추출하여 문장을 단어의 벡터로 표현하고, 문서 내 단어의 출현빈도와 말뭉치 내 단어의 사용빈도를 이용하여 각 문장의 중요도를 계산한다. 그리고 중요도가 높은 상위 몇 위의 문장을 요약문장으로 추출한다. 실험 결과, 문서내 단어빈도의 중요도를 낮추고, 말뭉치내 일반 사용빈도를 단어의 가중치에 추가했을 때 가장 좋은 효율을 보였다. 또 요약하고자 하는 문서와 유사한 말뭉치를 사용했을 때 높은 효율을 보였다.

## 1. 서론

인터넷의 급격한 사용과 컴퓨터의 보급으로 대부분의 문서들이 디지털화되고 있다. 또한 회사에서도 문서 결재 시스템의 사용이 증가하여 종이가 없는 사무실이 늘어나고 있다. 이렇게 급격히 증가하는 디지털 문서를 효율적으로 검색하기 위해 사용자는 정보검색 시스템을 사용한다. 그러나 정작 정보검색 시스템을 사용하여 결과를 얻었다고 하더라도, 결과에서 자신이 원하는 문서를 찾기 위해선 결과 내에 포함된 모든 문서의 내용을 살펴 보아야만 한다. 게다가 검색된 문서의 양도 상당하여 검색 결과로 얻어진 모든 문서의 내용을 모두 살펴볼 수 없다. 시스템이 문서의 요약문(문서의 첫 몇 문장이나 수작업을 통한 요약문)을 보여줌으로써 이를 해결하고 있다.

그러나 문서를 요약하기 위해서는 상당한 비용이 든다. 문서를 요약하는 것은 전적으로 수작업에 의존해야 하며, 요약하고자 하는 문서의 양이 상당할 경우에는 엄청난 시간이 필요하다. 따라서 보다 적은 비용으로 적절한 효율성을 가진 자동 문서 요약 시스템을 도입한다면, 검색의 효율을 높일 수 있을 뿐만 아니라, 수작업에 의존한 문서요약 비용을 감소시킬 수 있다. 그러나 자동 문서 요약은 수작업에 의한 요약문에 비해 미려하지 못하고 정확도가 떨어지는 등의 단점은 있지만 비용 측면에서 충분히 보상될 수 있다.

이 논문에서 제시한 문서 요약 시스템은 문서로부터 요약문이 될 수 있음직한 문장을 추출하는 기능을 가진다. 문서 내부에서 사용되는 단어의 빈도만을 사용했을 때 발생하는 단점을 보완하기 위해 말뭉치에 나타난 단어의 빈도를 사용한다. 예를 들어 한 문서 내에 많이 나타난 단어가 있다면 이는 문서 내에서는 중요한 의미를 가지는 단어라고 가정할 수 있다. 하지만 그 단어가 일반적인 문서에서도 많이 나타나는 단어라면, 그 문서 뿐만 아니라 다른 문서에도 충분히 많이 나타날 가능성이 존재하므로,

그 단어는 그 문서 내에서 중요하지 않은 단어일 가능성이 크다. 이런 단어의 중요도는 말뭉치를 이용하여 보완 조정할 수 있다.

따라서 본 논문에서는 문서 내의 사용빈도와 말뭉치 내의 일반 사용빈도를 고려하여 요약문장을 추출하고자 한다.

본 논문은 다음과 같이 구성된다. 2장에서는 기존의 문서 요약 연구에 대해 기술하고, 3장에서는 문서 요약 기법에 대해 설명한다. 4장에서는 이 논문에서 제시한 문서요약기법에 대한 실험과 결과에 대해서 설명하겠다. 그리고 5장에서는 결론 및 향후 연구 방향에 대해 설명한다.

## 2. 기존 연구

문서 요약 시스템은 이미 60년대 중반부터 개발 및 연구되기 시작했다[1]. 그러나 인터넷의 사용이 급격히 증가하기 시작하고, 문서의 디지털화가 급속화되기 시작한 90년대 이래로 문서 요약 시스템의 연구 및 개발이 보다 활성화되고 있다.

문서 요약 기법은 크게 통계적인 방법, 계산 언어학적 방법, 수사학적 방법 등으로 나눌 수 있다.

통계적인 방법은 통계적인 자료를 이용하거나, 학습데이터를 이용하여 요약 문장이 될 수 있는 문장의 특성을 추출하거나 학습한 후, 이를 요약하고자 하는 문서에 적용하는 기법이다.[3]

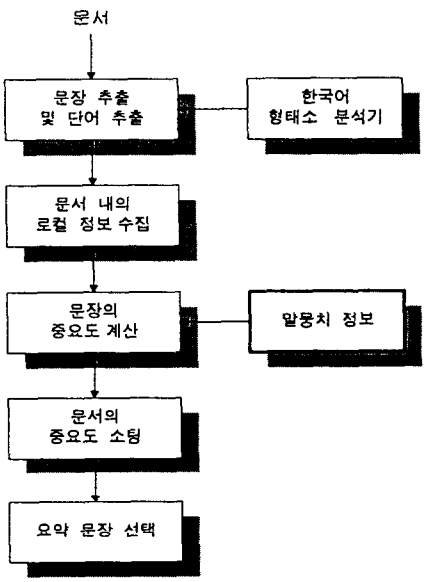
계산언어학(computational linguistics)적 방법은 자연언어 처리기법을 기반으로 담화 구조(discourse structure)나 문서 구조(text structure)를 분석하여, 문서의 내용을 파악한 후 요약 문장을 추출하거나 요약문을 생성하는 기법이다.[6,7]

수사학적(rhetorical) 방법은 수사학적 지식을 기반으로 문장의 패턴이나 문장 간의 접속관계(rhetorical relations) 등을 이용하여 문서를 분석하고, 분석된 문서구조나 담화구조를 이용하여 요약문장을 추출하는 기법이다.[2]

또 문장 간의 유사성을 비교하여 문장을 그룹화하여 분류한 후, 요약문장이나 요약 문단을 추출하는 방법[5]이나 제목, 헤드라인(headline) 등 그 문서를 요약할 단서가 될 수 있는 정보를 골(goal)로 선택하여 골과 가장 근접한 문장을 선택하는 기법[6, 7]도 연구되었다.

## 3. 문서 요약

이 논문에서는 통계적 방법을 사용하여 문서로부터 요약문에 포함될 수 있는 문장을 추출하는 시스템에대해서 기술한다.




[그림 1] 문서 요약 시스템의 구성

이 논문에서 제시한 문서 요약 시스템의 구조는 [그림1]과 같다. 한국어 형태소 분석기를 사용하여 문장을 분리하고, 문장에서 단어(명사)를 추출한다. 추출된 단어를 기반으로 문서 내에 단어의 사용빈도를 계산한다. 문서 내의 단어 정보를 수집한 후, 말뭉치(corpus)에서 나타난 단어 정보를 이용하여 문장의 중요도(유사도)를 계산한다. 문장의 유사도는 문서와 문장 간의 유사한 정도로써 코사인 유사계수(cosine measure)를 이용하여 계산된다. 이 유사도 값을 이용하여 소팅한 후 상위 몇 위까지의 문장을 요약문장으로 추출한다.

### 3.1. 문서 요약 기법

문서에서 각 문장이 차지하는 중요도는 그 문서와 각 문장에 대한 유사도를 이용해서 계산할 수 있다. 즉 유클리디언 벡터 공간에서 문서의 벡터는 $D=(d_1, d_2, \ldots, d_n)$으로 나타내고, $S_i$ 문장의 벡터는 $S_i = (s_{i1}, s_{i2}, \ldots, s_{im})$으로 나타낸다. 여기서

$d_i$는 문서 D에 포함된 단어 i가 가지는 가중치를 나타내고, $S_i$ 문장에서 $s_{ik}$는 문장에 포함된 단어 k가 가지는 가중치이다. 이때 그 문서와 가장 유사한 문장을 구하여 이를 그 문서의 요약 문장으로 구성할 수 있다.

즉, 문서가 k개의 문장으로 구성되어 있다면

$$D = (d_1, d_2, \ldots, d_n)$$

$$S_1 = (s_{11}, s_{12}, \ldots, s_{1m})$$
$$S_2 = (s_{21}, s_{12}, \ldots, s_{2m})$$
$$\cdots\cdots$$
$$\cdots\cdots$$
$$S_k = (s_{k1}, s_{k2}, \ldots, s_{km})$$

으로 나타낼 수 있다.

문서와 각 문장의 유사도는 코사인 유사계수를 이용하여 다음과 같이 계산할 수 있다.

$$SIM(D, S_i) = \frac{\sum(d_k \cdot s_{ik})}{\sqrt{\sum(d_k^2) \cdot \sum(s_{ik}^2)}}$$

코사인 유사계수함수를 이용해서 계산된 값은 문서와 문장 간의 유사도이다. 문서와 문장이 유사하다면 그 문장은 문서를 대표할 수 있는 요약문장에 포함시킬 수 있으며, 문서를 대표할 수 없는 문장이라면 유사도 값이 낮을 것이다. 이 유사도 값을 이용하여 각 문장의 중요도를 계산하고, 요약문장을 추출한다.

### 3.2. 말뭉치 정보의 이용

문서에 나타나는 단어의 로컬빈도만을 이용하여 문장의 유사도를 계산하기 위해 문서와 문장에 나타나는 단어가 가지는 가중치는 다음과 같다.

$d_i$ = (문서 내의 단어 $i$의 빈도)

$s_{ik}$ = ($i$문장 내의 단어 $k$의 빈도)

문서 내에 나타나는 단어의 로컬빈도만을 이용하여 요약문장을 추출하면, 문서 내에서 높은 빈도로 나타나는 단어에 의존적이다. 문서를 요약할 때 문서 내에 나타나는 단어의 사용빈도가 높은 영향을 발휘하지만, 불필요한 단어의 사용빈도가 높아서 효율이 떨어질 수 있다. 예를 들어 '이용', '대표', '사실', '필요' 등 일반적으로 다른 문서에서도 높은 빈도로 나타나는 단어가 포함된 문장은 그렇지 않은 문장에 비해 유사도가 높게 나타난다. 따라서 이런 단어의 중요도는 낮추어 주어야 한다. 또 말뭉치에 나타나지 않는 단어는 그 문서에 특수하게 사용된 것이므로 중요도를 높여주어야 한다. 문서 내의 로컬 중요도가 높게 나타나지만 실제 일반적인 중요도가 낮은 단어의 영향을 줄이기 위해서, 그리고 로컬 중요도는 낮지만, 일반 사용빈도가 낮은 단어의 영향을 높이기 위해 문서 집합인 말뭉치를 사용한다.

학습데이터를 이용해서 학습을 통해 통계적인 정보를 얻을 수 있다. 그러나 학습데이터를 구성하는 것이 어렵고, 학습데이터 구성과정에서 많은 수작업을 요구한다. 대신 말뭉치 내의 단어 사용빈도를 이용하여 문서를 요약한다면, 검색 시스템을 이용해 말뭉치를 데이터베이스에 저장하고, 이를 검색함으로 수작업이 필요없다. 단어빈도의 검색을 위해 기존 정보검색 시스템을 그대로 이용하므로, 학습용 프로그램을 구성할 필요도 없다는 장점을 가지고 있다.

말뭉치에 나타나는 단어의 빈도를 그 단어가 일반적인 문서에 나타날 사용빈도라고 가정한다. 말뭉치의 크기에 따라 단어의 중요도가 달라진다. 말뭉치의 크기가 크면, 단어는 일반사용빈도와 유사한 빈도를 가질 것이다. 그러나 말뭉치의 크기가 작으면, 일반사용빈도와의 차이가 커질 수 있다.

또 말뭉치의 종류에 따라 단어의 중요도가 달라진다. 요약하고자 하는 문서와 유사한 종류의 문서를 보유한 말뭉치라면 그 문서에 나타나는 단어의 일반 사용빈도는 보다 정확해질 것이다. 그러나 요약하고자 하는 문서와 다른 특성을 가진 말뭉치라면 그 단어가 가진 일반적인 사용빈도와는 거리가 먼 사용빈도를 얻을 것이다. 이 논문에서는 신문기사 말뭉치와 KTSET 테스트데이터 말뭉치를 사용한 실험 결과를 통해 말뭉치의 종류에 따른 효율의 차이를 보인다.

위에서 기술한 것처럼 말뭉치의 일반적인 사용빈도를 고려한 단어의 가중치는 다음과 같다.

$d_i$ = (문서 내 단어 $i$의 빈도)
* (말뭉치 단어 $i$의 빈도)

$s_{ik}$ = ($i$문장 단어 $k$의 빈도)
* (말뭉치 내 단어 $k$의 빈도)

말뭉치는 특정단어의 일반 사용빈도를 가지므로, 문서 내에서 높은 빈도를 가진다고 하더라도 말뭉치 내에서 그 단어의 빈도가 높으면, 그 단어의 가중치를 낮추어 주어서

불필요하게 유사도가 높아지는 문장의 유사도를 낮춘다.

말뭉치에 나타나는 단어 중에서 10% 이상의 문서에 나타나는 단어는 그 중요성이 적다. 영어권에서는 불용어 사전을 이용하여 흔히 사용되는 단어는 제거하지만, 한국어의 특성상 이런 불용어(stop word)를 정의하기는 쉽지 않다. 그러나 조사, 부사 등 명사가 아닌 단어를 불용어로 처리하여 색인단어에서 제외한다. 말뭉치에 나타난 단어 중 10%이상의 문서에 나타난 단어는 불용어로써 판단할 수 있고, 본 논문에서는 이를 적용하여 10%이상의 문서에 나타난 단어를 제외하고 문서에서 요약문장을 추출하는 방법도 제시한다.

## 4. 실험 및 결과

말뭉치 단어의 사용빈도를 추출하기 위해 KTSET V2.0(약 4천4백건)과 동아일보사 1년치 분량의 신문기사 데이터(약 7만건)를 사용하였다. 실험은 말뭉치를 사용한 경우와 말뭉치를 사용하지 않고 문서 내 단어빈도만을 이용한 경우를 비교 분석하고, 말뭉치의 종류에 따른 효율의 차이를 비교해 본다.

### 4.1. 실험 방법

실험에 사용한 데이터는 정보과학회 및 기타 정보과학회 관련 학회에 제출된 논문으로, 서론과 결론 부분만을 사용하였다. 데이터의 건수는 25건, 문서에 포함된 문장의 평균 개수는 22.04개, 수작업을 통해 추출한 요약문장의 평균 개수는 3.32개이다. 전산과 대학원생 3명, 전산과 대학생 1명의 수작업을 통하여 추출한 요약 문장과 시스템이 추출한 요약 문장을 비교하여 문서 요약의 효율성을 비교 분석하였다.

말뭉치를 사용하지 않고 문서내 단어 빈도만을 이용하여 문장을 추출하였을 때 나타나는 재현율과 동아일보 말뭉치와 KTSET 테스트 데이터 말뭉치를 이용했을 때 나타나는 재현율을 비교하였다.

### 4.2. 실험 결과

[표 1]의 방법은 단어의 가중치를 구할 때 문서 내 사용빈도와 문장내 사용빈도에 제곱근(root)을 사용하여 문서내 단어의 중요도를 줄이고 말뭉치 사용빈도의 중요성을 높인 것이다. 또 사전에 나타나지 않는 단어의 말뭉치 사용빈도는 중간값 (Max idf + Min idf)/2를 사용하였다.

[표 1]은 말뭉치를 사용하지 않은 것에 비해 신문기사 말뭉치와 KTSET 말뭉치를 사용하는 것이 효율적임을 보여준다.

| 말뭉치 종류 / 추출 문장 수 | No Corpus | 신문기사 | KTSET |
|---|---|---|---|
| 2 | 0.2592 | 0.3160 (+21.9%) | 0.3396 (+31.0%) |
| 4 | 0.4230 | 0.4640 (+9.7%) | 0.5108 (+20.8%) |
| 6 | 0.6368 | 0.6680 (+3.8%) | 0.6548 (+2.8%) |
| 8 | 0.7068 | 0.7268 (+2.8%) | 0.7448 (+5.4%) |
| 10 | 0.7480 | 0.7644 (+2.2%) | 0.8080 (+8.0%) |
| 12 | 0.7920 | 0.7848 (-0.9%) | 0.8472 (+7.0%) |
| 14 | 0.8380 | 0.8620 (+2.8%) | 0.8988 (+7.3%) |
| 증가율 | - | + 5.3 % | + 10.3 % |

[표 1] 추출 문장 수에 따른 요약문장의 재현율

신문기사 말뭉치를 사용한 경우 평균 5.3% 재현율 증가를 보이고 있으며, KTSET 말뭉치의 경우, 평균 10.3% 재현율 증가를 보이고 있다.

요약하고자 하는 문서에서 문서크기의 대략 20%(위의 표에서 네 개 문장)를 추출하였을 때 나타날 요약문장의 재현율은 51.08 %이고, 이때 정확도는 42.4 %이다.

[표 2]는 말뭉치에 나타난 단어에 대해서 말뭉치 내의 10%이상의 문서에 나타난 단어를 제거하는 방법을 사용한 결과이다.

10%이상의 문서에 나타나는 단어를 불용어로 처리하여 문장의 유사도 계산에서 제외한다. [표 2]에서 보는 것처럼 말뭉치의 종류에 따라 재현율에서 차이가 나는 것을 알 수 있다. 실험데이터와 유사한 종류의 말뭉치를 사용했을 때는 3.7 %의 증가율을 나타냈고, 실험데이터와 관련이 없는 신문기사 말뭉치를 사용한 경우 오히려 증가율이 0.68 %로 떨어졌다. 이는 말뭉치에 따라 서로 다른 단어의 빈도를 나타내므로 신문기사 말뭉치에서 10 %이상의 문서에 나타나는 단어는 요약 문장 추출시에 중요하게 사용된 단어를 포함하고 있다는 것을 의미한다.

| 말뭉치 / 추출 종류 / 문장 수 | No Corpus | 신문기사 | KTSET |
|---|---|---|---|
| 2 | 0.2592 | 0.3028 (+16.8%) | 0.3028 (+16.8%) |
| 4 | 0.4230 | 0.4160 (-1.6%) | 0.4428 (+4.7%) |
| 6 | 0.6368 | 0.5808 (-9.6%) | 0.5672 (-12.3%) |
| 8 | 0.7068 | 0.6588 (-7.3%) | 0.7100 (+0.4%) |
| 10 | 0.7480 | 0.7280 (-2.7%) | 0.7912 (+5.8%) |
| 12 | 0.7920 | 0.7712 (-2.7%) | 0.8528 (+7.7%) |
| 14 | 0.8380 | 0.8520 (+1.7%) | 0.8900 (+6.3%) |
| 증가율 | - | - 0.68 % | + 3.7 % |

[표 2] 추출한 문장 수에 따른 요약문장의 재현율

따라서 요약하고자 하는 문서에서 중요한 역할을 했던 단어가 10%이상의 문서에 나타나는 단어집합에 포함되어 유사도 계산에서 제외되므로 효율이 떨어진다. [표 2]는 요약 문장 추출시 요약하고자 하는 문서와 특성이 상이한 종류의 말뭉치를 사용하면, 요약의 효율성이 감소한다는 것을 보여준다.

말뭉치의 단어 사용빈도를 조절하기 위해 여러가지 다양한 방법을 사용하여 실험하였다. 방법은 다음과 같다.

(1) 사전에 없는 단어는 유사도 계산에서 제외하는 경우
(2) 사전에 없는 단어는 가장 낮은 말뭉치 빈도를 적용한 경우
(3) 사전에 없는 단어는 가장 높은 말뭉치 빈도를 적용한 경우
(4) (3)방법을 적용하고, 말뭉치에서 10%, 5%, 1%이상의 문서에 나타나는 단어는 제거하는 경우
(5) 말뭉치에 나타나지 않는 단어의 말뭉치 내 사용빈도는 중간값을 적용하고, 신문기사 말뭉치에서 10%, 5%, 1%이상의 문서에 나타나는 단어는 문장과 문서에서 제거하고, 그 외의 단어는 유사 종류인 KTSET 말뭉치에 나타나는 사용빈도를 적용한 경우
(6) 문서 내 단어 정보의 효과를 감소시키고(문서 내 단어 빈도에 제곱근을 적용해 빈도를 계산), 말뭉치에 없는 단어는 중간 말뭉치 빈도를 적용한 경우
(7) 문서 내 단어 정보의 효과를 감소시키고(문서 내 빈도는 제곱근을 이용해 빈도를 계산), 말뭉치에 없는 단어는 가장 높은 말뭉치 빈도를 적용한 경우

아래의 표는 말뭉치를 사용하지 않았을 때 재현율과 말뭉치를 사용했을 때 재현율과의 차이를 퍼센트로 표시한 것이다.

| 말뭉치 / 방법 | 신문기사 | KTSET |
|---|---|---|
| (1)방법 | - 1.2 % | - 0.9 % |
| (2)방법 | - 0.9 % | - 1.2 % |
| (3)방법 | - 0.4 % | + 2.7 % |

[표 3] 사전에 없는 단어의 사용빈도를 조정한 방법의 재현율 차이

[표 3]은 말뭉치에 존재하지 않는 단어가 문서에 나타났을 때, 문서와 문장의 유사도계산 시 단어가 가지는 일반 사용빈도를 다르게 계산하였다. 말뭉치에 존재하지 않는 단어는 말뭉치 내 사용빈도를 가장 높은 값으로 설정하였을 때, 가장 좋은 효율을 나타냈다.

| 말뭉치 / 방법 | 신문기사 | KTSET |
|---|---|---|
| (4)방법, 10%제거 | + 0.3 % | + 3.7 % |
| (4)방법, 5%제거 | + 0.03 % | + 0.2 % |
| (4)방법, 1%제거 | - 0.1 % | + 0.2 % |
| (5)방법, 10%제거 | NA | + 1.5 % |
| (5)방법, 5%제거 | NA | + 1.6 % |
| (5)방법, 1%제거 | NA | - 0.7 % |

[표 4] 빈도가 높은 단어를 제거하는 방법에 대한 재현율 차이

[표 4]에서 (4)방법은 말뭉치에 없는 단어의 처리는 [표 3]에서 가장 좋은 결과를 나타낸 방법을 사용하고, 말뭉치 내 빈도가 10%, 5%, 1%이상의 문서에 나타나는 단어는 제거하고 문서와 문장의 유사도를 계산한 결과이다. (5)방

법은 말뭉치에 없는 단어는 중간 말뭉치 빈도를 적용하고, 특성이 다르지만 보다 일반적인 신문기사 말뭉치 중에서 10%, 5%, 1%이상의 문서에 나타나는 단어를 문장과 문서에서 제거하고, 그 외의 단어는 특성이 유사한 KTSET 말뭉치에 나타나는 사용빈도를 적용한다. [표 4]에서 NA는 신문기사 말뭉치에서 보다 일반적인 단어를 선택한 것으므로 값을 계산할 수 없다. [표 4]에서는 (4)번 방법과 10%이상의 문서에 나타나는 단어를 제거했을 때 가장 좋은 효율을 나타냈다.

| 말뭉치 / 방법 | 신문기사 | KTSET |
|---|---|---|
| (6)방법 | + 5.3 % | + 10.3 % |
| (7)방법 | + 3.5 % | + 7.4 % |

[표 5] 문서 내 단어빈도는 줄이고, 말뭉치내 빈도의 중요도를 높인 방법의 재현율 차이

[표 5]는 문서 내의 단어 빈도의 중요성을 낮추고, 말뭉치내 일반사용빈도를 조절하는 방법의 결과이다. (6)방법은 문서내 단어빈도는 제곱근을 이용하여 사용빈도를 조정하였고, 중간 말뭉치빈도를 적용하였다. (7)방법은 문서내 단어빈도는 제곱근을 이용하여 사용빈도를 조정하였고, 가장 높은 말뭉치빈도를 적용하였다. 문서내 사용빈도를 조절하였을 때, 말뭉치에 나타나지 않는 단어의 말뭉치내 일반사용빈도는 중간치를 사용하는 것이 보다 좋은 검색 효율을 나타냈다. (6)방법을 적용했을 때 가장 높은 효율을 나타냈다.

## 5. 결론 및 향후 연구 방향

이 논문에서는 문장 유사도 측정 기법과 말뭉치 정보를 이용하여 문서에서 요약 문장을 추출하는 문서 요약 시스템을 구현하였다. 시스템은 문장 및 문서 내의 단어빈도를 추출하는 부분과 문장의 중요도를 계산하는 부분, 그리고 소팅 및 선택 부분으로 나눌 수 있다. 문장의 중요도는 문서와 문장 간의 코사인 유사계수 함수를 이용하여 계산하고, 이 유사도를 문장의 중요도 값으로 선택하였다. 단어의 가중치는 문서 내의 로컬 사용빈도와 말뭉치 내의 일반 사용빈도를 이용하여 계산하였다. 문서크기의 대략 20%로 문장을 추출하였을 때 재현율은 51.08 %, 정확도는 42.4 %이다. 요약하고자 하는 문서데이터와 유사한 종류의 말뭉치인 KTSET을 사용했을 때 보다 높은 효율성을 나타냈다.

통계적인 정보 외에 문서가 가진 언어적인 특징도 고려해야 한다. 문장의 패턴, 좌우 문장의 이용, 문장의 위치정보, 단서 단어의 선택, 시소러스를 이용한 문장의 확장 등의 기법도 도입해야 할 것이다.

## [참고문헌]


[1] H. P. Edmundson, "Problems in Automatic Abstracting," *Communications of the ACM*, Vol.7, No.4, June 1964, pp98-101

[2] S. Miike, E. Itoh, K. Ono and K. Sumita, "A Full-Text Retrieval System with a Dynamic Abstract Generation Function," *Proceedings of 17-th SIGIR Conference*, pp152-161, 1994

[3] J. Kupiec, J. Pedersen, and F. Chen, "A Trainable Document Summarizer," *Proceedings of 18-th SIGIR Conference*, pp68-73, 1995

[4] K. McKeown and D. R. Radev, "Generating Summaries of Multiple News Articles," *Proceedings of 18-th SIGIR Conference*, pp74-82, 1995

[5] M. Mitra, A. Singhal, and C. Buckley, "Automatic Text Summarization by Paragraph Extraction," *Intelligent Scalable Text Summarization*, pp39-46, 1997

[6] C. D. Paice, "Constructing Literature Abstracts by Computer: Techniques and Prospects," *Information Processing and Management*, 26(1), pp171-186, 1990

[7] R. Ochitani, Y. Nakao, and F. Nishino , "Goal-Directed Approach for Text Summarization" *Intelligent Scalable Text Summarization*, pp47-50, 1997

[8] G. Salton and M. J. McGill, *Introduction to Modern Information Retrieval*, McGraw-Hill, 1983

[9] W. B. Frakes and R. Baeza-Yates, *Information Retrieval:Data Structures & Algorithms*, Prentice- Hall Inc., 1992

[10] G. Salton, *Automatic Text Processing*, Addison-Wesley Publishing Company, 1989